University of Oulu
Department of Civil Engineering
Structural Engineering Laboratory
Publication 56

PROCEEDINGS OF THE 6TH FINNISH MECHANICS DAYS

September 5-6, 1997
Oulu, Finland

Jukka Aalto and Tapio Salmi (eds.)

University of Oulu
Structural Engineering Laboratory
P. O. Box 191
90101 Oulu
FINLAND
Tel. +358-8-5534370
Fax. +358-8-5534322

Scientific committee

Professor Jukka Aalto (Chairman)
University of Oulu

Professor Martti Mikkola
Helsinki University of Technology

Professor Pekka Neittaanmäki
University of Jyväskylä

Professor Erkki Niemi
Lappeenranta University of Technology

Professor Antti Pramila
University of Oulu

Associate Professor Eero-Matti Salonen
Helsinki University of Technology

Professor Markku Tuomala
Tampere University of Technology

Organizing committee

Professor Jukka Aalto (Chairman)
Library Assistant Eila Keränen
Assistant Kai Kuula
Professor Antti Pramila
Associate Professor Tapio Salmi
Associate Professor Stig-Göran Sjölind

Organizers

University of Oulu
Structural Engineering Laboratory
Engineering Mechanics Laboratory

ISBN 951-42-4712-4
ISSN 0782-3096

Preface

These proceedings contain the papers presented at the Sixth Finnish Mechanics Days held in Oulu, Finland, 5-6 September 1997. The First Finnish Mechanics Days were held in Oulu in 1982, the Second in Tampere in 1985, the Third in Espoo in 1988, the Fourth in Lappeenranta in 1991 and the Fifth in Jyväskylä in 1994.

The purpose of the Finnish Mechanics Days is to bring together researchers, post-graduate students, teachers and practising engineers, whose interest lies in the field of mechanics and computational methods. Although the invited lectures of the Finnish Mechanics Days have mainly been given by foreign researchers, the organizers decided this time to invite four prominent finnish experts M. Mikkola, M. A. Ranta, J. Koski, and R. Stenberg as invited speakers. In addition to these 29 contributed papers were presented.

The Organizing Committee wishes to express its sincere gratitude to all the authors for their hard and successfull work in preparing their contributions. The economical support of Tauno Tönning foundation for publishing this volume is also gratefully acknowledged.

Jukka Aalto and Tapio Salmi

Contents

Structural models

Structural analysis

Nonlinear structures

Experimental methods

Soil Mechanics

Finite element method

Dynamics

THERMOMECHANICAL MODEL OF FREEZING SOIL BY USE OF THE THEORY OF MIXTURES

J.Hartikainen & M.Mikkola

Laboratory of Structural Mechanics

Helsinki University of Technology

P.O.Box 2100, FIN-02015 HUT, FINLAND

ABSTRACT

A thermomechanical theory for arbitrary mixtures based on the theory of mixtures, the principles of continuum mechanics and the macroscopic thermodynamics is introduced. A thermomechanical model of the freezing of saturated soil is established by applying the general theory. The saturated soil is considered as a mixture of skeleton, water and ice. The proposed model is capable of describing the thermal, hydraulic and mechanical features of the frost phenomenon. The complete system is reduced to a computationally convenient set of equations and the finite element implementation is done in one-dimensional case. Numerical simulation of a soil freezing test is carried out and the results are presented.

1 INTRODUCTION

During the last few decades, comprehensive research has been in progress to improve the understanding of physical nature of freezing soil. A well established feature is that water in fine-grained soil freezes over a temperature range, see Fig. 1, [1], [17], [24], [32]. This leads to the so-called frozen fringe, a "mushy" region, in which water coexists in solid, liquid and gaseous phases. Under certain conditions the frost phenomenon is initiated. It is a rather complicated combination of several interactive physical

processes: freezing of water in soil under temperature gradient creates a cryogenic suction, a driving force, which induces migration of water from the unfrozen soil to the frozen fringe; on one hand the ice formation expands pores of the soil causing the frost heave, but on the other hand the negative pore-water pressure may induce consolidation of the unfrozen soil; in transient freezing, the frozen fringe advances continuously. The characteristics of the frost phenomenon depend mainly on soil type, applied loads and freezing conditions.

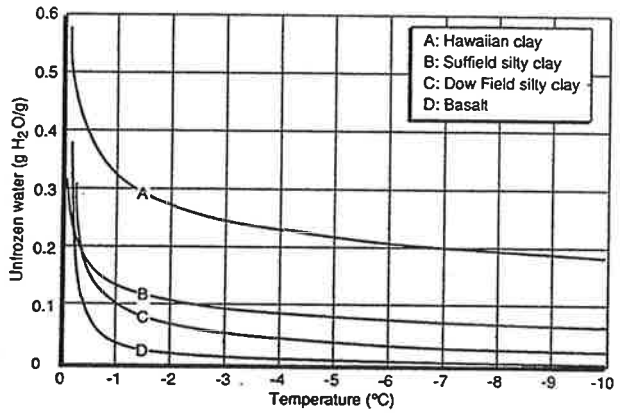


Figure 1: Typical unfrozen water content-temperature relationships in frozen soils [1].

As a result of the research several models, well summarized by [19] and by [25], [26], have been developed to predict the outcome of a particular frost process produced by certain conditions as an input. Kujala [25], [26] has classified them as empirical, semi-empirical, hydrodynamic, rigid-ice and thermomechanical models. The empirical models, e.g. [2], are based on purely empirical observations using field observations and frost heave tests. In semi-empirical models the physical nature of frost heave is also used as a basis, e.g. [5]. Hydrodynamical models, e.g. [16], [11], [14], [15], couple heat and mass transfer in freezing soil based on an analogy between water transport in unsaturated soils and water transport in partially frozen soil. The rigid-ice model, e.g. [13], [27], [28], is based on the theory of secondary frost heave. The segregation potential concept is defined as the ratio of the water migration rate to the overall gradient in the frozen fringe, e.g. [21], [22]. The most advanced models are the thermomechanical ones, e.g. [7], [9], which are based on the theory of mixtures, continuum mechanics and macroscopic thermodynamics. They take into account thermal, hydraulic and mechanical aspects being able to describe heat transfer, freezing, suction and migration

of water, and deformations of skeleton.

This work is a direct continuation to the work of Frémond and Mikkola [9]. The thermomechanical theory is extended and verified to arbitrary mixtures. It is shown that not only the conservation laws but also the constitutive relations can be established in a general form without specifying the material beforehand. With the aim of the application of the general theory to freezing of saturated soil, the physicochemical structure of the soil is described and the nature of corresponding material parameters is explained. The resulting model is capable of describing the relevant processes of the frost phenomenon. The model is reduced to a computationally convenient set of equations and the finite element implementation is done in one-dimensional case. Numerial simulation of a soil freezing test is carried out and the results are presented.

2 THERMOMECHANICAL THEORY

The key idea in the application of continuum mechanics to heterogeneous media like a mixture of several constituents is to extend the conventional definition of continuum. The discrete structure of matter is disregarded and each constituent is assumed to be spread over the spatial domain in a continuous manner. Hence, all the constituents are able to coexist at any point of the region. As a consequence of the smoothing the variables and functions are continuous and differentiable describing the state of the medium in some average sense.

The volume fraction β^k of constituent k is defined as the relation of the apparent density ρ^k and a constant reference (bulk) density $\bar{\rho}^k$:

$$\beta^k = \frac{\rho^k}{\bar{\rho}^k} \tag{1}$$

In addition, the volume fractions satisfy the obvious conditions

$$\sum_k \beta^k = 1\,,\ \beta^k \geq 0. \tag{2}$$

2.1 Basic concepts of kinematics

The state of motion of a constituent k at an arbitrary instant of time t is described by a velocity field $\vec{U}^k(\vec{x},t)$. For a constituent j of solid type, it is more convenient to describe the motion by its displacement field $\vec{u}^j(\vec{x},t)$; the velocity $\vec{U}^j(\vec{x},t)$ is obviously the

material time derivative $\frac{d^j}{dt}\vec{u}^j$. Because constituents have in general different velocities at the same macroscopic point of the mixture, a reference velocity field $\vec{U}^*(\vec{x},t)$ is introduced in order to establish the fundamental principles for the mixture. In the classical mixture theory [30], the barycentric velocity is used as a reference while in the models based on Biot's theory [3], the movement of the solid is taken as the reference movement. The approach chosen in this paper, allows the reference movement to be arbitrary and independent of the motion of the material.

The deformations are described either by the rate of deformation

$$\mathcal{D}^k = \tfrac{1}{2}\left[\nabla\vec{U}^k + (\nabla\vec{U}^k)^T\right] \tag{3}$$

or by the strain

$$\varepsilon^j = \tfrac{1}{2}\left[\nabla\vec{u}^j + (\nabla\vec{u}^j)^T\right]. \tag{4}$$

The material time derivative of a quantity following the movement of constituent k is determined by the operator

$$\frac{d^k}{dt} = \frac{\partial}{\partial t} + \vec{U}^k\cdot\nabla. \tag{5}$$

When the material time derivative with respect to the reference movement is introduced, eq. (5) is replaced by

$$\frac{d^k}{dt} = \frac{d^*}{dt} + \vec{V}^{k*}\cdot\nabla, \tag{6}$$

where $\vec{V}^{k*}$ is the relative velocity of the constituent k with respect to the reference velocity: $\vec{V}^{k*} = \vec{U}^k - \vec{U}^*$.

The material time derivative of a quantity defined by a volume integral $\int\limits_{\mathcal{V}^k(t)} \Omega^k\, d\mathcal{V}$ is established next. Consider a control volume $\mathcal{CV}$ bounded by a control surface $\mathcal{CS}$, which encloses the moving material region $\mathcal{V}^k(t)$ at an arbitrary time t. Using the Eulerian approach the definition of material time derivative of such a quantity can then be expressed by

$$\frac{D^k}{Dt}\int\limits_{\mathcal{V}^k(t)} \Omega^k\, d\mathcal{V} = \int\limits_{\mathcal{CV}} \frac{\partial}{\partial t}\Omega^k\, d\mathcal{V} + \int\limits_{\mathcal{CS}} \Omega^k\vec{U}^k\cdot\vec{\mathrm{n}}\, d\mathcal{S}. \tag{7}$$

When the control volume is following the reference movement eq. (7) takes the form

$$\frac{D^k}{Dt}\int\limits_{\mathcal{V}(t)} \Omega^k\, d\mathcal{V} = \frac{D^*}{Dt}\int\limits_{\mathcal{CV}(t)} \Omega^k\, d\mathcal{V} + \int\limits_{\mathcal{CS}(t)} \Omega^k\vec{V}^{k*}\cdot\vec{\mathrm{n}}\, d\mathcal{S}, \tag{8}$$

which can be considered as an Arbitrary Lagrangian-Eulerian (ALE) description of the material time derivative of a quantity defined as a volume integral.

2.2 Conservation laws and entropy inequality

The general conservation law for a constituent k

$$\frac{D^k}{Dt}\int\limits_{\mathcal{V}^k(t)} \mathfrak{Q}^k\, d\mathcal{V} = -\int\limits_{\mathcal{S}^k(t)} \mathfrak{J}^k\cdot\vec{\mathrm{n}}\, d\mathcal{S} + \int\limits_{\mathcal{V}^k(t)} \mathfrak{B}^k\, d\mathcal{V} + \int\limits_{\mathcal{V}^k(t)} \mathfrak{C}^k\, d\mathcal{V} \tag{9}$$

states that the rate of change of a quantity $\mathfrak{Q}^k$ within a material domain $\mathcal{V}^k$ is the sum of the external supply given by the flux $\mathfrak{J}^k$ through the material surface $\mathcal{S}^k$ and of the external supply $\mathfrak{B}^k$ within the material domain $\mathcal{V}^k$, and of the rate of production $\mathfrak{C}^k$ due to interaction of different constituents. Employing eq. (7) and applying the Gauss's theorem to the surface integrals, eq. (9) can be brought into the form

$$\int\limits_{\mathcal{CV}} \mathfrak{C}^k\, d\mathcal{V} = \int\limits_{\mathcal{CV}} \left[\tfrac{\partial}{\partial t}\mathfrak{Q}^k + \nabla\cdot\left(\mathfrak{Q}^k\vec{U}^k\right) + \nabla\cdot\mathfrak{J}^k - \mathfrak{B}^k\right] d\mathcal{V}. \tag{10}$$

Its local form is directly

$$\mathfrak{C}^k = \tfrac{\partial}{\partial t}\mathfrak{Q}^k + \nabla\cdot\left(\mathfrak{Q}^k\vec{U}^k\right) + \nabla\cdot\mathfrak{J}^k - \mathfrak{B}^k. \tag{11}$$

The corresponding conservation law for the mixture is obtained simply by adding the contributions of different constituents and demanding that the sum of rate of productions $\mathfrak{C}^k$ vanishes:

$$\sum_k \mathfrak{C}^k = 0. \tag{12}$$

Introducing the density ρ^k and applying eq. (11) the rate of production of mass of constituent k is obtained

$$\theta^k = \tfrac{\partial}{\partial t}\rho^k + \nabla\cdot\left(\rho^k\vec{U}^k\right), \tag{13}$$

while the conservation law of mass for the mixture is

$$\sum_k \theta^k = 0. \tag{14}$$

It is convenient to introduce a specific quantity $\mathfrak{q}^k = \frac{\mathfrak{Q}^k}{\rho^k}$, and on account of (13) to rewrite the local conservation law (11) into the form

$$\mathfrak{C}^k = \rho^k \frac{d^k}{dt}\mathfrak{q}^k + \mathfrak{q}^k\theta^k + \nabla\cdot\mathfrak{J}^k - \mathfrak{B}^k. \tag{15}$$

The rate of production of the linear momentum of constituent k is obtained by application of (15)

$$\vec{m}^k = \rho^k \frac{d^k}{dt}\vec{U}^k + \theta^k\vec{U}^k - \nabla\cdot\sigma^k - \vec{f}^k, \tag{16}$$

where σ^k is the stress tensor and $\vec{f}^k$ the body force. The balance of linear momentum for the mixture is

$$\sum_k \vec{m}^k = 0. \tag{17}$$

Similary, introducing the specific internal energy e^k, the heat flux $\vec{q}^k$ and the external heat supply r^k and making use of (15) the rate of production of energy of constituent k becomes

$$\ell^k = \rho^k \frac{d^k}{dt}e^k + \left(e^k - \tfrac{1}{2}\vec{U}^k\cdot\vec{U}^k\right)\theta^k - \sigma^k : \nabla\vec{U}^k + \vec{m}^k\cdot\vec{U}^k + \nabla\cdot\vec{q}^k - r^k, \tag{18}$$

and the corresponding balance law of energy for the mixture reads

$$\sum_k \ell^k = 0. \tag{19}$$

Analogous to the establishment of conservation laws, the second principle of thermodynamics is stated. Hence, the rate of production of entropy of constituent k is

$$\gamma^k = \rho^k \frac{d^k}{dt}s^k + s^k\theta^k + \nabla\cdot\left(\frac{\vec{q}^k}{T}\right) - \frac{r^k}{T}, \tag{20}$$

where s^k is the specific entropy and T the thermodynamic temperature. Taking into account eqs. (16) and (18) and introducing the specific free energy ψ^k, defined by

$$\psi^k = e^k - Ts^k \tag{21}$$

the rate of production of entropy (20) can be brought into the form

$$T\gamma^k = \sigma^k : \nabla\vec{U}^k - \rho^k\left(\frac{d^k}{dt}\psi^k - s^k\frac{d^k}{dt}T\right) - \left(\psi^k - \tfrac{1}{2}\vec{U}^k\cdot\vec{U}^k\right)\theta^k - \vec{m}^k\cdot\vec{U}^k - \frac{\nabla T}{T}\cdot\vec{q}^k \tag{22}$$

The second principle for a set of constituents postulates that the total rate of entropy production must be non-negative:

$$\sum_k \gamma^k \geq 0. \tag{23}$$

If the process is reversible the equality in (23) holds, otherwise the process is irreversible and the left hand side of (23) is strictly positive.

2.3 Constitutive relations

The fundamental laws established so far are valid for a mixture of any number of arbitrary constituents. In order to characterize the physical behaviour of the material and to treat specific problems, so-called constitutive relations are needed. These equations describing the response of the material are determined by the variables defining the state and the dissipative properties of the material and by the expressions of specific free energy and dissipation functions of the constituents. Here, the thermodynamic temperature T and the strains ε^k and the volume fractions β^k of constituents are chosen as the state variables. They are suitable for defining the state of a mixture of liquids and solids whereas in mixtures of gases it is more convenient to replace the strain by a density and the volume fraction by a mass fraction.

The specific free energy ψ^k of a constituent k is a function of the state variables

$$\psi^k = \psi^k \left(T, \varepsilon^k, \beta^j\right). \tag{24}$$

containing all the information wanted of the nondissipative behaviour of the material. Especially, the influence of different constituents on the behaviour of constituent k is taken into account by the volume fractions.

The volume fractions, however, are not independent state variables but constrained by the conditions (2). These constraints are taken into account by including the indicator function $\mathcal{I}(\beta^1, \ldots, \beta^n)$ into the free energy of the mixture

$$\sum_k \rho^k \psi^k = \sum_k \rho^k \tilde{\psi}^k + T\mathcal{I}(\beta^1, \ldots, \beta^n). \tag{25}$$

Hence, the specific free energy ψ^k of constituent k is represented by

$$\psi^k = \tilde{\psi}^k + \tfrac{T}{\tilde{\rho}^k}\mathcal{I}(\beta^1, \ldots, \beta^n), \tag{26}$$

where $\tilde{\psi}^k$ is the specific free energy of constituent k without the contribution due to the constraint (2). The indicator function $\mathcal{I}$ is multiplied byt the temperature in order to have the specific internal energy e^k which does not contain $\mathcal{I}$. The indicator function is a function $\mathbb{R}^n$ to $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ [8], [10],

$$\mathcal{I}(\beta^1, \ldots, \beta^n) = \begin{cases} 0 & \text{if } (\beta^1, \ldots, \beta^n) \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases} \tag{27}$$

The set $\mathcal{C} \subset \mathbb{R}^n$ is a convex set defined by the internal constraints (2)

$$\mathcal{C} = \left\{(\beta^1, \ldots, \beta^n) \in \mathbb{R}^k | \sum_k \beta^k = 1, \quad \beta^k \geq 0\right\}. \tag{28}$$

By means of the indicator function the free energy is forced to take only the values which comply with the constraints.

The essential variables to describe the dissipative behaviour of the material are the heat flow vectors $\vec{q}^k$, the rates of deformation $\mathcal{D}^k$, the rates of production of mass θ^k and the relative velocities $\vec{V}^{k*}$ of constituents. The corresponding function is the dissipation function per unite volume

$$\Phi = \Phi\left(\vec{q}^k, \mathcal{D}^k, \theta^k, \vec{V}^{k*}, T, \varepsilon^k, \beta^j\right), \tag{29}$$

which can also depend on the state variables.

The second principle (23) does not directly give any constitutive laws but rather is a condition to be satisfied in admissible evolutions of the process. There are two main approaches to derive them: the one, introduced by Germain [12] and applied in the previous work of Frémond and Mikkola [9], is based on the theory of pseudo-potentials; the other one, introduced by Ziegler [33] and applied in this work, makes use of the principle of maximal rate of entropy production.

To begin with, the stress tensors are devided into deviatoric and spherical parts

$$\sigma^k = \sigma'^k - p^k \mathrm{I}, \quad p^k = -\tfrac{1}{3}\mathrm{tr}\sigma^k. \tag{30}$$

On account of the conservation law of mass (13) and the definition (3), the material time derivative of volume fraction is represented by

$$\frac{d^k}{dt}\beta^k = \frac{\theta^k}{\bar{\rho}^k} - \beta^k \mathrm{tr}\mathcal{D}^k. \tag{31}$$

For the nonsmooth indicator function the concepts of subgradient and subdifferential are employed (see [10])

$$\left(\hat{B}^1, \ldots, \hat{B}^n\right) \in \partial\mathcal{I}\left(\beta^1, \ldots, \beta^n\right), \tag{32}$$

i.e. the subgradient defined by the set $\left(\hat{B}^1, \ldots, \hat{B}^n\right)$ is an element of the subdifferential $\partial\mathcal{I}$.

Making use of (6), (31) and (32) the sum of the material time derivates $\frac{d^k}{dt}\psi^k$ is developed with respect to the state variables into the form

$$\sum_k \rho^k \frac{d^k}{dt}\psi^k = \sum_k \left\{ \rho^k \frac{\partial \tilde{\psi}^k}{\partial T}\frac{d^k}{dt}T + \rho^k \frac{\partial \tilde{\psi}^k}{\partial \varepsilon'^k} : \mathcal{D}'^k + \left[\rho^k \frac{\partial \tilde{\psi}^k}{\partial (\mathrm{tr}\,\varepsilon^k)} - \beta^k \left(\hat{B}^k + \sum_j \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \right] \mathrm{tr}\,\mathcal{D}^k \right.$$
$$\left. + \left(\hat{B}^k + \sum_j \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \frac{\theta^k}{\bar{\rho}^k} + \sum_j \left[\left(\beta^k \hat{B}^j + \rho^k \frac{\partial \tilde{\psi}^k}{\partial \beta^j} \right) \nabla \beta^j - \left(\beta^j \hat{B}^k + \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \nabla \beta^k \right] \cdot \vec{V}^{k*} \right\} \tag{33}$$

It follows that the free energy ψ^k represents a potential: its partial derivatives with respect to the state variables are forces and its material time derivative are the products of those forces and the corresponding velocities. The dissipation function Φ plays the role of a potential as well. Introducing the definition

$$\Phi \stackrel{def}{=} \sum_k T\gamma_i^k \tag{34}$$

its development with respect to the essential dissipative variables according to Ziegler [33], [34] states

$$\Phi = \nu \frac{\partial \Phi}{\partial \vec{q}^k} \cdot \vec{q}^k + \nu \frac{\partial \Phi}{\partial \mathcal{D}'^k} : \mathcal{D}'^k + \nu \frac{\partial \Phi}{\partial \mathrm{tr}\,\mathcal{D}^k} \mathrm{tr}\,\mathcal{D}^k + \nu \frac{\partial \Phi}{\partial \theta^k}\theta^k + \nu \frac{\partial \Phi}{\partial \vec{V}^{k*}} \cdot \vec{V}^{k*} \geq 0, \tag{35}$$

where

$$\nu = \Phi \left(\frac{\partial \Phi}{\partial \vec{q}^k} \cdot \vec{q}^k + \frac{\partial \Phi}{\partial \mathcal{D}'^k} : \mathcal{D}'^k + \frac{\partial \Phi}{\partial \mathrm{tr}\,\mathcal{D}^k} \mathrm{tr}\,\mathcal{D}^k + \frac{\partial \Phi}{\partial \theta^k}\theta^k + \frac{\partial \Phi}{\partial \vec{V}^{k*}} \cdot \vec{V}^{k*} \right)^{-1}. \tag{36}$$

Finally, on account of (22), (33), (34) and (35) eq. (23) obtains the form

$$\sum_k \left\{ \left(\sigma'^k - \rho^k \frac{\partial \tilde{\psi}^k}{\partial \varepsilon'^k} - \nu \frac{\partial \Phi}{\partial \mathcal{D}'^k} \right) : \mathcal{D}'^k \right.$$
$$+ \left[p^k + \rho^k \frac{\partial \tilde{\psi}^k}{\partial (\mathrm{tr}\,\varepsilon^k)} + \nu \frac{\partial \Phi}{\partial \mathrm{tr}\,\mathcal{D}^k} - \beta^k \left(\hat{B}^k + \sum_j \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \right] \mathrm{tr}\,\mathcal{D}^k$$
$$- \rho^k \left(\frac{\partial \tilde{\psi}^k}{\partial T} + s^k \right) \frac{d^k}{dt}T - \left(\frac{\nabla T}{T} + \nu \frac{\partial \Phi}{\partial \vec{q}^k} \right) \cdot \vec{q}^k$$
$$- \left[\vec{m}^k + \sum_j \left(\beta^k \hat{B}^j + \rho^k \frac{\partial \tilde{\psi}^k}{\partial \beta^j} \right) \nabla \beta^j - \sum_j \left(\beta^j \hat{B}^k + \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \nabla \beta^k + \nu \frac{\partial \Phi}{\partial \vec{V}^{k*}} \right] \cdot \vec{V}^{k*}$$
$$\left. - \left(\tilde{\psi}^k + \frac{\hat{B}^k}{\bar{\rho}^k} + \sum_j \frac{\rho^j}{\bar{\rho}^k} \frac{\partial \tilde{\psi}^j}{\partial \beta^k} - \tfrac{1}{2}\vec{U}^k \cdot \vec{U}^k + \nu \frac{\partial \Phi}{\partial \theta^k} \right) \theta^k \right\} = 0 \tag{37}$$

It is required that (37) is satisfied for any real evolution, i.e. for any independent velocities $\vec{q}^k$, $\mathcal{D}'^k$, $\mathrm{tr}\mathcal{D}^k$, θ^k and $\vec{V}^{k*}$. This yields the relationships

$$\sigma'^k = \rho^k \frac{\partial \tilde{\psi}^k}{\partial \varepsilon'^k} + \nu \frac{\partial \Phi}{\partial \mathcal{D}'^k}, \tag{38}$$

$$p^k = -\rho^k \frac{\partial \tilde{\psi}^k}{\partial (\mathrm{tr}\varepsilon^k)} - \nu \frac{\partial \Phi}{\partial \mathrm{tr}\mathcal{D}^k} + p^k_{th} \tag{39}$$

$$p^k_{th} = \beta^k \left(\hat{B}^k + \sum_j \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \tag{40}$$

$$s^k = -\frac{\partial \tilde{\psi}^k}{\partial T} \tag{41}$$

$$-\frac{\nabla T}{T} = \nu \frac{\partial \Phi}{\partial \vec{q}^k} \tag{42}$$

$$-\left[\vec{m}^k + \sum_j \left(\beta^k \hat{B}^j + \rho^k \frac{\partial \tilde{\psi}^k}{\partial \beta^j} \right) \nabla \beta^j - \sum_j \left(\beta^j \hat{B}^k + \rho^j \frac{\partial \tilde{\psi}^j}{\partial \beta^k} \right) \nabla \beta^k \right] = \nu \frac{\partial \Phi}{\partial \vec{V}^{k*}}. \tag{43}$$

The last equation cannot be written conventionally since the velocities θ^k are not independent but related according to (14). This problem is clarified by an example. Consider a case, in which e.g. a phase change occurs between the constituents p and q so that $\theta^q = -\theta^p$. The corresponding constitutive equation reads then

$$\tilde{\psi}^q + \frac{p^q_{th}}{\rho^q} - \tfrac{1}{2}\vec{U}^q \cdot \vec{U}^q - \left(\tilde{\psi}^p + \frac{p^p_{th}}{\rho^p} - \tfrac{1}{2}\vec{U}^p \cdot \vec{U}^p \right) = \nu \frac{\partial \Phi}{\partial \theta^p}. \tag{44}$$

Considering the equations above some remarks are made:

- The third term on the right hand side in (39) indicates the reaction due to differences between potentials of constituents.
- The equation (43) can be regarded as a generalization of various diffusion laws as e.g. the Fick's law of diffusion and the law of Darcy.
- The equation (44) can be considered as a generalization of the Clausius-Clapeyron equation, respectively. What is interesting in this equation is that it contains also the contribution of kinetic energies, which cannot be neglected in defining the state, if constituents have significant velocities, e.g. high speed flows in aerodynamics.

aerobisessa prosessissa syntyy maitohappoa, jota vain aerobinen prosessi voi hajottaa. Koska väsymyksen tunne voimistuu veren maitohappopitoisuuden lisääntyessä, asettaa tämä rajan anaerobisen energian kehittämisprosessin kestolle. Ihminen ei myöskään kykene jatkuvasti ponnistelemaan maksimaalisella aerobisella teholla. Lihasrakenteesta ja harjoituksesta riippuen suurin mahdollinen jatkuva teho eli metabolinen teho on 40-80% maksimaalisesta aerobisesta tehosta.

Koska aerobinen teho tuotetaan hapen avulla, on se suorassa suhteessa käytettyyn happeen. Käytetty happimäärä eli hapenottokyky on taas helppo määrittää sisään- ja uloshengitysilman happipitoisuuksien erosta. Jotta vielä ruumiin koko otettaisiin huomioon ja saataisiin vertailukelpoisia arvoja, ilmoitetaan hapenottokyky aikayksikössä käytettynä happimääränä ruumiin massayksikköä kohden. Tavallisin yksikkö on $O_2 - ml/kg\ min$ Suurimmat hapenottokyvyn arvot ovat parhailla hiihtäjillä. Ne ovat suuruusluokkaa 100 *ml/kg min.* Tavallisilla ihmisillä hapenottokyky on vain noin puolet edellisestä eli 40-50 *g min.*

Ottamalla huomioon eri ravintoaineiden kehittämä energiamäärä käytettyä happimäärää kohti saadaan, jos hapenottokyky on 100 ml/kg min, seuraavat tehoarvot ko. urheilijalle
- hiilihydraateista 35,31 *W/kg* (teho/ruumiin massayksikkö)
- rasvoista 32,80 " "
- valkuaisaineista 31,05 " "

Koska arvojen erot ovat näinkin pienet, ei ole yllättävää, että maksimaalinen hapenottokyky korreloi vahvasti kilpailumenestyksen kanssa kestävyyslajeissa. Joskus onkin ehdotettu, että palkinnot jaettaisiin laboratoriokokeiden perusteella. Miten kävisi katsojien tällöin?

Energiantuoton matemaattinen malli ([8], [9] ja [10])

Edellä olleen varsin suppean energiantuottoa koskevan esityksen perusteella voidaan asia pelkistää seuraavasti:
- Ihmisellä on ennen urheilusuoritusta elimistössään latenttina tietty energiamäärä E_0, jota ei voida ylittää
- Urheilusuorituksen aikana hän ponnistellessaan voimalla $\mathbf{F} \leq \mathbf{F}_{max}$ kuluttaa energiavarastoaan teholla $\mathbf{F} \cdot \mathbf{v}$
- Urheilusuorituksen aikana aerobinen prosessi täydentää energiavarastoa korkeintaan maksimaalisella vakioteholla Σ.

Energiatasapaino antaa tehoyhtälön

$$\frac{dE}{dt} = \Sigma - \mathbf{F} \cdot \mathbf{v} \qquad (2)$$

Energian $E(t)$ on toteutettava alkuehto

$$E(0) = E_0 \ .$$

Integroimalla tehoyhtälö (2) voidaan energian rajoitusehto kirjoittaa muotoon

$$E_0 \geq E(t) \equiv E_0 + \Sigma t - \int_0^t \mathbf{F} \cdot \mathbf{v}\, dt \geq 0 \ . \tag{3}$$

NEUROMUSKULAARINEN TOIMINTA

Voima ([1] ja [14])

Lihassäie jännittyy supistuessaan ja synnyttää voiman F, joka on jännityksen σ ja lihassäikeen poikkipinta-alan A tulo

$$F = \sigma A \ .$$

Jos yksinkertaisuuden vuoksi pidämme lihassäiettä tai jopa koko lihasta prismana, jonka pituus L, poikkipinta-ala A ja tiheys ρ ovat vakioita, saamme lihaksen massaksi m

$$m = \rho L A$$

Eliminoidaan näistä A, jolloin voimalle F seuraa kaava

$$F = \frac{\sigma}{\rho}\frac{m}{L} \tag{4}$$

Koska lihaksen pituus ei täysikasvuisella enää muutu, voidaan voimaa lisätä vain
- parantamalla harjoituksella lihaksen kvaliteettia σ/ρ
- lisäämällä lihasmassaa poikkipinta-alaa kasvattamalla.

Kaavassa (4) käytetään usein massan m tilalla lihaksen painoa $G = mg$, jolloin kaava kuuluu

$$F = \frac{\sigma}{\rho g}\frac{G}{L} \ . \tag{5}$$

Useissa tapauksissa on helppo löytää sopiva referenssipituus R, jonka avulla voidaan määritellä pituusparametri $p = L/R$. Tätä käyttäen voidaan kaavan (5) perusteella määritellä dimensioton lihasvoiman kvaliteettifunktio k

$$k = \frac{F}{G/p} \qquad \left(= \frac{\sigma}{\rho g R} \right) . \tag{6}$$

Kaava (6) johdettiin yksittäiselle lihakselle tai oikeastaan lihassäikeelle. Sitä voidaan kuitenkin soveltaa useampien lihasten yhteisvaikutukseen eli urheilijaan kokonaisuutena. Kvaliteetin k avulla voidaan keskenään vertailla eri kokoisten ja painoisten henkilöiden lihasvoimaa samassa urheilusuorituksessa.

Saman urheilulajin harrastajille pätee kokemuksen mukaan ([11] ja [14]) varsin tarkasti yksinkertaistava otaksuma, että erikokoiset harrastajat ovat yhdenmuotoiset. Tällöin saadaan seuraavat verrannollisuudet

$$G \propto p^3 \qquad \text{eli} \qquad p \propto G^{1/3}$$

$$F \propto p^2 \qquad \text{eli} \qquad F \propto G^{2/3} .$$

Voima ja nopeus ([1])

Jokapäiväisissä askareissa ihminen voi tietyissä rajoissa vapaasti valita käyttämänsä voiman ja liikkeidensä nopeuden. Kun urheilusuorituksissa toimitaan suorituskyvyn äärirajoilla, ei enää voida vapaasti valita voimaa ja nopeutta vaan ne kytkeytyvät toisiinsa.

Jotta liike olisi mahdollinen täytyy lihaksen voittaa ns. sisäinen kitkavoima a. Jos F on kitkavoiman ylittävä osa lihasvoimaa ja lihaksen supistumisnopeus on v, kehittää lihas tehon

$$P = (F + a)v \ .$$

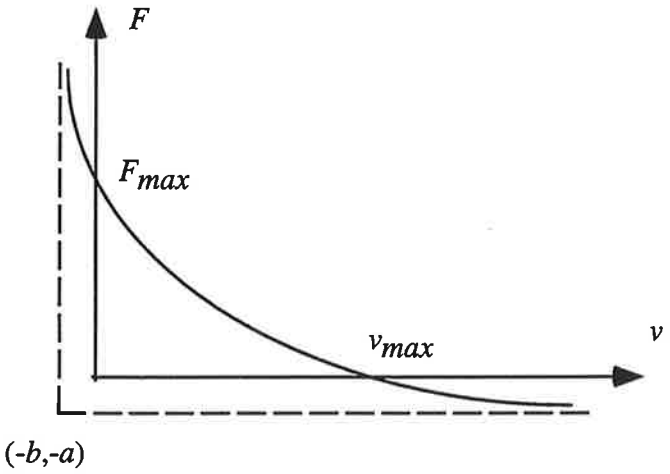
Jos $F_{\max}$ on lihaksen isometrisesti (ilman liikettä) kehittämä maksimivoima, on isotoonisissa (liike mukana) kokeissa todettu seuraavan tehoyhtälön olevan voimassa

$$P = b\left(F_{\max} - F\right) \ ,$$

jossa b on kokeellinen vakio. Yhdistämällä nämä kaksi tehoyhtälöä saadaan

$$(F + a)(v + b) = \left(F_{\max} + a\right)b \equiv P_{\max} \ , \tag{7}$$

jonka mukaan lihaksen maksimaalinen (kokonais)teho on vakio $P_{\max}$.



Kuva 1. Lihasvoiman ja liikenopeuden välinen skemaattinen yhteys

Yhtälö (7) esittää hyperbeliä, joka leikkaa F-akselin pisteessä $F_{\max}$ ja v-akselin pisteessä $v_{\max} = F_{\max} b/a$ ja jolla on asymptootit $F = -a$ ja $v = -b$.

Ratkaistaan yhtälö (7) vielä voiman suhteen

$$F = \frac{P_{\max}}{v+b} - a \quad . \tag{8}$$

Yhtälö (8) on johdettu yksittäiselle lihakselle, mutta sitä voitaneen soveltaa laajemminkin. Todetaan vielä, että yhdenmuotoisuusotaksuma johtaa verrannollisuuksiin

$$P_{\max} \propto L^3 \quad , \quad a \propto L^2 \quad \text{ja} \quad b \propto L \quad .$$

PSYKOLOGISET TEKIJÄT

Kilpaurheilun motivaatiota on pyritty kaikin tavoin parantamaan. Taloudellisten etujen lisääminen kuten hyväpalkkaiset toimet, erilaiset stipendit ja runsaat palkkiorahat sekä muut etuudet vaikuttavat omalta osaltaan. Tärkein motivaatio ja uran alussa ainoa on kunnianhimo. Sitä voidaan usein kohottaa lisäämällä merkittävien kilpailujen voiton arvostusta ja voittajalle osoitettua huomiota. Myös suggestiota ja hypnoosia voidaan käyttää apuna.

URHEILUSUORITUSTEN TEKNIIKKA

Paitsi että valmennusmenetelmät ovat kehittyneet, on myös itse urheilusuoritusten tekniikka kehittynyt. Tästä muutamia esimerkkejä:

- painonnostossa kyykkytyyli on paljolti sivuuttanut saksityylin
- korkeushypyssä ei enää harrasteta muuta kuin ns. Floppaus-tyyliä (yli selkä rimaan päin)

- heittolajeissa (kuulantyönnössä) pyörähdystyyli on voittamassa alaa; keihäänheitossa se kiellettiin.
- mäkihypyssä ns. V-tyyli on syrjäyttänyt täysin kaikki muut hyppytyylit [20].

KILPAILUTAKTIIKKA JA SUORITUKSEN OPTIMOINTI

Kilpailussa noudatettavaa taktiikkaa on tutkittu niin psykologiselta kuin myös fyysiseltä kannalta. Jos urheilusuoritukselle voidaan rakentaa toimiva matemaattinen malli, niin sen optimoiminen antamalla tiettyjen parametrien eli ohjailusuureiden muuntua johtaa luonnostaan parhaaseen suoritukseen niin taktiselta kuin fyysiseltäkin kannalta.
Tästä esimerkkejä ovat

- oikea voimien ja vauhdin jako hiihdossa, juoksussa [9] ja uinnissa [12]
- sopivat aloituspainot ja lisäykset painonnostossa [13]
- edullisimman lähtökulman valinta keihäänheitossa [21] ja kuulantyönnössä [28].

URHEILUVÄLINEIDEN KEHITTÄMINEN

On urheilusuorituksia, joissa välineiden kehittäminen on ollut tuloksen parantumisen kannalta ratkaisevaa. Tästä muutamia esimerkkejä:

Keihäänheitossa ([11] ja [21]) on itse keihäs tehty liitävämmäksi lisäämällä sen paksuutta ja muuttamalla sen painon jakaumaa. Koska tällöin painopisteen ja aerodynaamisten voimien vaikutuspisteen välistä etäisyyttä on lyhennetty, on keihään lennon aerodynaaminen vakavuus huonontunut. Tästä taas on seurannut, että keihäs putoaa entistä useammin lappeelleen.

Moukarinheitossa on kokeiltu teräspallon painopisteen siirtämistä geometrisestä keskiöstä kiinnityspisteen vastakkaiselle puolelle päin. Näin on saatu moukarin painopisteen pyörähdyskaaren säde ja kulmaliikemäärä suuremmiksi.

Kiekonheitossa on kiekon muotoa paranneltu liito-ominaisuuksien lisäämiseksi.

Seiväshypyssä tulokset paranivat oleellisesti, kun seipään materiaalia vaihdettiin: bambu →teräs →lasikuitu →hiilikuituvahviste. Samalla on lisääntynyt seipään kyky absorboida vauhdista saatavaa liike-energiaa ja muuttaa se hyppääjän potentiaalienergiaksi [3].
Hiihdossa ja mäenlaskussa on suksien muoto kehittynyt ja materiaali muuttunut puusta erilaisiin kuituvahvisteisiin muoveihin. Tällöin on lumen hankauskitka pienentynyt ja pitoa ponnistuksessa on voiteilla parannettava ([3] ja [19]). Oman lisänsä tuloksiin antavat kevyet pienen ilmanvastuksen omaavat sileäkankaiset urheiluvaatteet sekä kengät ja sauvat. Mäkisuksien aerodynaamiset ominaisuudet ovat myös muotoilulla parantuneet. Laskijan ruumiin koolla on myös vaikutusta tulokseen ([17] ja [18]).

Kilpailu urheiluvälineiden kehittämisessä on ollut kovaa ja alati lisääntyvää. Monessa lajissa on täytynyt tarkistaa sääntöja, jotta tilanne pysyisi hallinnassa.

ULKOISTEN OLOSUHTEIDEN VAIKUTUS URHEILUSUORITUKSIIN

On selvää, että ulkoiset olosuhteet vaikuttavat urheilutuloksiin. Yleensä pitää pyrkiä harjoittelemaan kilpailutilannetta vastaavissa olosuhteissa riittävän kauan ennen kilpailuja (ns. akklimatisoitumisefekti). Seuraavassa esitetään eräitä pääpiirteitä ulkoisten olosuhteiden vaikutuksesta:

Painovoiman pienentyminen lisää heittojen pituutta sekä hyppyjen pituutta tai korkeutta.
Ilman tiheyden pienentyminen pienentää ilmanvastusta ja parantaa anaerobisten suoritusten tulosta. Tämä koskee erityisesti pikajuoksua ja pituushyppyä, miksei myös kolmiloikkaa, kuulantyöntöä ja moukarinheittoa. Kiekon- ja keihäänheitossa pienenee samassa suhteessa myös ilman nosto, joten välineen liitokyky (nosto/vastus) säilyy ennallaan. Aerobisissa suorituksissa hapensaannin pienentyminen vaikuttaa voimakkaammin kuin vastuksen pienentyminen. Kestävyyslajeissa siis tulokset huonontuvat.
Koska ilmanpaineen alentuminen lämpötilan säilyessä vakiona pienentää ilman tiheyttä, pätee edellä sanottu suoraan ilmanpaineen vaihteluihin esimerkiksi tapauksissa, jolloin kilpailuja suoritetaan huomattavan korkealla merenpinnasta.
Ilmankosteus lienee urheilijoiden kannalta edullisin, kun suhteellinen kosteus on 50-60%. Ilmankosteudella on merkitystä ennenkaikkea kestävyyslajeissa, joissa jäähdytysjärjestelmä ja hapenottokyky joutuvat kovalle rasitukselle. Suuret poikkeamat edullisimmasta kosteudesta huonontavat tuloksia kestävyyslajeissa.
Ilman lämpötilalla on myös urheilijoiden kannalta optimiarvonsa [22]. Anaerobisissa suorituksissa suotuisin lämpötila lienee

- suomalaisille urheilijoille 22-24 °C
- Etelä-Euroopan urheilijoille 25-27 °C
- tropiikin maiden urheilijoille 28-30 °C.

Aerobisissa suorituksissa optimilämpötila lienee muutamia asteita alempi. Poikkeamat molempiin suuntiin huonontavat tulosta. Tuuli vaikuttaa pääasiassa mekaanisesti. Vastatuuli lisää ja myötätuuli pienentää ilmanvastusta. Lisäksi tuuli edistää ruumiin lämmönhukkaa ja vaikuttaa siten kuin lämpötilan lasku.

SOVELLUTUSESIMERKKEJÄ MATEMAATTISISTA MALLEISTA

Edellä olleita perusmalleja voidaan soveltaa hyvinkin erityyppisiin urheilusuorituksiin [23]. Yksinkertainen malli riittää usein kuvaamaan urheilusuoritusta yllättävän hyvin. Seuraavassa muutamia esimerkkejä.

KILPAJUOKSU ([8], [9] ja [23]...[27])

Prof. Keller on julkaisuissaan ([8], [9] ja [10]) käsitellyt kilpajuoksua ja ratkaissut periaatteessa optimaalisen vauhdinjaon eripituisilla juoksumatkoilla. Hän on käyttänyt liikeyhtälössä (1) lineaarista vastusmallia. Analyysissään Keller totesi, että jos juostava matka on ns. kriittistä matkaa pidempi, on optimaalinen juoksustrategia seuraava:

- Juostavan matkan pituudesta riippuen heti lähdön jälkeen n. 1-2 s kiihdytysvaiheen aikana vauhti kiihdytetään maksimivoimalla matkanopeuteen ,
- joka valitaan niin, että energiavarasto on vakionopeusvaiheen jälkeen tyhjä eli $e = 0$ juostavan matkan pituudesta riippuen 1-5 s ennen maaliintuloa.
- Viimeiset metrit eli hidastusvaihe juostaan vakioteholla $fv = \sigma$ siten, että $e \equiv 0$. Tällöin vauhti hidastuu melkoisesti kohden metabolista nopeutta.

Tutkimuksissa [25] ja [26] on taas käytetty neliöllistä vastusmallia. Vastusmalli ei vaikuta nopeimpaan loppuaikaan tähtäävään juoksustrategiaan eli vauhdinjaon optimointiin, ainoastaan numeeriset yksityiskohdat hieman muuttuvat. Lähtöyhtälöiksi voidaan ottaa liikeyhtälö (1) ja tehoyhtälö (2) massalla jaettuna. Merkitään

$$f(t) = F(t)/m \le f_{\max} \quad , \quad e_0 \ge e(t) = E(t)/m \ge 0 \quad , \quad \sigma = \Sigma/m$$

ja oletetaan vastuksen olevan muotoa

$$\frac{D}{m} = k_R v^2 + k_D (v - v_t)^2 \ ,$$

missä k_R on jalkojen edestakaisesta liikuttamisesta (rotaatiosta) tuleva vastusosuus ja k_D on ilmanvastuksen osuus. Myötätuuli on positiivinen eli $v_t > 0$. Lähteessä [26] esitetään kertoimille arviot $k_R = 0.0464\,m^{-1}$ ja $k_D = 0.0033\,m^{-1}$, joten jalkojen rotaatiovastus on tyynellä säällä ilmanvastukseen nähden yli kymmenkertainen. Liikeyhtälö (1) kuuluu nyt

$$\frac{dv}{dt} = f - k_R v^2 - k_D (v - v_t)^2 \quad , \quad v(0) = 0 \tag{9}$$

ja energiayhtälö (2)

$$\frac{de}{dt} = \sigma - fv \quad , \quad e(0) \equiv e_0 \ge e(t) \ge 0 \ . \tag{10}$$

Liikeyhtälön (9) integrointi (vertaa [27]) onnistuu analyyttisesti vain jos f on vakio, joten on parasta turvautua numeeriseen integrointiin.

Tyynellä säällä on $v_t \equiv 0$, jolloin liikeyhtälö yksinkertaistuu muotoon ($k = k_R + k_D$) muotoon

$$\frac{dv}{dt} + kv^2 = f \quad . \tag{11}$$

Liikeyhtälö (11) voidaan integroida analyyttisesti, kun $f = f_{\max}$ on vakio, jolloin saadaan

$$t = \frac{1}{kV} ar\tanh\left(\frac{v}{V}\right) \quad , \tag{12}$$

$$x = -\frac{1}{2k}\ln\left[1-\left(\frac{v}{V}\right)^2\right] \equiv \frac{1}{k}\ln\cosh(kVt) \cong Vt - \frac{1}{k}\ln 2 \quad , \tag{13}$$

missä V on määritelty yhtälössä (15). Yhtälöistä (10) ja (11) seuraa eliminoimalla voima f (massaa kohti) uusi tehoyhtälö

$$\frac{d}{dt}\left(\tfrac{1}{2}v^2 + e\right) = \sigma - kv^3 \quad . \tag{14}$$

Suurin mahdollinen nopeus

$$V = \left(f_{\max}/k\right)^{1/2} \tag{15}$$

suoralla radalla saavutetaan raja-arvona yhtälöstä (11), kun $f = f_{\max}$. Suurin mahdollinen jatkuva nopeus eli metabolinen nopeus

$$U = \left(\sigma/k\right)^{1/3} \tag{16}$$

saadaan raja-arvona yhtälöstä (14). Edellä kuvattu juoksumalli sisältää neljä parametria:

k yhdistetty vastus
V tai $f_{\max}$ äärimmäinen lyhytaikainen suorituskyky
U tai σ äärimmäinen jatkuva suorituskyky
e_0 perusenergiavarasto.

Pisin pikajuoksumatka eli kriittinen matka

Tarkastellaan seuraavaa kysymystä: kuinka pitkän ajan T_c eli miten pitkän matkan D_c juoksija kykenee juoksemaan maksimaalisella voimalla ennenkuin energiavarasto on tyhjä? Oletetaan yksinkertaisuuden vuoksi, että juoksu tapahtuu suoralla radalla.

Integroidaan energiayhtälö (10), jolloin kaavojen (13)-(16) avulla voidaan johtaa likimääräinen mutta hyvin tarkka energiayhtälö probleeman ratkaisemiseksi:

$$e(T_c) = e_0 + \sigma T_c - f_{\max}\frac{1}{k}\ln\cosh(kVT_c) \approx e_0 + V^2\ln 2 - k\left(V^3 - U^3\right)T_c = 0 \ .$$

Tästä seuraa kriittinen aika

$$T_c = \frac{e_0 + V^2\ln 2}{k\left(V^3 - U^3\right)} \tag{17}$$

ja vastaava kriittinen matka

$$D_c = \frac{e_0 V + U^3\ln 2}{k\left(V^3 - U^3\right)} \ . \tag{18}$$

Laskut osoittavat, että 25 s < T_c < 30 s ja 250 m < D_c < 300 m.

Juoksu kaarteisella radalla

Vain sadan metrin juoksussa koko matka juostaan suoraan. Muita matkoja pituudeltaan D juostaan niin, että maali on aina viimeisen 100 m suoran päässä. Tällöin matkan pituudesta riippuen lähdetään joko kaarteeseen tai suoralle. Koska radan pituus on 400 m ja kaarteita on kaksi, on kaikilla 200 m:llä jaollisilla matkoilla lähtö kaarteeseen. Toisin sanoen, jos D/100m on parillinen kokonaisluku, tapahtuu lähtö kaarteesta. Jos taas D/100m on pariton kokonaisluku, tapahtuu lähtö suoralta.Tavallisimmista kilpailumatkoista vain 300 m ja 1500 m lähtevät suoralle. Mailimatkoja ei tässä tarkastella.

Jos radat numeroidaan sisäradalta alkaen, on normaalikokoisen urheilukentän radan n kaarevuussäde $R(n)$ metreinä

$$R = R(n) = 100/\pi + 1.22(n-1) \ .$$

Lähteen [24] mukaan kaarrejuoksussa tulee liikeyhtälössä (11) voima f korvata voimalla

$$f_t = \sqrt{f^2 - \left(v^2/R\right)^2} \ ,$$

joka ottaa huomioon kaarrejuoksussa syntyvän keskipakoisvoiman kompensoimisen lihasvoimalla. Tarkastelu kannattaa aloittaa 400 m juoksulla, jossa matka voidaan jakaa kuuteen eri vaiheeseen. Tilayhtälöt ovat optimivauhdinjaon mukaan seuraavat:

1.1 Kiihdytysvaihe ensimmäisessä kaarteessa ($0 \le t \le t_0$)

$$\frac{dv}{dt} + kv^2 = \sqrt{f_{\max}^2 - \left(\frac{v^2}{R}\right)^2} \quad , \qquad v(0) = 0 \tag{19a}$$

tai

$$\frac{1}{2}\frac{dv^2}{dx} + kv^2 = \sqrt{f_{\max}^2 - \left(\frac{v^2}{R}\right)^2} \quad , \qquad v(0) = 0 \tag{19b}$$

$$\frac{dx}{dt} = v \quad , \qquad x(0) = 0 \tag{20}$$

$$e(t) = e_o + \sigma t - f_{\max}\, x(t) \quad , \qquad e(0) = e_0 \quad . \tag{21}$$

Liikeyhtälöä (19a) ei voida analyyttisest integroida, vaan täytyy tyytyä numeeriseen ratkaisuun. Liikeyhtälön (19b) alkuehtoihin $v(0) = 0$ ja $x(0) = 0$ sovitettu ratkaisu – matka nopeuden funktiona – saadaan muotoon

$$x = \frac{1}{2k\left[1 + (1/kR)^2\right]}\left\{\frac{1}{kR}\arcsin\left(\frac{(v/V)^2}{kR}\right) - \ln\left[\sqrt{1 - \left(\frac{(v/V)^2}{kR}\right)^2} - \left(\frac{v}{V}\right)^2\right]\right\} \quad . \tag{22}$$

Kiihdytysmatkan pituus x_0 saadaan matkanopeusehdosta (johdetaan myöhemmin)

$$v = u = V\sqrt{s} \quad , \tag{23}$$

mikä antaa

$$x_0(s) = \frac{1}{2k\left[1 + (1/kR)^2\right]}\left\{\frac{1}{kR}\arcsin\left(\frac{s}{kR}\right) - \ln\left[\sqrt{1 - \left(\frac{s}{kR}\right)^2} - s\right]\right\} \quad . \tag{24}$$

Tässä s on parametri, joka kuvaa voiman käyttöä suoralla. Myöhemmin esiintyvä r on vastaava parametri kaarteessa.

1.2 Vakionopeusvaihe $u = v(t_0)$ ensimmäisessä kaarteessa ($t_0 \le t \le t_1$)

Tällöin on voimassa $dv/dt = 0$, $u = v(t_0)$, $r = f_{kaarre}/f_{\max}$ = vakio, jolloin

$$ku^2 = \sqrt{r^2 f_{\max}^2 - \left(\frac{u^2}{R}\right)^2} \tag{25}$$

$$x(t) = x(t_0) + (t - t_0)u \tag{26}$$

$$e(t) = e(t_0) + (t - t_0)\sigma - \left[x(t) - x(t_0)\right] r f_{\max} \quad . \tag{27}$$

2. Vakionopeus u ensimmäisellä suoralla $(t_1 \le t \le t_2)$

Tällöin on voimassa $dv/dt = 0$, $s = f_{suora}/f_{\max}$ = vakio, jolloin

$$ku^2 = s f_{\max} \tag{28}$$

$$x(t) = x(t_1) + (t - t_1)u \tag{29}$$

$$e(t) = e(t_1) + (t - t_1)\sigma - \left[x(t) - x(t_1)\right] s f_{\max} \quad . \tag{30}$$

3. Vakionopeus u toisessa kaarteessa $(t_2 \le t \le t_3)$

Nyt on $dv/dt = 0$, $r = f_{kaarre}/f_{\max}$ = vakio, jolloin

$$ku^2 = \sqrt{r^2 f_{\max}^2 - \left(\frac{u^2}{R}\right)^2} \tag{31}$$

$$x(t) = x(t_2) + (t - t_2)u \tag{32}$$

$$e(t) = e(t_2) + (t - t_2)\sigma - \left[x(t) - x(t_2)\right] r f_{\max} \quad . \tag{33}$$

4.1 Vakionopeus u toisella suoralla $(t_3 \le t \le t_4)$

Tällöin on taas $dv/dt = 0$, $s = f_{suora}/f_{\max}$ = vakio, jolloin

$$ku^2 = s f_{\max} \tag{34}$$

$$x(t) = x(t_3) + (t - t_3)u \tag{35}$$

$$e(t) = e(t_3) + (t - t_3)\sigma - \left[x(t) - x(t_3)\right] s f_{\max} \quad . \tag{36}$$

Yhtälöistä (28) tai (34) saadaan yhtälön (15) avulla matkanopeus

$$u = V\sqrt{s} \quad . \tag{37}$$

Voima $r f_{\max}$ kaarteessa verrattuna voimaan $s f_{\max}$ suoralla eli suhde r/s, jotta nopeus pysyisi vakiona, saadaan yhtälöistä (25) tai (31) yhtälön (37) avulla. Sen suuruus vaihtelee

$$\underset{n=1}{1.120} \geq r/s = \sqrt{1+(1/kR)^2} \geq \underset{n=8}{1.076} \tag{38}$$

Tätä (tai sen neliöjuurta) voidaan pitää kaarteen haitta- tai rasituskertoimena suoraan rataan verrattuna.

Aika ja matka ennen hidastusvaihetta

Juoksuradan $D = 400$ m geometriasta voidaan päätellä matkat

$$x_1 \quad = x(t_1) \quad = \tfrac{1}{2}D - R(n)\pi \tag{39}$$

$$x_2 - x_1 = x(t_2) - x(t_1) = \tfrac{1}{4}D \tag{40}$$

$$x_3 - x_2 = x(t_3) - x(t_2) = R(n)\pi \tag{41}$$

$$x_3 \quad = x(t_3) \quad = \tfrac{3}{4}D \quad . \tag{42}$$

Laskemalla yhteen energiayhtälöt (21), (27), (30), (33) ja (36) hidastumisvaiheen alussa, kun $t = t_4$, saadaan energia

$$e(t_4) = e_o + \sigma t_4 - f_{\max}\left[(1-r)x_0 + sx_4 + (r-s)\tfrac{1}{2}D\right] = 0 \quad . \tag{43}$$

Samalla tavalla laskemalla yhteen matkayhtälöt (26), (29), (32) ja (35) saadaan matka

$$x_4 = x(t_4) = x_0 + (t_4 - t_0)V\sqrt{s} \quad . \tag{44}$$

Koska ehto hidastusvaiheen alkamiselle on $e(t_4) = 0$, seuraa yhtälöistä (43) ja (44) aika

$$t_4(s) = \frac{e_o + kV^2\left[Vs^{3/2}t_0 - (1-r+s)x_0 - (r-s)\frac{1}{2}D\right]}{k\left[\left(V\sqrt{s}\right)^3 - U^3\right]} \quad . \tag{45}$$

Vaikka teoria on johdettu 400 m juoksulle, pätee yhtälö (45) sisärataa pitkin kaikille pidemmille matkoille, jotka ovat 200 m:lla jaollisia. Lähtö on kaarteeseen ja x_0 saadaan yhtälöstä (24), mutta t_0 täytyy ratkaista yhtälöstä (19a) numeerisesti.

Jos suure $D/100$ m on pariton kokonaisluku, tulee yhtälö (45) korvata yhtälöllä

$$t_4(s) = \frac{e_o + kV^2\left[Vs^{3/2}t_0 - x_0 - (r-s)\frac{1}{2}(D - 100\text{ m})\right]}{k\left[\left(V\sqrt{s}\right)^3 - U^3\right]} \tag{46}$$

ja lähtö on suoralle, jolloin t_0 ja x_0 saadaan yhtälöistä (12) ja (13)

$$t_0(s) = \frac{1}{kV}\,ar\tanh\sqrt{s} \tag{47}$$

$$x_0(s) = -\frac{1}{2k}\ln(1-s)\ . \tag{48}$$

Kaikissa tapauksissa yhtälö (44) antaa hidastusvaiheen alkuun juostun matkan $x_4(s)$.

4.2 Hidastusvaihe suoralla ($t_4 \le t \le T$)

Koska $e(t) \equiv 0$ hidastusvaiheessa, tehoyhtälöstä (14) seuraa kaksi samanarvoista liikeyhtälöä

$$\frac{1}{3}\frac{dv^3}{dx} + kv^3 = \sigma \tag{49a}$$

$$\frac{1}{2}\frac{dv^2}{dt} + kv^3 = \sigma \tag{49b}$$

Liikeyhtälöstä (49a) voidaan alkuehtoon $v(x_4) = u = V\sqrt{s}$ sovittaen integroida nopeus matkan funktiona. Tulos on

$$v(x,s) = U\left\{1 + \left[\left(\frac{V}{U}\right)^3 s^{3/2} - 1\right]e^{-3k[x - x_4(s)]}\right\}^{1/3}\ . \tag{50}$$

Liikeyhtälön (49b) integraalifunktio, aika nopeuden funktiona, on

$$t(v) = \frac{1}{kU}\left\{\frac{1}{6}\ln\frac{v^2 + vU + U^2}{(v-U)^2} - \frac{1}{\sqrt{3}}\arctan\frac{2v+U}{U\sqrt{3}}\right\}\ . \tag{51}$$

Matkan $x - x_4(s)$ juoksemiseen käytetty aika on nopeuksien avulla lausuttuna

$$t - t_4(s) = t[v(x,s)] - t\left[V\sqrt{s}\right]\ . \tag{52}$$

Voimaparametrin s määrittäminen ja nopein loppuaika T

Kun juoksija ajan T kuluttua on maalissa ja juossut matkan D, tulee olla

$$D = x(T) \ ,$$

jolloin vastaava loppuaika voimaparametrin s funktiona seuraa kaavasta (52)

$$T(s) = t_4(s) + t[v(D,s)] - t\left[V\sqrt{s}\right] \ . \tag{53}$$

Loppuaika tulee sitten minimoida suureen s suhteen. Hyvänä alkuarvauksena voimaparametrille voidaan pitää (katso (44)) yhtälön

$$x_4(S) = x_0(S) + \left[t_4(S) - t_0(S)\right]V\sqrt{S} = D \tag{54}$$

juurta S eli että energiavarasto olisi tyhjä juuri maaliviivalla. Optimaalisen juoksun loppuaika T sekä vastaava arvo s optimaaliselle voimankäytölle $sf_{\max}$ matkajuoksun suoralla osalla saadaan ehdosta

$$T = \min_{s>S} T(s) \ . \tag{55}$$

Kaarteissa saman vauhdin ylläpitämiseksi tarvitaan kaavan (38) mukaan hieman suurempaa voimankäyttöä $rf_{\max}$. Mitä suorempi rata on, sitä vähemmän lisävoimaa tarvitaan kaarteissa. Ulkorata olisi siis nopein, mutta sieltä ei ole helppoa seurata muiden juoksijoiden matkantekoa.

Maailmanennätystilastoon sovittaminen

Sovitettaessa mallia maailmanennätystilastoon oletetaan, että kriittistä matkaa lyhyemmät matkat juostaan täydellä voimalla ja sitä pidemmät optimivauhdin jaolla. Parametrit k ja V määräävät ajan kriittistä matkaa lyhyemmillä matkoilla. Jos maailmanennätystilasto välillä 50yd (45.72 m) ja 220 yd (201.168 m) puhdistetaan tuulen ja kaarteen vaikutuksesta ja käytetään pienimmän neliösumman menetelmää, saadaan parametreille arvot

$$k = 0.0621609 \text{ m}^{-1}$$
$$V = 11.1587 \text{ m/s}$$
$$f_{\max} = 7.74002 \text{ N/kg} \ .$$

Matkojen 400 - 10000 m maailmanennätyksistä on samoin pienimmän neliösumman menetelmällä saatu

$e_0 = 1740.0$ J/kg
$U = 6.33$ m/s
$\sigma = 15.7663$ W/kg .

Kaavoista (17) ja (18) seuraa kriittiselle matkalle

$T_c = 25.89$ s
$D_c = 277.50$ m .

Kellerin laskema ([8]-[10]) kriittinen aika oli 27.36 s ja matka 291 m. Vaikka eroa on jonkin verran, se ei häiritse tilastoon sovittamista.

Taulukot 1 ja 2 on laskettu edellä esitetyn kvadraattiseen vastuslakiin perustuvan teorian perusteella. 100 m juoksussa on otaksuttu 2 m/s myötätuulta, 200 m matkalla on käytetty 3. rataa, jota yleisesti pidetään mieluisimpana ja 400 m matkalla on käytetty 1. rataa, joka on hitain. Jos juostava matka on pidempi kuin 400 m, ajatellaan koko matka juostavan sisärataa eli 1. rataa.

TAULUKKO 1. Maailmanennätysajat eri matkoilla verrattuna teorian mukaan laskettuihin optimaalisiin aikoihin

Matka D (m)	T-ennätys (min:sek)	T-teoria (min:sek)	s-arvo (%)	ε_i-virhe (%)
100	9.84	9.87	100.0	-0.334462
200	19.32	19.34	100.0	-0.113972
400	43.29	42.58	73.8653	+1.66373
800	1:41.73	1:41.31	50.7773	+0.416961
1000	2:12.18	2:12.07	46.5051	+0.0800151
1500	3:27.27	3:30.21	41.1318	-1.39687
2000	4:50.81	4:49.87	38.3833	+0.324153
3000	7:25.11	7:29.74	35.8217	-1.02858
5000	12:44.40	12:50.89	33.8287	-0.841429
10000	25:38.09	26:15.53	32.3720	+1.43192

Neliösummana on käytetty virhealkioiden

$$\varepsilon_i = \left(\frac{T_{ennätys} - T_{teoria}}{T_{teoria}} \right)_i \cdot 100\%$$

summaa $\varepsilon^2 = \frac{1}{N-1} \sum_{i=1}^{10} \varepsilon_i^2$.

Virhejakauman keskiarvo on −0.022823 ja hajonta 0.950326 sekä neliöllinen keskiarvo 0.789985 ja vastaava hajonta 0.833517.

TAULUKKO 2. Juoksun eri vaiheiden kesto, pituus ja nopeus

Matka	Kiihdytysvaihe		Matkavaihe			Hidastumisvaihe		
D	t_0	x_0	$t_4 - t_0$	$x_4 - x_0$	u	$T - t_4$	$D - x_4$	$v(T)$
(m)	(s)	(m)	(s)	(m)	(m/s)	(s)	(m)	(m/s)
400	1.989	11.911	39.670	380.446	9.59	0.922	7.643	7.40
800	1.306	5.833	98.734	785.078	7.95	1.268	9.089	6.69
1000	1.214	5.121	129.462	985.153	7.61	1.399	9.726	6.57
1500	1.096	4.262	207.468	1484.750	7.16	1.642	10.993	6.45
2000	1.051	3.935	287.008	1984.170	6.91	1.811	11.899	6.40
3000	1.001	3.598	446.464	2981.760	6.68	2.270	14.646	6.35
5000	0.963	3.346	767.442	4980.820	6.49	2.481	15.834	6.34
10000	0.935	3.167	1569.480	9964.470	6.35	5.110	32.365	6.33

Nopeimpaan aikaan tähtäävä juoksustrategia edellyttää suurimmalla osalla matkaa tasaista vauhtia u, jonka ylläpitämiseen vaadittava voima on suoralla s ja kaarteessa r =1.12 s maksimaalisesta voimasta (Taulukko 1). Tämä tiedetään jo ennestään kokemuksesta. Yllättävämpää sen sijaan voi olla, että lopussa ei otetakaan nopeaa kiriä, kuten useimmiten nähdään , vaan tapahtuu maitohapoille meno eli kangistuminen tai jopa "sammuminen". Loppukiri merkitsee, että matkalla ei ole haluttu tai tarvinnut käyttää koko kapasiteettia. Taulukosta 2 nähdään, että hidastumisvaihe on lyhyt koko matkaan verrattuna. Jos se pääsee pienenkin virhearvioinnin takia alkamaan liian aikaisin, voi koko juoksu mennä piloille. Laskennallisesti voidaan tutkia, millaiseksi loppuaika muodostuu, jos energiavarasto tyhjenee vasta maaliviivalla: 400 m:llä on aika 0.11 s huonopi, 5000 m:llä 0.01s huonompi ja 10000 m:llä ei eroa enää huomaa.

KILPAUINTI ([12])

Liike- ja energiayhtälöt ovat samanlaiset kuin juoksussa, vastuskerroin on kuitenkin eri. Juoksun kiihdytysvaiheen tilalle tulee uinnissa joko veden yläpuolella olevalta lähtöjalustalta alkava ilmalento ja sitä seuraava liukuvaihe vedessä tai vedessä tapahtuvassa lähdössä vain ponnistusta seuraava liukuvaihe. Lisäksi voidaan ottaa huomioon käännökset altaan päissä. Muuten voidaan laskea kuin juoksussa.

Seuraavat tulokset on syytä huomioida:

- miehillä energiavaraston alkuarvo e_0 on sama kuin juoksussa
- naisilla e_0 on 67% miesten arvosta

- miehillä energiavaraston täydennysnopeus σ on 83% juoksijoiden arvosta
- naisilla σ on 95% miesten arvosta

PAINONNOSTO ([10], [11], [13]-[16])

Koska painonnostossa on kyseessä lyhytaikainen maksimaalinen ponnistus, tuotetaan energia täysin anaerobisesti. Aerobinen prosessi hoitaa vain nostajan palautumisen uuteen yritykseen. Jos arvostellaan vain nostotulosta eikä nostoa kokonaisuudessan dynaamisena tapahtumana, voidaan eri nostajien saman nostomuodon (tempaus tai työntö) tuloksia vertaillessa käyttää kaavassa (6) määriteltyä kvaliteettifunktiota k.

$$k = \frac{F}{G/p} \quad . \tag{6}$$

Luonteva kaikille sama referenssipituus on tangon suurimman levyn säde R = 22.5 cm, joka ilmoittaa millä korkeudella nostotangon painopiste on alkujaan lavasta. Jos kaavaan (6) sijoitetaan $F = Mg$ eli tangon paino ja $G = mg$ eli nostajan paino, saadaan kvaliteettifunktio ilmaistua ns. Riegert'in lukuna eli *Ri*-lukuna

$$Ri = p\frac{M}{m} \quad . \tag{56}$$

Aikaisemmin mainittu yhdenmuotoisuusotaksuma pätee erityisen hyvin painonnostajiin. Nostajan massan m ja pituuden $L = pR$ välillä vallitsee tilastollisesti likimääräinen yhteys

$$m \approx 0.2 \cdot p^3 \,[kg] \tag{57}$$

Riegert'in luvun vaihteluväli on tempauksessa $14 < Ri < 16$ ja työnnössä $17 < Ri < 19$.

Painonnoston dynaaminen malli ([15] ja [16])

Painonnoston dynaaminen malli voidaan helpoiten johtaa energiaperiaatteella. Tangon nostamisessa tehty työ muuttuu lopulta sen potentiaalienergiaksi.

Jos h on tarvittava korkeus lavan pinnasta, jotta L-pituinen nostaja pääsisi tangon alle, määritellään nostotekniikkaa kuvaava parametri $0.4 < \eta = h/L < 0.6$. Jos nostaja tangon alle syöksyessään siirtää painopistettään alaspäin matkan ℓ keskimääräisellä nopeudella V, voidaan vielä määritellä toinen tekniikkaparametri $0.3 < \lambda = \ell/L < 0.5$ sekä nopeusparametri $1 < u = V/\sqrt{gR/2} < 3$. Vielä määritellään yhdistetty parametri $q = (\lambda/u)^2$. Näiden lisäksi nostajan voimaa kuvaa suhteellinen voima $f(z)$, joka suhteellisen korkeuden $0.1 < z = y/L < 0.7$ funktiona on $15 < f(z) = F(y)p/mg < 30$. Tässä $F(y)$ on nostajan todellinen nostovoima, kun tanko on korkeudella y.

Tämän jälkeen voidaan määritellä uusi dynaamisen noston hyvyysluku eli ns. performanssi-indeksi PI ($\approx$ tehty työ/nostajan massa). Sille voidaan johtaa nostotuloksesta M lauseke

$$PI = (\eta p - 1)\frac{M}{m} \tag{58}$$

tai suhteellisesta voimasta $f(z)$ integraalilauseke

$$PI = \int_{1/p}^{\eta - qp} f(z)dz \ . \tag{59}$$

Performanssi-indeksin vaihteluväli on tempauksessa $5.8 < PI < 7.1$ ja työnnössä $5.9 < PI < 7.0$. Sen avulla voidaan verrata tempaus- ja työntötuloksia toisiinsa. Performanssi-indeksi riippuu näet periaatteesa vain voimasta ja nostotekniikan hallinnasta. Malli kuvaa todellisuutta yli 96.5% tarkkuudella. Mallin avulla voidaan spekuloida kaavoista (57) ja (59) ihmisen koon rajat [16]

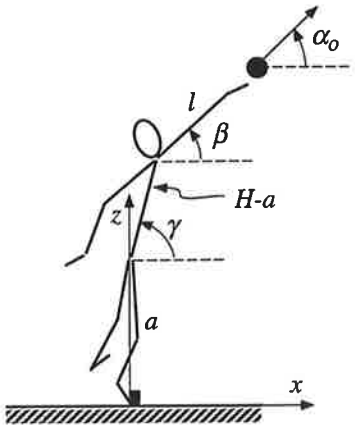
pienin koko: pituus 0.5 m ja massa 2 kg
suurin koko: pituus 6 m ja massa 4 tonnia.

Pienin koko muistuttaa hämmästyttävästi vastasyntyneen lapsen kokoa. Suurin koko taas on lähinnä suurimpien nisäkkäiden suuruusluokkaa.

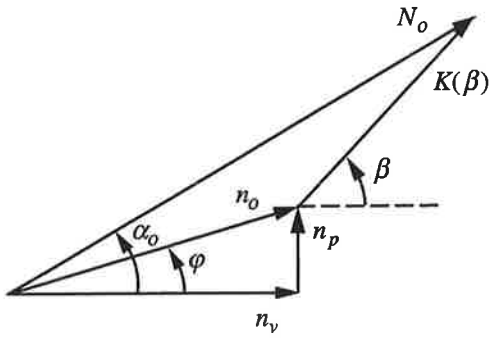
KUULANTYÖNTÖ ([1], [2], [23] ja [28])

Kuulantyönnöstä on julkaistu paljon. Julkaisut pohjautuvat pääasiassa tunnettuihin kaavoihin "heittoliike tyhjiössä". Niissä optimoidaan kuulakaaren pituus lähtökulman α_0 suhteen pitämällä kuulan irtoamispistettä kädestä tunnettuna.

Näitä ajatuksia mukaellen voidaan sekä klassilliselle työntötyylille että pyörähdystyylille esittää samantapainen malli. Ajatellaan, että työntäjä vie kuulaa olkapäällään korkeudella h vaakasuoralla nopeudella n_v. Vartaloaan ja jalkojaan oikaisemalla työntäjä kohottaa olkapäänsä korkeudelle H ja antaa samalla kuulalle pystysuoran nopeuskomponentin n_p sekä ojentaa l-pituisen käsivartensa suuntaan β (vaakatasosta) voimalla F kiihdyttäen kuulan lähtönopeuteen N_o lisänopeuden $K(\beta)$ avulla. Merkitään $n_o^2 = n_v^2 + n_p^2$, jossa n_o on resultoiva suuntaan $\varphi = Arc\tan(n_p/n_v)$ osoittava kuulan kuljetusnopeusvektori $\mathbf{n}_o$ ennen käden oikaisemista. Mallissa [28] on siis kaksi kulmaa, olkanivelen nousukulma φ ja käsivarren työntökulma β, joille kummallekin voidaan löytää optimaalinen arvo työnnön pituuden maksimoimiseksi pitäen olkanivelen sijaintia tunnettuna mutta kuulan irtoamispiste sijaitsee ympyrän, jonka säde on käsivarren pituinen, kehällä. Tilannetta selventävät kuvat 2 ja 3.



Kuva 2. Työntäjän skemaattinen malli sekä lähtökulma α_0, työntökulma β ja vartalokulma γ.



Kuva 3. Kuljetusnopeuden $\mathbf{n_0}$ ja lisänopeuden $\mathbf{K}(\beta)$ vektoraalinen yhteenlasku.

Kantaman optimointi työntökulman suhteen ([28])

Lähdetään yleisestä kantaman lausekkeesta tyhjiössä ja kirjoitetaan se muotoon

$$\frac{X(\beta)}{H} = \frac{x_o(\beta)}{H} + 2S(\beta)\left[\sin\alpha_o(\beta) + \sqrt{\sin^2\alpha_o(\beta) + \frac{z_o(\beta)/H}{S(\beta)}}\right]\cos\alpha_o(\beta) \qquad (60)$$

sekä kirjoitetaan vielä sen rinnalle tarvittavat lisäyhtälöt:

lähtöpisteen koordinaatit

$$\frac{x_o(\beta)}{H} = \left(1 - \frac{a}{H}\right)\cos\gamma + \frac{l}{H}\cos\beta \tag{61}$$

$$\frac{z_o(\beta)}{H} = \frac{a}{H}(1 - \sin\gamma) + \sin\gamma + \frac{l}{H}\sin\beta \ , \tag{62}$$

korkeusparametri $S(\beta)$

$$S(\beta) = \frac{n_o^2}{2gH} + \frac{l}{H}\left[\frac{M}{m}\frac{L}{l} - \sin\beta\right] \ , \tag{63}$$

kuulan effektiivinen kiihdytysmatka $\lambda = L/l$

$$\lambda = \frac{L}{l} = \frac{\ell}{l} + \frac{m}{M}\left(1 - \frac{\ell}{l}\right)\sin\beta + \frac{n_o}{\sqrt{gH(l/H)}}\sqrt{2\frac{m}{M}\frac{\ell}{l}\left(1 - \frac{m}{M}\sin\beta\right)}\cos(\beta - \varphi) \ , \tag{64}$$

kyynärkulman δ vaikutus

$$\frac{\ell}{l} = 1 - \sin\left(\frac{\delta}{2}\right) \ , \tag{65}$$

käsivarren oikaisusta kuulan saama lisänopeus $K(\beta)$

$$K(\beta) = \sqrt{2gH(\ell/l)(l/H)(M/m - \sin\beta)} \tag{66}$$

$$\equiv \sqrt{\left(n_p\sin\beta + n_v\cos\beta\right)^2 + 2gH\left(\frac{l}{H}\right)\left[\frac{M}{m}\frac{L}{l} - \sin\beta\right]} - \left(n_p\sin\beta + n_v\cos\beta\right) \tag{67}$$

sekä kuulan lähtökulma α_0

$$\alpha_o(\beta) = Arc\cot\left[\frac{n_v + K(\beta)\cos\beta}{n_p + K(\beta)\sin\beta}\right] \ . \tag{68}$$

Näistä ilmenee eri suureiden riippuvuus työntökulmasta β. Tämän yhtälösysteemin optimointi on parasta suorittaa numeerisesti ottaen huomioon työntäjän ruumiinrakenteesta riippuvat parametrit eli ruumiin osien pituudet a, H ja l sekä kyynärkulman δ. M on työntäjän työntökäden punnerrusvoima.

Työntökulman suhteen optimoidun mallin soveltaminen ([28])

Tarkastellaan työntäjää, jonka hartian korkeus on $H = 1.75$ m, käsivarren pituus $l = 0.75$ m ($\eta = l/H = 0.43$), jalan pituus $a = 1.05$ m, alkukyynärkulma $\delta = 40°$, vartalokulma $\gamma = 80°$ ja yhdenkäden punnerrustulos $M = 43.3$ kg. Vauhdinoton lopulla kuulan kuljetusnopeus olkoon $n_o = 6.15\,\mathrm{ms}^{-1}$ ja suunta saakoon arvot $\varphi = 0.0°$, 30°, $\varphi_{opt} \approx 43.0°$ ja 50°. Tulokset on esitetty Taulukossa 3.

Taulukosta nähdään, että jos kuljetusnopeuden suuruus n_0 on vakio, sen suunnalla on erittäin suuri merkitys. Optimisuunta φ_{opt} lienee helpompi toteuttaa klassillisella työntötyylillä kuin pyörähdystyylillä. Pyörähdystyylillä työnnettäessä kuljetusnopeus-vektori $\mathbf{n_0}$ näyttää olevan lähempänä vaakatasoa, mutta itse nopeus on suurempi kuin klassillisessa tyylissä.

TAULUKKO 3. Työntökulman suhteen optimoitu työntö eri kulman φ arvoilla tyhjiössä.

Suure	Symboli	Kuljetusnopeuden suunta φ			
		0.0°	30°	$\varphi_{opt} \approx 43.0°$	50°
Optimityöntökulma	β_{opt} (°)	48.37	42.12	38.81	36.95
Lähtökulma	α_o (°)	26.05	36.52	40.74	42.95
Lambda	λ	1.36	1.68	1.70	1.68
Irtoamispisteen koordinaatit	x_o (m)	0.62	0.68	0.71	0.72
	z_o (m)	2.30	2.24	2.21	2.19
Lähtönopeus	N_o (ms^{-1})	12.10	13.24	13.33	13.27
Nopeudenlisäys	K (ms^{-1})	7.11	7.16	7.19	7.21
Kantama tyhjössä	X (m)	16.01	20.40	20.91	20.75

Ilmanvastus ja tuuli

Optimisektori, johon työntösuunnan tulee langeta, on korkeudeltaan pari astetta eli $\beta_{opt} \pm 1°$. Ilmanvastus vähentää tuulettomassa säässä noin 0.5% edellä laskettuja tyhjiö-työntötuloksia, ja β_{opt} on noin asteen pienempi kuin tyhjiössä. Myötätuuleen 10 m/s työnnettäessä kasvaa β_{opt} noin puoli astetta ja vastatuuleen 10m/s vähenee vastaavasti saman verran. Sivutuuli 10 m/s ei vaikuta kulmaan β_{opt}, mutta lyhentää edelleen työnnön pituutta hieman.

LOPPUSANAT

Biomekaniikka ja erityisesti urheilun biomekaniikka tarjoavat kiintoisan tutkimuskentän alasta kiinnostuneille mekaniikan harrastajille. Kaikki urheilusuoritukset ovat mekaniikan lakien alaisia. Usein jo varsin yksinkertainenkin malli kuvaa suoritusta yllättävän hyvin. Mekaniikan ja matematiikan taitajien tulisi muodostaa tutkimusryhmiä yhdessä urheilun ja fysiologian asiantuntijoiden kanssa. Täten voitaisiin laatia kehittyneempiä malleja, joista varmasti saataisiin irti myös enemmän tietoa. Unohtaa ei sovi myöskään, että urheilututkimuksissa tarvitaan aina suoritusten analysoitia varten erilaisia mittauksia ja rekisteröintejä. Se edellyttää myös tarvittavien laitteiden kehittämistä ja rakentamista, ellei niitä ole valmiina saatavilla.

Biomekaniikan alalla on vuodesta 1967 järjestetty joka toinen vuosi kansainvälinen kongressi. Vuosina 1975 ja 1995 tämä kongressi oli Jyväskylässä. Urheilun ja biomekaniikan tutkimus onkin Suomessa keskittynyt juuri Jyväskylään. Siellä on yliopistossa liikuntatieteellinen tiedekunta sekä sen ulkopuolella itsenäinen Kilpaurheilun Tutkimuskeskus (KIHU). Myös muiden yliopistojen lääketieteellisissä tiedekunnissa harrastetaan urheilua tukevaa tutkimusta. Lisäksi maassamme on olemassa muitakin tutkimuslaitoksia, joiden tutkimusohjelmassa urheilu on mukana tavalla tai toisella.

Suomen Olympiakomitealla on apunaan mm.
- teknis-luonnontieteellinen asiantuntijaryhmä
- urheilulääketieteellinen asiantuntijaryhmä
- urheilupsykologinen asiantuntijaryhmä
- valmennuksen johtoryhmä.

Näiden ryhmien tavoitteena on varmistaa Suomen kansainvälisen huippu-urheilun menestys. Sen toteuttamiseksi ryhmät käsittelevät valmennuksen kehittämismääräraha-anomuksia ja antavat lausuntoja erilaisista tutkimussuunnitelmista sekä järjestävät tutkijaseminaareja ja osallistuvat muiden järjestämiin urheilututkimukseen liittyviin tilaisuuksiin. Asiantuntijaryhmiä voitaisiin käyttää näkyvämmin yleisen mielipiteen muokkaukseen, jotta urheilun tutkimiselle painopistealoilla myönnettäisiin enemmän varoja.

Biomekaniikan ja urheilun niinkuin minkä tahansa alan tutkimuksessa poikkitieteellisyys ja kansainvälisyys ovat tärkeitä näkökohtia. Eräs mahdollinen ulkomainen yhteistyökumppani tällä alalla olisi Tallinnan teknillinen korkeakoulu, jossa on jo aloitettu biomekaniikan opetus ([25]) insinöörikoulutuksen yhteydessä. Vastaavanlaista koulutusta kannattaisi harkita myös meillä Suomessa.

LÄHTEET

1. Hochmuth, G., *Biomechanik sportlicher Bewegungen*. Wilhelm Limpert-Verlag GMbH, Frankfurt (Main) 1967. 232 s.
2. Dyson, G, *The Mechanics of Athletics*. University of London Press LTD, London 1967. 224 s.
3. Tuokko, R., *Urheilija luonnonlakien kahleissa*. WSOY, Porvoo 1965. 121 s.
4. O'Shea, J.P., *Scientific Principles and Methods of Strength Fitness*. Addison-Wesley Publishing Company, USA 1968. 165 s.
5. Seppänen, L. & Oikarinen, E., *Kestävyysvalmennus*. Suomen Valtakunnan Urheiluliitto, Helsinki 1976. 235 s.
6. Kairento, A., *Biomekaaninen analyysi*. Arkhimedes 30 (1978), s. 213-217.
7. Liljeström, T., *Uimahyppyponnistuksen biomekaniikka*. Arkhimedes 30 (1973), s. 218-221.
8. Keller, J.B., *A Theory of Competitive Running*. Physics to Day 26 (1973) 9, s. 42-47.
9. Keller, J.B., *Optimal Velocity in Race*. American Mathematical Monthly 51 (1974) 5, s. 474-480.
10. Keller, J.B., *Mechanical Aspects of Athletics*. Proceedings of the Seventh U.S. National Congress of Applied Mechanics Boulder, Co, June 1974, s. 22-26.
11. Keller, J.B., *Mechanical Aspects of Athletics*, (Javelin Throwing, Weightlifting, Rowing). Optimal Strategies in Sports. North-Holland Publishing Company, 1975. s. 136-224.
12. Francis, P.R. and Dean, N., *A Biomechanical Model for Swimming Performance*. International Series on Sport Sciences. University Park Press. 2 (1975) Swimming II, s. 118-124.
13. Lilien, L., *Optimal Weightlifting*. Management Science in Sports. NorthHolland Publishing Company, 1976. s. 101-112.
14. Ranta, M.A., *Voimaurheilun teoriaa*. Oulun yliopisto, koneinsinööriosasto. Raportti No. 2, 1969. 52 s.
15. Ranta, M.A., *Simple Mathematical Model of Weightlifting*. Biomechanics V-B. International Series of Biomechanics. University Park Press 1B (1976), s. 337-343.
16. Ranta, M.A., *A Mathematical Model of Weightlifting*. Invited general lecture in the General Assembly of IUTAM at Herrenalb BRD on 5th September 1978, 29 s.
17. Pramila, A., *On the Effect of the Size of the Skier on the Time Needed to Clear a Downhill*. Helsingin teknillinen korkeakoulu, yleinen osasto, mekaniikan laitos. Julkaisu No. 4, 1979, 17 s.
18. Pramila, A., *A Novel Model for Speed Skiing*. Journal of Structural Mechanics, Vol. 30, 1997, Nro 2. s. 65-74.
19. Keinonen, J. ja Palosuo, E., *Suksen ja lumen välisestä kitkasta*. Arkhimedes 31 (1979) 4, s. 211-215.
20. Holmlund, U., *Nopeuslaskuun, murtomaahiihtoon ja mäkihyppyyn liittyvän mäenlaskun matemaattinen mallintaminen,* Lisensiaatintyö. Teknillinen

korkeakoulu, tietotekniikan osasto, laskennallisen dynamiikan laboratorio, Otaniemi 1993.

21. Olkinuora, P., *Keihään aerodynaamisten kertoimien mittaus ja lentoradan laskenta.* Diplomityö. Helsingin teknillinen korkeakoulu, koneinsinööriosasto, Otaniemi 1979. 120 s.
22. Franssila, M., *Koleus kotikenttäetu. Iltasäät suosivat Suomen juoksijoita.* Uusi-Suomi, 27. toukokuuta 1975.
23. Angelo, A., Jr., *The Physics of Sports.* American Institute of Physics, New York, second printing, 1993.
24. Alexandrov, I. and Luchtin, P., (1984) *Physics of sprinting*, s. 254-257 lähteessä [23].
25. Ranta, M., A., *Biomecanics and Sport, especially Running*, Vierailuluento Tallinnan teknillisessä korkeakoulussa, 18. joulukuuta 1995.
26. Holmlund, U., von Hertzen, R. ja Ranta, M.A., *Eräs pikajuoksun matemaattinen malli.* Arkhimedes 3/96, s. 8-13.
27. Holmlund, U., von Hertzen, R., *Models of Sprinting based on Newton's second Law of Motion and their Comparison*, Journal of Structural Mechanics, Vol. 30, 1997, Nro 2. s. 7-16.
28. Ranta, M.,A., von Hertzen, R. ja Holmlund, U., *Kuulantyönnön dynaaminen malli.* Teknillinen korkeakoulu, Teknillisen fysiikan ja matematiikan osato, Laskennallinen dynamiikka. Tutkimusraportti No 44, 1996, 10 s.

SPECIAL TOPICS IN STRUCTURAL OPTIMIZATION
MULTICRITERIA DESIGN

J. KOSKI
Laboratory of Applied Mechanics
Tampere University of Technology
P.O. Box 589
FIN-33101 Tampere, FINLAND

ABSTRACT

Structural optimization problem with several conflicting and noncommensurable criteria is considered. A motivation for the multicriteria approach is discussed and the corresponding vector optimization problem is formulated. Pareto optimal solutions in the design space and minimal solutions in the criteria space are defined. The basic techniques to generate Pareto optima are briefly described. A special attention is paid to the concept of conflict and its consequences in multicriteria optimization. Both local and global conflict in the case of two criteria are introduced and illustrated.

1. INTRODUCTION

Usually a scalar objective function, which in most cases is the weight of the structure, is optimized in the feasible set defined by the equality and inequality constraints. In practical applications, however, the weight rarely represents the only measure of the performance of a structure. In fact, several conflicting and noncommensurable criteria usually exist in real-life design problems. This situation forces the designer to look for a good compromise design by performing trade-off studies between the conflicting requirements. Consequently, he must take a decision-maker's role in an interactive design process where generally several optimization problems must be solved. Multicriteria optimization offers one flexible approach for the designer to treat this overall decision-making problem in a systematic way.

Multicriteria (multicriterion, multiobjective, Pareto, vector) optimization has recently achieved an established position also in structural design. One reason for the introduction of this approach is its natural property of allowing partici-

pation in the design process after the formulation of the optimization problem. It is generally considered that multicriteria optimization in its present sense originated towards the end of the last century when Pareto (1848 - 1923) presented a qualitative definition for the optimality concept in economic problems with several competing criteria [1]. Some other even earlier contributors have been discussed for example by Stadler [2]. A wider interest in this subject concerning the fields of optimization theory, operations research and control theory was aroused at the end of the 1960s and since then the research has been very intensive also in engineering design [3, 4, 5]. Especially in structural optimization, the first applications in the English-language literature appeared in the late 1970s [6 - 10] giving an impetus to emphasizing the decision-maker's viewpoint also in the design of load-supporting structures.

The purpose of this article is to introduce the basic concepts and methods used in multicriteria structural optimization. Special attention has been paid to those fundamental matters which are common to most of the published applications. These general ideas have been illustrated by an example problem where the emphasis is rather on the multicriteria view than on the numerical solution techniques. Specifically, the multicriteria problem formulation and the generation of Pareto optima as well as the concept of conflict are considered. The decision-making process for finding the best Pareto optimal design is also briefly discussed.

2. CRITERIA AND CONFLICT

In structural optimization one is faced by the question of which criteria are suitable for measuring the economy and performance of a structure. Such a quantity that has a tendency to improve or deteriorate is actually a criterion in nature. On the other hand, those quantities which must only satisfy some imposed requirements are not criteria but they can be treated as constraints. Most of the commonly used design quantities have a criterion nature rather than a constraint nature because in the designer's mind they usually have better or worse values. As an example of a strict constraint the structural analysis equations or any physical laws governing the system can be mentioned. They represent equality constraints whereas different official regulations and norms generally impose inequality constraints. For example such matters as space limitations, strength and manufacturing requirements are often treated as inequality constraints. One difficulty appears in choosing the allowable constraint limits which may be rather fuzzy in real-life problems. If these allowable values cannot be determined it seems reasonable to treat the quantity in question as a criterion.

An important basic property in the multicriteria problem statement is a conflict between the criteria. Only those quantities which are competing should be treated as criteria whereas the others can be combined as a single criterion or

one of them may represent the whole group. In the literature the concept of the conflict has deserved only a little attention while on the contrary the solution procedures have been studied to a great extent. In the problem formulation, however, it is useful to consider the conflict properties because it helps to create a good optimization model. For example in [11] this topic is discussed in general terms and in [12], where truss design is studied, the concepts of a local and global conflict have been proposed. According to the latter presentation the local conflict between two criteria can be defined as follows. Functions f_i and f_j are called collinear with no conflict at point x if there exists $c>0$ such that $\nabla f_i(x) = c\nabla f_j(x)$. Otherwise, the functions are called **locally conflicting** at x.

Consequently, any two criteria are locally conflicting at a point in the design space if their maximum improvements are achieved in different directions. The angle between the gradients can be used as a natural measure of the local conflict. Even if two criteria are locally conflicting almost everywhere in the design space they still can achieve their optimum value at the same point. Thus it seems necessary to consider separately the concept of the global conflict where also the feasible set is involved. Functions f_i and f_j are called **globally conflicting** in Ω if the optimization problems $\min_{x\in\Omega} f_i(x)$ and $\min_{x\in\Omega} f_j(x)$ have different solutions.

These concepts have been illustrated in Figure 1 where the relevant situations in the design space are shown. In structural optimization, usually the weight and any chosen displacement are both locally and globally strongly conflicting quantities. Displacements often achieve their minima at the same point but still they may be locally conflicting in that part of Ω where the best design locates.

3. MULTICRITERIA PROBLEM AND PARETO OPTIMALITY

The choice of the design variables, criteria and constraints certainly represents the most important decisions in structural optimization because the designs which will be available in the continuation are fixed at this very early stage. For example in scalar optimization the minimization of a single criterion in the feasible set usually results in one optimal solution only. Certainly numerical computations are needed to get that optimum design but, as a matter of fact, all the decisions have been made already in the problem formulation.

The multicriteria problem inherently offers a possibility to perform a systematic sensitivity analysis for the chosen criteria. The difference between criteria and constraints is that the designer wants to improve the value of a criterion whereas this kind of desire is not associated with the constraints. As a natural consequence of the separation of the criteria f_i, $i = 1,2,\ldots,m$, and the constraints the following multicriteria problem is obtained:

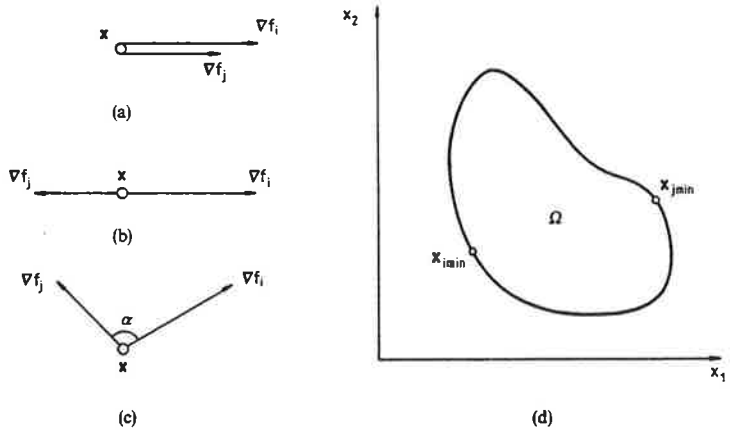


FIGURE 1: Conflict of criteria: (a) no conflict; (b) local complete conflict; (c) local conflict; (d) global conflict in the case of two design variables. Points $x_{i\,\min}$ and $x_{j\min}$ represent the individual minima of the criteria f_i and f_j in Ω, respectively.

$$\min_{x\in\Omega}\left[\, f_1(x)\; f_2(x) \ldots f_m(x)\,\right]^T. \tag{1}$$

Here $x=\left[x_1 x_2 \cdots x_n\right]^T$ represents a design variable vector and Ω is the feasible set in design space R^n. It is defined by inequality and equality constraints in the form

$$\Omega=\left\{\, x\in R^n \,|\, g(x)\le 0,\; h(x)=0 \,\right\}. \tag{2}$$

By using the notation $f(x) = \left[\, f_1(x)\; f_2(x) \ldots f_m(x)\,\right]^T$ for the vector objective function, which contains the m conflicting and possibly noncommensurable criteria as the components, the image of the feasible set in criteria space R^m is expressed as

$$\Lambda=\left\{\, z\in R^m \,|\, z = f(x),\; x\in\Omega \,\right\}. \tag{3}$$

This is called the attainable set and apparently it is more interesting for the decision-maker than the feasible set. Usually there exists no unique point which would give an optimum for all m criteria simultaneously. Thus the common optimality concept used in scalar optimization must be replaced by a new one, especially adapted to the multicriteria problem.

Only a partial order exists in criteria space R^m and thus the concept of Pareto optimality offers the most natural solution in this context. A vector $x^* \in \Omega$ is **Pareto optimal** for problem (1) if and only if there exists no $x \in \Omega$ such that $f_i(x) \leq f_i(x^*)$ for $i = 1,2,...,m$ with $f_j(x) < f_j(x^*)$ for at least one j.

This definition states that x^* is Pareto optimal if there exists no feasible vector x which would decrease some criterion without causing a simultaneous increase in at least one other criterion. In the literature also some other terms have been used instead of the Pareto optimality. For example words such as nondominated, noninferior, efficient, functional-efficient and EP-optimal solution have the same meaning. Here only the mathematical programming problem has been shown and applied but the corresponding control theory formulation can be found for example in [4, 5].

Two different spaces R^n and R^m, called the design and the criteria space, appear in a multicriteria problem. In order to avoid any confusion it is necessary to distinguish the optimal solutions in these separate spaces. Consequently, the vector $z^* = f(x^*)$, which represents the image of the Pareto optimum x^* in the criteria space, is called the **minimal solution**. Optimality concepts in both spaces have been illustrated in Figure 2 where the bicriteria case has been considered. So-called weak solutions, which are also shown in the figure, and their existence in structural optimization have been discussed in [13]. In scalar optimization, one optimal solution is usually characteristic of the problem, whereas there generally exists a set of Pareto optima as a solution to the multicriteria problem. Mathematically, problem (1) can be regarded as solved immediately

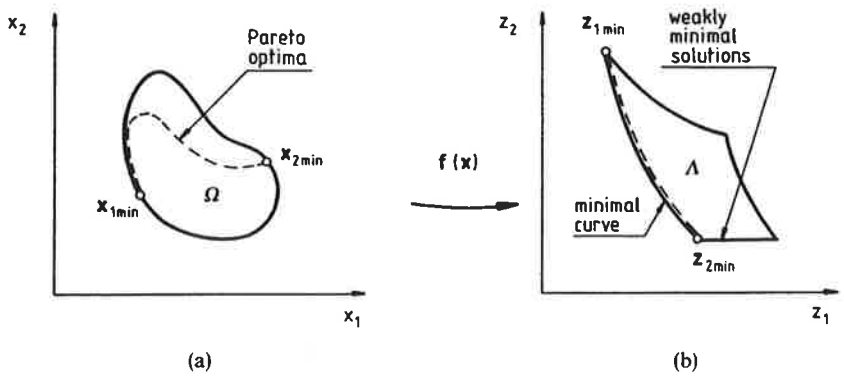


FIGURE 2: Pareto optimal and minimal solutions in bicriteria case with two design variables: (a) feasible set and Pareto optimal curve; (b) attainable set and minimal curve. Points $x_{1\min}$ and $x_{2\min}$ in the design space correspond to points $z_{1\min}$ and $z_{2\min}$ in the criteria space.

after the Pareto optimal set has been determined. In practical applications, however, it is necessary to order this set further because only one final solution is wanted by the designer. Thus he must take a decision-maker's role and introduce his own preferences to find the best compromise solution among Pareto optima.

4. WHY MULTICRITERIA APPROACH?

The formulation of a vector optimization problem, where several competing criteria are minimized simultaneously, may arouse confusion and criticism because a set of Pareto optima rather than one optimal design is obtained as a solution to problem (1). Questions concerning the advantages of the multicriteria approach compared with the traditional single criterion or scalar optimization can spring up in different contexts. Next some major drawbacks of the scalar approach are introduced by using a simple two-bar truss example.

The truss, loading and numerical design data are given in Figure 3a. Two competing criteria, the material volume V of the truss and the vertical nodal displacement Δ of the loaded node, are chosen for minimization. They are most evidently conflicting because a light structure is obtained by using small member areas and a stiff structure by using large member areas. These cross-sectional or member areas A_1 and A_2 are the design variables, ie the components of the design variable vector $\boldsymbol{x}=[A_1\, A_2]^T$. Stress constraints, where σ^u is the upper and σ^l the lower limit for stresses, are imposed. Upper limits A^u for the member areas are given for convenience, just to make the feasible set compact. The bicriteria problem is

$$\min[V(\boldsymbol{x})\ \Delta(\boldsymbol{x})]^T \tag{4}_1$$

subject to

$$\begin{aligned} &\sigma^l \le \sigma_i(\boldsymbol{x}) \le \sigma^u, \\ &\qquad A_i \le A^u, \quad i=1,2\,. \end{aligned} \tag{4}_2$$

The feasible set Ω, which in this isostatic case is the rectangle ABCD shown in Figure 3b, consists of those member area combinations which do not violate the above constraints. Point A corresponds to the minimum material volume solution and point C to the minimum of displacement Δ in the feasible set. These criteria are both locally and globally strongly conflicting achieving their minima at the extreme vertices of the feasible set.

Pareto optima of problem (4) consist of the polygonal line AEC depicted in Figure 3b and the corresponding minimal curve shown in Figure 3c lies on the front boundary of the attainable set Λ. These solutions offer large flexibility for

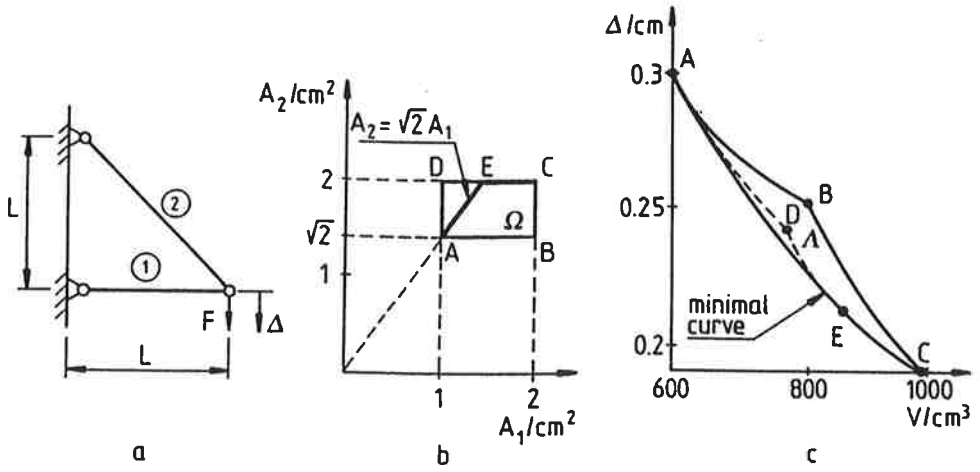


FIGURE 3: Bicriteria truss example: (a) Structure, loading, and displacement criterion Δ; (b) feasible set Ω in design space and Pareto optimal polygonal line AEC; (c) attainable set Λ in criteria space and minimal curve AEC. The broken line inside Λ corresponds to part ADE on the boundary of Ω. Design data for the problem, given in kN and centimeters: F = 10 kN, σ^u = 10 kN/cm^2, A^u = 2 cm^2, L = 200 cm, σ^l = -10 kN/cm^2, $E = 2 \cdot 10^4$ kN/cm^2.

a designer who is looking for the best compromise solution among Pareto optima. It should be noticed that every attainable solution inside Λ or on the boundary ABC is dominated by (ie. is worse than) some of the minimal solutions. At this stage of the design process it is also possible to consider other aspects, say aesthetics or manufacturing cost for example, which were not included in the optimization problem.

If the truss design problem at hand were treated by scalar optimization, probably one of the two commonest approaches discussed in the sequel would be applied. The first approach is to combine both the criteria linearly into one objective function and solve the scalar problem

$$\min_{x \in \Omega} \; \lambda V(x) + (1-\lambda)\Delta(x), \qquad 0 \le \lambda \le 1, \tag{5}$$

where a fixed weighting factor is used. It is obvious that the optimal solution of problem (5) strongly depends on the value of parameter λ, varying from the

minimum material volume solution (point A in Figure 3, obtained by $\lambda = 1$) to the minimum displacement solution (point C in Figure 3, obtained by $\lambda = 0$). The first thing that goes wrong is the evaluation of correct weights. The numerical values of the material volume and the nodal displacement have different magnitudes and for example a choice $\lambda = 0.5$ gives almost the minimum material volume solution if cm units are used. Different normalizations like

$$\tilde{f}_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x})}{f_{i\min}}, \qquad i = 1,2,\ldots,m, \tag{6}$$

can be used to alleviate this difficulty. Here $f_{i\min}$ represents the minimum value of criterion $f_i(\boldsymbol{x})$ in Ω. Even if some normalization is applied, this approach may give poor results because weights λ_i actually are parameters, which also have better or worse values, in the overall decision-making process. Multicriteria optimization treats them as parameters which the designer should choose in an optimal way. In addition, it is extremely difficult to recognize a possible duality gap in some nonconvex problems by just applying formula (5). Only the multicriteria approach reveals the conditions under which Pareto optima may be lost by solving scalar problem (5), no matter what weights are used.

Another frequently used scalar approach in the case of several criteria is to choose one criterion into the objective function and to remove the others into constraints. Usually a criterion which is considered as the most important, is chosen as a scalar objective function. In the truss example the material volume of the structure could be a single criterion to be minimized and the nodal displacement Δ is restricted by some chosen upper limit ε. Then the scalar optimization problem

$$\min_{\boldsymbol{x}\in\Omega} V(\boldsymbol{x}) \tag{7_1}$$

subject to

$$\Delta(\boldsymbol{x}) \le \varepsilon, \tag{7_2}$$

where Ω represents the original feasible set defined by stress and member area constraints (4_2), is formulated. Again the optimal solution totally depends on the chosen allowable value ε. Often in practical design the value of ε is rather fuzzy and difficult to fix in advance. More freedom is obtained if it is used as a parameter, which also has some optimal value, in the design process like the multicriteria approach does.

As this bicriteria truss example shows, the traditional single criterion approach may give poor results in the case where two or more conflicting criteria exist. No matter what scalar optimization formulation is used, there always appear parameters which should be fixed before optimization. If they are used as pa-

rameters and varied heuristically during optimization, the scalar method do not offer any systematic procedure to find good values for these parameters. The Pareto optimality concept offers a collection of optimal designs which should be pursued by any chosen method. The multicriteria decision-making based on the vector optimization problem evidently gives a systematic approach to rationally search for good designs. Moreover, it should be stressed here that if the number of the criteria increases ($m > 2$) then the choice of a correctly working scalar problem for optimization becomes more difficult than in this bicriteria truss problem. The graphical reasoning cannot be applied to verify if the used scalar approach produces Pareto optima or somethig else.

5. GENERATION OF PARETO OPTIMAL SOLUTIONS

5.1 Linear weighting method

Several methods for generating Pareto optima to a multicriteria optimization problem have been developed. Usually their application leads to the solution of several scalar problems which include certain parameters. Typically, each parameter combination corresponds to one Pareto optimum and by varying their values it is possible to generate the Pareto optimal set or its part. In the sequel those fundamental methods, which have been applied repeatedly in the structural optimization literature, are briefly described. They are also illustrated graphically in the criteria space in order to show the reasons for their different potential to cover the Pareto optimal set.

The linear weighting method combines all the criteria into one scalar objective function by using the weighted sum of the criteria. If the weighting coefficients are denoted by w_i, i = 1,2,...,m, this scalar optimization problem takes the form

$$\min_{x\in\Omega} \sum_{i=1}^{m} w_i f_i(\boldsymbol{x}) \tag{8}$$

where the normalization

$$\sum_{i=1}^{m} w_i = 1 \tag{9}$$

can be used without losing generality. By varying these weights it is now possible to generate Pareto optima for problem (1). The main disadvantage of this method is the fact that only in convex problems it can be guaranteed to generate the whole Pareto optimal set. According to the author's experience, such non-convex cases where the weighting method fails to generate all Pareto optima are not typical of structural optimization. Some simple truss examples, which dem-

onstrate that this phenomenon really exists in applications, have been reported in the literature [14]. The geometrical interpretation of the weighting method in a bicriteria problem is shown in Figure 4a where it corresponds to the case p = 1. It is interesting to notice that problem (8) expressed in the criteria space has the form

$$\min_{z \in \Lambda} \sum_{i=1}^{m} w_i z_i \tag{10}$$

where $z_i = f_i(\boldsymbol{x})$ for $i = 1,2,\ldots,m$. Thus a linear objective function is minimized in the attainable set.

5.2 Constraint method

One natural technique is to replace the original multicriteria problem by a scalar problem where one criterion f_k is chosen as the objective function and all the other criteria are removed into the constraints. By introducing parameters ε_i into these new constraints an additional feasible set

$$\Omega_k(\varepsilon_i) = \left\{ \boldsymbol{x} \in \boldsymbol{R}^n \mid f_i(\boldsymbol{x}) \leq \varepsilon_i, \ i = 1,2,\ldots,m, \ i \neq k \right\} \tag{11}$$

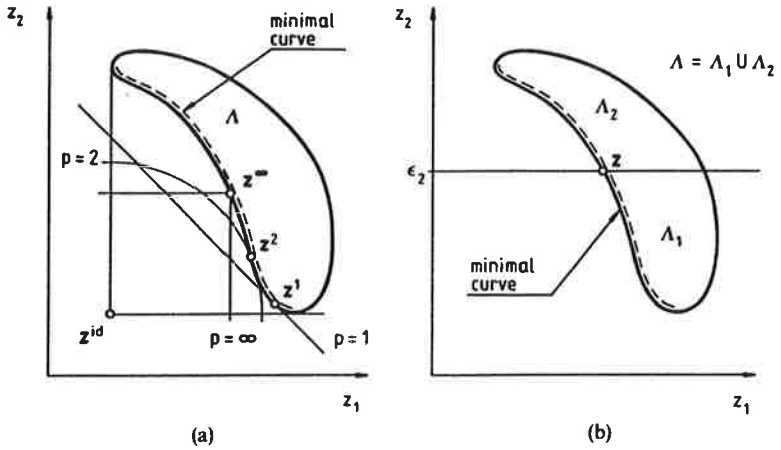


FIGURE 4: Geometrical interpretation of norm and constraint methods in bicriteria case: (a) linear weighting method (p = 1), weighted quadratic (p = 2) and minimax (p = ∞) methods illustrated in the criteria space; (b) constraint method where $f_1(\boldsymbol{x})$ is chosen as the scalar objective function and $f_2(\boldsymbol{x})$ is removed into the constraints.

is obtained. If the resulting feasible set is denoted by $\overline{\Omega}_k = \Omega \cap \Omega_k$, the parametrized scalar problem can be expressed as

$$\min_{x \in \overline{\Omega}_k} f_k(\boldsymbol{x}). \tag{12}$$

Here each parameter combination yields a separate problem usually corresponding to one Pareto optimum. This technique, called the constraint method, can generate the whole Pareto optimal set also in nonconvex cases and it has been applied to some extent in structural optimization. If so-called weak solutions [13] shown in Figure 2 exist, the constraint method can be modified to cover that case as well but then equality constraints appear and several new scalar problems must be solved for just one Pareto optimum. The geometric interpretation of the method is given in Figure 4b where $\mathbf{z}_1$ is minimized in the set Λ_1 which is the image of the set $\overline{\Omega}_1$.

5.3 Norm methods

Norm methods are based on the minimization of the distance between the attainable set and some chosen reference point in the criteria space. In the literature they have also been called metric, distance, and global criterion methods. The scalar problem is

$$\min_{\boldsymbol{x} \in \Omega} d_p(\boldsymbol{x}) \tag{13}$$

where the distance function

$$d_p(\boldsymbol{x}) = \left\{ \sum_{i=1}^{m} w_i \left[f_i(\boldsymbol{x}) - \hat{z}_i \right]^p \right\}^{1/p} \tag{14}$$

has been widely used in structural optimization. The reference point $\hat{\boldsymbol{z}} \in \boldsymbol{R}^m$ may be chosen by the designer and often the so-called ideal or utopia point

$$\boldsymbol{z}^{id} = \left[f_{1\min} \; f_{2\min} \; \cdots f_{m\min} \right]^T \tag{15}$$

can be found in the applications. This ideal vector contains all the individual minima of the criteria in Ω as components. Thus it is necessary to solve m scalar optimization problems

$$\min_{\boldsymbol{x} \in \Omega} f_i(\boldsymbol{x}), \qquad i = 1,2,\ldots m \tag{16}$$

if z^{id} is used as a reference point $\hat{z}$. The normalization given in equation (9) is also applicable for the weights w_i here. Usually $\hat{z}$ and p are fixed and w_i are the only parameters but also other possibilities exist [13]. If the origin is used as a reference point, i.e. $\hat{z}=\mathbf{0}$, then the extreme case $p=\infty$ in equation (14) corresponds to the weighted minimax problem

$$\min_{x\in\Omega} \max_{i} [w_i f_i(\boldsymbol{x})], \qquad i = 1,2,\ldots,m \tag{17}$$

which is capable of generating all Pareto optima also in nonconvex problems. The other extreme case $p = 1$ can be interpreted as the linear weighting method. Correspondingly, the case $p = 2$ might be called as a weighted quadratic method. All these three cases have been illustrated together in Figure 4a. In practical applications, where the numerical values of the noncommensurable criteria may have huge variations with respect to each other, it is useful to normalize all the criteria before computations. One possibility is to use the formula

$$\tilde{f}_i(\boldsymbol{x}) = \frac{f_i(\boldsymbol{x}) - f_{i\min}}{f_{i\max} - f_{i\min}} \tag{18}$$

where all the nondimensional criteria are limited to an equal range, i.e. $\tilde{f}_i(\boldsymbol{x}) \in [0,1]$, $i = 1,2,\ldots,m$. The quadratic case $p = 2$ seems to be the most popular choice in the literature but also both of the extreme cases have been used frequently in structural design applications.

6. CONCLUSION

The vector optimization problem and the Pareto optimality concept offer a clear and sound basis for any further development in the multicriteria optimum design theory. The proper control of the decision-making process becomes possible in this way because then the choice can be restricted to designs which are optimal in an undisputed mathematical sense. Also, the comparison of different methods is facilitated by these concepts. For example the question of which technique should be applied in a certain problem to generate good designs, may be often answered by comparing how many Pareto optima different methods can achieve. Whatever method the designer may choose, he should check whether it computes Pareto optima or something else. In the latter case the method is useless, and in the former case it should be able to generate as many Pareto optima as possible. If the designer wants to generate good points by just varying the constraint limits in a scalar problem, he has several possibilities. Both equality and inequality constraints can be applied in many different ways. The choice among these alternative methods can be made reliably only by checking which technique gives Pareto optima. The choice of the method by intuition may lead to

undesired designs. Sometimes it has been suspected that the solution of a multicriteria problem depends on the chosen decision-making procedure. This is true in the sense that usually experience is being acquired by the decision-maker during the design process. If the decision-maker's preferences remain unchanged during the process and the Pareto optimality concept is used, then any available procedure should give the same result provided that the designer makes consistent decisions. The natural requirement is that the chosen method must be able to generate also the best compromise solution. Thus in the case of unchanged preferences the result can be regarded as a function of the person making decisions rather than a function of the design procedure.

The future of the multicriteria approach looks promising also in structural optimization where it has recently reached industrial applications. A multicriteria module, which generates Pareto optimal solutions, can be expected to be an essential part in any new optimization-oriented finite element package. The computation of a covering collection of Pareto optima may be expensive for large scale problems because several structural and sensitivity analyses must be performed for each solution. This matter together with the graphic or numerical representation of the results, which should be modified into a form that is suitable for the decision making, will be one challenge in the field.

REFERENCES

1. V. Pareto, *Cours d'Economic Politique*, Volumes 1 and 2, Rouge, Lausanne (1896 - 1897).
2. W. Stadler, *Initiators of multicriteria optimization - recent advances and historical development of vector optimization, in Proceedings of an International Conference on Vector Optimization* (eds. J. Jahn and W. Krabs), Lecture Notes in Economics and Mathematical Systems 294, Springer, Berlin, Darmstadt (1986).
3. A. Osyczka, *Multicriterion Optimization in Engineering*, Ellis Horwood, Chichester (1984).
4. W. Stadler (ed.), *Multicriteria Optimization in Engineering and in the Sciences*, Mathematical Concepts and Methods in Science and Engineering 37, Plenum, New York (1988).
5. H. Eschenauer, J. Koski and A. Osyczka (eds.), *Multicriteria Design Optimization - Procedures and Applications*, Springer, Berlin (1990).
6. W. Stadler, *Natural structural shapes of shallow arches*, Journal of Applied Mechanics, 44, (1977), 291-298.
7. G. Leitmann, *Some problems of scalar and vector-valued optimization in linear viscoelasticity*, Journal of Optimization Theory and Applications, 23, (1977), 93-99.
8. W. Stadler, *Natural structural shapes (the static case)*, Quarterly Journal of Mechanics and Applied Mathematics, 31, (1978), 169-217.

9. E. N. Gerasimov and V. N. Repko, *Multicriterial optimization*, Soviet Applied Mechanics, 14, (1978), 1179-1184.
10. J. Koski, *Truss Optimization with Vector Criterion*, Publication No. 6, Tampere University of Technology (1979).
11. J. L. Cohon, *Multiobjective Programming and Planning*, Academic Press, New York (1978).
12. J. Koski, *Bicriterion optimum design method for elastic trusses*, Acta Polytechnica Scandinavica, Mechanical Engineering Series No. 86, Dissertation, Helsinki (1984).
13. J. Koski and R. Silvennoinen, *Norm methods and partial weighting in multicriterion optimization of structures*, International Journal for Numerical Methods in Engineering, 24, (1987), 1101-1121.
14. J. Koski, *Defectiveness of weighting method in multicriterion optimization of structures*, Communications in Applied Numerical Methods, 1, (1985), 333-337.

ERROR ANALYSIS OF THE STABILIZED MITC PLATE ELEMENTS

Mikko Lyly
Faculty of Mechanical Engineering
Helsinki University of Techonlogy
02150 Espoo
mikko.lyly@hut.fi

Rolf Stenberg
Institut für Mathematik und Geometrie
Universität Innsbruck
Technikerstrasse 13, A-6020 Innsbruck
rolf.stenberg@uibk.ac.at

Abstract

Three new families of finite element methods for the Reissner-Mindlin plate bending model are described. The methods are based on a combination of the stabilized formulation presented in [29] and the MITC reduction technique [7]. The families use identical basis functions for the deflection and the rotation. Optimal order of convergence, independent of the plate thickness, is proved.

1 Introduction

The purpose of this paper is to present an error analysis of our stabilized MITC plate bending elements. In earlier communications [22, 21] we have presented results of numerical calculations with these elements.

In the methods we combine the shear projection technique of the original MITC elements [7, 5] with recent stabilized formulations [17, 29]. The advantage of this, compared to both the MITC elements and the previous stabilized formulations, is that identical shape functions can be used for all unknows. Compared to more traditional methods, a stabilized formulation gives a more well conditioned stiffness matrix.

Another big advantage of these new families of methods is that they include convergent triangular linear and quadrilateral bilinear elements. These lowest order elements were introduced in [10] in connection with a general analysis of the MITC elements. In that context, the modification is in the spirit of the "trick" introduced by Fried and Yang already in 1973 [14], and more recently analyzed by Pitkäranta [27]. This is, however, not more the case when the methods are viewed as a stabilized formulations. Then, they arise from a very systematic approach, cf. [17, ?, 22] and the presentation below. Recentely we have used the same approach for designing methods for the Naghdi shell modell in a bending dominated state [12].

We have also been given the opportunity to implement our methods in the SHIPFEM code of the Ship Laboratory, Technical Research Center of Finland (cf. [23, 25]), which is greatfully acknowledged.

Recently, Lyly has observed [20] that our linear triangular element is equivalent to an ealier formulation (which from the outset looks different) given by Tessler and Hughes [31]. Later the formulation of Hughes and Taylor has been rediscovered by Xu, Aurichhio and Taylor [33, 30, 4] (in a form that is quite easily seen to be equivalent to that of Tessler and Hughes). The equivalent of the linear stabilized method and that of Xu et al. has independently been proved by Lovadina [19]. Due to this equivalence the analysis of [10, 20, 19] justifies these other metods.

In this paper we will first recall our methods and then derive the basic error estimates, which show that the methods converge optimally and independently of the relative thickness of the plate. In an engineering vocabulary we show that the elements are completely free from locking.

2 Notation and preliminaries

We consider the Reissner-Mindlin plate bending model and assume that the plate is clamped along its boundary. Denoting the midsurface of the plate by $\Omega \subset \mathbb{R}^2$, the variational problem is: find the deflection $w \in H_0^1(\Omega)$ and the rotation vector $\boldsymbol{\beta} \in [H_0^1(\Omega)]^2$ such that

$$Gt^3 a(\boldsymbol{\beta}, \boldsymbol{\eta}) + G\kappa t(\nabla w - \boldsymbol{\beta}, \nabla v - \boldsymbol{\eta}) = (f, v) \qquad \forall (v, \boldsymbol{\eta}) \in [H_0^1(\Omega)]^3. \tag{2.1}$$

Here G is the shear modulus and κ denotes the shear correction factor. f is the transverse load and t is the thickness of the plate. The bilinear form a is defined as

$$a(\boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{6}\{(\boldsymbol{\varepsilon}(\boldsymbol{\beta}), \boldsymbol{\varepsilon}(\boldsymbol{\eta})) + (\frac{\nu}{1-\nu})(\operatorname{div} \boldsymbol{\beta}, \operatorname{div} \boldsymbol{\eta})\}, \tag{2.2}$$

where $\boldsymbol{\varepsilon}(\cdot)$ is the small strain tensor and ν is the Poisson ratio. As usual, the L_2-inner products are denoted by $(\cdot, \cdot)_D$ and the corresponding norms by $\|\cdot\|_{0,D}$, with the subscript D dropped when $D = \Omega$.

The shear force $\boldsymbol{Q}$ and bending moment $\boldsymbol{M}$ are obtained from

$$\boldsymbol{Q} = G\kappa t(\nabla w - \boldsymbol{\beta}) \tag{2.3}$$

and

$$\boldsymbol{M} = \frac{Gt^3}{6}\{\boldsymbol{\varepsilon}(\boldsymbol{\beta}) + (\frac{\nu}{1-\nu})\mathrm{div}\,\boldsymbol{\beta}\ \boldsymbol{I}\}, \tag{2.4}$$

respectively.

For the theoretical analysis one assumes that the load is proportional to the third power of the plate thickness, i.e. $f = Gt^3 g$ with g fixed independent of t. With this assumption the problem (2.1) has a finite and non-trivial solution in limit when $t \to 0$ (cf. [9]). Hence, the problem becomes: find $(w, \boldsymbol{\beta}) \in [H_0^1(\Omega)]^3$ such that

$$a(\boldsymbol{\beta}, \boldsymbol{\eta}) + \kappa t^{-2}(\nabla w - \boldsymbol{\beta}, \nabla v - \boldsymbol{\eta}) = (g, v) \qquad \forall (v, \boldsymbol{\eta}) \in [H_0^1(\Omega)]^3. \tag{2.5}$$

Introducing the scaled shear force

$$\boldsymbol{q} = \kappa t^{-2}(\nabla w - \boldsymbol{\beta}) \tag{2.6}$$

as an independent unknown, the mixed form of (2.5) is: find $(w, \boldsymbol{\beta}, \boldsymbol{q}) \in [H_0^1(\Omega)]^3 \times [L_2(\Omega)]^2$ such that

$$\begin{aligned} a(\boldsymbol{\beta}, \boldsymbol{\eta}) + (\boldsymbol{q}, \nabla v - \boldsymbol{\eta}) &= (g, v) \qquad \forall (v, \boldsymbol{\eta}) \in [H_0^1(\Omega)]^3 \\ \kappa^{-1} t^2 (\boldsymbol{q}, \boldsymbol{s}) - (\nabla w - \boldsymbol{\beta}, \boldsymbol{s}) &= 0 \quad \forall \boldsymbol{s} \in [L_2(\Omega)]^2. \end{aligned} \tag{2.7}$$

The strong form corresponding to this system is obtained by integrating by parts:

$$\begin{aligned} \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{q} &= 0 \quad \text{in } \Omega, & (2.8)\\ -\mathrm{div}\,\boldsymbol{q} &= g \quad \text{in } \Omega, & (2.9)\\ -\kappa^{-1} t^2 \boldsymbol{q} + \nabla w - \boldsymbol{\beta} &= 0 \quad \text{in } \Omega, & (2.10)\\ w &= 0 \quad \text{on } \partial\Omega, & (2.11)\\ \boldsymbol{\beta} &= 0 \quad \text{on } \partial\Omega. & (2.12) \end{aligned}$$

Above the differential operator $\boldsymbol{L}$ is defined through

$$\boldsymbol{L}\boldsymbol{\eta} = \frac{1}{6}\,\mathrm{div}\,\{\boldsymbol{\varepsilon}(\boldsymbol{\eta}) + (\frac{\nu}{1-\nu})\mathrm{div}\,\boldsymbol{\eta}\ \boldsymbol{I}\}, \tag{2.13}$$

where **div** stands for the divergence of second order tensors and $\boldsymbol{I}$ is the unit tensor.

3 The finite element methods

We let $\mathcal{C}_h$ be the finite element partitioning of $\bar{\Omega}$ into triangles or convex quadrilaterals and define the finite element subspaces for the deflection and rotation vector with the index $k \geq 1$ as

$$\begin{aligned} W_h &= \{v \in H_0^1(\Omega) \mid v_{|K} \in R_k(K),\ \forall K \in \mathcal{C}_h\}, & (3.1)\\ \boldsymbol{V}_h &= \{\boldsymbol{\eta} \in [H_0^1(\Omega)]^2 \mid \boldsymbol{\eta}_{|K} \in [R_k(K)]^2,\ \forall K \in \mathcal{C}_h\}, & (3.2) \end{aligned}$$

where $R_k(K)$ is a space of polynomials of degree $\leq k$ defined on K. We point out that this means that equal basis functions are used for the deflection and both components of the rotation.

The shear energy will be modified by interpolating with the MITC technique. For an element $K \in \mathcal{C}_h$ an auxiliary space $\boldsymbol{\Gamma}_k(K)$ and an MITC reduction operator $\boldsymbol{R}_h : [H^1(K)]^2 \to \boldsymbol{\Gamma}_k(K)$ are introduced.

The finite element methods for the problem (2.1) are then defined as follows: find $(w_h, \boldsymbol{\beta}_h) \in W_h \times \boldsymbol{V}_h$ such that

$$\mathcal{B}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) = (f, v) \quad \forall (v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h. \tag{3.3}$$

The bilinear form $\mathcal{B}_h$ is defined as

$$\begin{aligned} \mathcal{B}_h(w, \boldsymbol{\beta}; v, \boldsymbol{\eta}) &= Gt^3 a(\boldsymbol{\beta}, \boldsymbol{\eta}) - \alpha \sum_{K \in \mathcal{C}_h} h_K^2 (\boldsymbol{L}\boldsymbol{\beta}, \boldsymbol{L}\boldsymbol{\eta})_K \\ &+ \sum_{K \in \mathcal{C}_h} \Big(\frac{G\kappa t^3}{t^2 + \kappa \alpha h_K^2} \Big) (\nabla w - \boldsymbol{R}_h \boldsymbol{\beta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}, \nabla v - \boldsymbol{R}_h \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K. \end{aligned} \tag{3.4}$$

Here h_K denotes the diameter of the element $K \in \mathcal{C}_h$ and α is a positive constant for which an upper bound will be defined below.

The different methods will then be defined by specifying the spaces $R_k(K)$ and $\boldsymbol{\Gamma}_k(K)$ together with the reduction operator $\boldsymbol{R}_h$.

From the solution $(w_h, \boldsymbol{\beta}_h)$ to (3.3), the approximations for the shear force and bending moment are obtained from

$$\boldsymbol{Q}_{h|K} = \Big(\frac{G\kappa t^3}{t^2 + \kappa \alpha h_K^2} \Big) (\nabla w_h - \boldsymbol{R}_h \boldsymbol{\beta}_h - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}_h)_{|K} \quad \forall K \in \mathcal{C}_h \tag{3.5}$$

and

$$\boldsymbol{M}_h = \frac{Gt^3}{6} \{ \boldsymbol{\varepsilon}(\boldsymbol{\beta}_h) + (\frac{\nu}{1-\nu}) \operatorname{div} \boldsymbol{\beta}_h \, \boldsymbol{I} \}, \tag{3.6}$$

respectively. An alternative way to determine the approximate shear force is to calculate it through the equilibrium equation

$$\boldsymbol{Q}_{h|K} = -Gt^3 \boldsymbol{L}\boldsymbol{\beta}_{h|K}, \quad \forall K \in \mathcal{C}_h. \tag{3.7}$$

This is, of course, reasonable only when $k \geq 2$.

Next, let us define the different methods.

Method I

We let K be a triangle, $R_k(K) = P_k(K)$ with $k \geq 1$ and denote by

$$\boldsymbol{\Gamma}_k(K) = [P_{k-1}(K)]^2 \oplus (y, -x) \tilde{P}_{k-1}(K), \tag{3.8}$$

the rotated Raviart-Thomas space [28]. Here $\tilde{P}_{k-1}(K)$ is the space of homogeneous polynomials of degree $k-1$. The reduction operator is defined through the conditions

$$\int_E [(\boldsymbol{R}_h \boldsymbol{\eta} - \boldsymbol{\eta}) \cdot \boldsymbol{\tau}] v \, ds = 0, \ \forall v \in P_{k-1}(E), \quad \text{for every edge } E \text{ of } K, \tag{3.9}$$

and for $k \geq 2$

$$\int_K (\boldsymbol{R}_h\boldsymbol{\eta} - \boldsymbol{\eta}) \cdot \boldsymbol{\tau}\, dx\, dy = 0,\ \forall \boldsymbol{\tau} \in [P_{k-2}(K)]^2. \tag{3.10}$$

Above $\boldsymbol{\tau}$ is the tangent to the edge E.

Remark 3.1 For linear elements with $k = 1$ it holds

$$\boldsymbol{L}\boldsymbol{\eta}_{|K} = 0,\ \ \forall K \in \mathcal{C}_h,\ \ \forall \boldsymbol{\eta} \in \boldsymbol{V}_h, \tag{3.11}$$

and so the bilinear form $\mathcal{B}_h$ reduces to

$$\mathcal{B}_h(w, \boldsymbol{\beta}; v, \boldsymbol{\eta}) = Gt^3 a(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{K \in \mathcal{C}_h} \Big(\frac{G\kappa t^3}{t^2 + \kappa\alpha h_K^2}\Big)(\nabla w - \boldsymbol{R}_h\boldsymbol{\beta}, \nabla v - \boldsymbol{R}_h\boldsymbol{\eta})_K. \tag{3.12}$$

This gives our linear element (introduced in [10]), which is equivalent to the elements of Tessler-Hughes [31] and Xu et al. [33, 30, 4]. Taking $\alpha = 0$ we get an unstable element introduced by Hughes and Taylor [18]. In the above mentioned papers we have not found any remark showing the near relationship between this element and the elements later considered by the same authors. ■

Remark 3.2 The MITC7 element [6] is obtained from Method I by choosing $\alpha = 0$, $k = 2$, and taking $R_2(K) = P_2(K) \oplus \text{span}\{\lambda_1\lambda_2\lambda_3\}$ in the rotation space $\boldsymbol{V}_h$. Here λ_i, $i = 1, 2, 3$, denote the barycentric coordinates of K. ■

Remark 3.3 With $\alpha = 0$ and $k = 2$ one obtains an element proposed in [26]. The element is unfortunately not optimally convergent. ■

Method II

Now K is a quadrilateral and $R_k(K) = Q_k(K)$ with $k \geq 1$. We let $\boldsymbol{J}_K$ be the Jacobian matrix of the mapping $\boldsymbol{F}_K : \hat{K} \to K$ ($\hat{K}$ is the unit square with coordinates ξ and η) and define

$$\boldsymbol{\Gamma}_k(K) = \{\boldsymbol{\eta} \mid \boldsymbol{\eta} = \boldsymbol{J}_K^{-T}\hat{\boldsymbol{\eta}} \circ \boldsymbol{F}_K^{-1},\ \hat{\boldsymbol{\eta}} \in \boldsymbol{\Gamma}_k(\hat{K})\}, \tag{3.13}$$

where $\boldsymbol{J}_K^{-T}$ is the transpose of $\boldsymbol{J}_K^{-1}$, and

$$\boldsymbol{\Gamma}_k(\hat{K}) = P_{k-1,k}(\hat{K}) \times P_{k,k-1}(\hat{K}). \tag{3.14}$$

This is the rectangular rotated Raviart-Thomas space with

$$P_{m,n}(\hat{K}) = \{v \mid v = \sum_{i=0}^{m}\sum_{j=0}^{n} a_{ij}\xi^i\eta^j \text{ for some } a_{ij} \in \mathbb{R}\}. \tag{3.15}$$

The reduction operator $\boldsymbol{R}_h : [H^1(K)]^2 \to \boldsymbol{\Gamma}_k(K)$ is now defined through

$$\boldsymbol{R}_h\boldsymbol{\eta} = \boldsymbol{J}_K^{-T}\boldsymbol{R}_{\hat{K}}\boldsymbol{J}_K^T\boldsymbol{\eta}, \tag{3.16}$$

where $\boldsymbol{R}_{\hat{K}} : [H^1(\hat{K})]^2 \to \boldsymbol{\Gamma}_k(\hat{K})$ is an operator satisfying the conditions

$$\int_{\hat{E}} [(\boldsymbol{R}_{\hat{K}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}) \cdot \boldsymbol{\tau}] v \, ds = 0, \ \forall v \in P_{k-1}(\hat{E}), \quad \text{for every edge } \hat{E} \text{ of } \hat{K}, \tag{3.17}$$

and in the case if $k \geq 2$

$$\int_{\hat{K}} (\boldsymbol{R}_{\hat{K}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}) \cdot \boldsymbol{r} \, d\xi \, d\eta = 0, \quad \forall \boldsymbol{r} \in P_{k-1,k-2}(\hat{K}) \times P_{k-2,k-1}(\hat{K}). \tag{3.18}$$

Remark 3.4 If $k = 1$ it is possible use the reduced bilinear form (3.12). By doing this we get the stabilized MITC4 element [24], and if we further choose $\alpha = 0$ we obtain the original MITC4 element of Bathe and Dvorkin [8]. ■

Method III

Again, K is a quadrilateral but now we choose $R_k(K) = Q_k^r(K) = Q_k(K) \cap P_{k+1}(K)$ (isoparametric) with $k \geq 1$. For this method we define

$$\boldsymbol{\Gamma}_k(\hat{K}) = [P_k(\hat{K})]^2 \setminus \text{span}\{(\xi^k, 0), (0, \eta^k)\}, \tag{3.19}$$

which is the rotated rectangular Brezzi-Douglas-Fortin-Marini (BDFM) space [11]. The operator $\boldsymbol{R}_h$ is defined as in (3.16) with $\boldsymbol{R}_{\hat{K}}$ satisfying

$$\int_{\hat{E}} [(\boldsymbol{R}_{\hat{K}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}) \cdot \boldsymbol{\tau}] v \, ds = 0, \ \forall v \in P_{k-1}(\hat{E}) \quad \text{for every edge } \hat{E} \text{ of } \hat{K}, \tag{3.20}$$

and for $k \geq 2$

$$\int_{\hat{K}} (\boldsymbol{R}_{\hat{K}}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}) \cdot \boldsymbol{r} \, d\xi \, d\eta = 0, \ \forall \boldsymbol{r} \in [P_{k-2}(\hat{K})]^2. \tag{3.21}$$

Remark 3.5 The MITC9 element [6] is obtained from Method III by taking $\alpha = 0$, $k = 2$ and $R_2(K) = Q_2(K)$ in the rotation space $\boldsymbol{V}_h$. ■

Remark 3.6 For all three methods it holds

$$\boldsymbol{R}_h \nabla v = \nabla v, \quad \forall v \in W_h.$$

This property is used in analysis below. ■

4 Error analysis

As mentioned the error analysis should be done for the scaled problem (2.5). Without any loss of generality we can also choose $\kappa = 1$. Therefore we consider the scaled finite element formulation: find $(w_h, \boldsymbol{\beta}_h) \in W_h \times \boldsymbol{V}_h$ such that

$$\mathcal{S}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) = (g, v), \quad \forall (v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h, \tag{4.1}$$

with

$$\begin{aligned} \mathcal{S}_h(w, \boldsymbol{\beta}; v, \boldsymbol{\eta}) &= a(\boldsymbol{\beta}, \boldsymbol{\eta}) - \alpha \sum_{K \in \mathcal{C}_h} h_K^2 (\boldsymbol{L}\boldsymbol{\beta}, \boldsymbol{L}\boldsymbol{\eta})_K \\ &+ \sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} (\nabla w - \boldsymbol{R}_h \boldsymbol{\beta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}, \nabla v - \boldsymbol{R}_h \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K. \end{aligned} \tag{4.2}$$

The approximation to the scaled shear force (2.6) is then defined by

$$\boldsymbol{q}_{h|K} = (t^2 + \alpha h_K^2)^{-1} (\nabla w_h - \boldsymbol{R}_h \boldsymbol{\beta}_h - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}_h)_{|K}. \tag{4.3}$$

The aim is now to derive error estimates which are independent of the plate thickness. To this end C we will denote various positive constants which do not depend on the thickness t or the global mesh parameter

$$h = \max_{K \in \mathcal{C}_h} h_K. \tag{4.4}$$

We will use standard finite element notation with $|\cdot|_{m,D}$ and $\|\cdot\|_{m,D}$ denoting the seminorms and norms in $H^m(D)$ and $[H^m(D)]^2$. Again, the subscript D is dropped when $D = \Omega$.

Under some (minor) restrictive assumptions on the mesh (see [32]) we have the following result which states that the operator $\boldsymbol{R}_h$ has optimal interpolation properties.

Lemma 4.1 [28, 9] *There exist a positive constant C such that for $1 \leq m \leq k$ and $\boldsymbol{\eta} \in [H^m(K)]^2$ it holds*

$$\|\boldsymbol{\eta} - \boldsymbol{R}_h \boldsymbol{\eta}\|_{0,K} \leq C h_K^m \|\boldsymbol{\eta}\|_{m,K}, \ \forall K \in \mathcal{C}_h. \ \blacksquare$$

We will also make use of the following inverse estimate which is valid since the space $\boldsymbol{V}_h$ consists of piecewise polynomials (cf. e.g. [15]).

Lemma 4.2 *There exists a constant $C_I > 0$ such that*

$$C_I \sum_{K \in \mathcal{C}_h} h_K^2 \|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 \leq a(\boldsymbol{\eta}, \boldsymbol{\eta}), \quad \forall \boldsymbol{\eta} \in \boldsymbol{V}_h. \ \blacksquare$$

Remark 4.1 The constant C_I of Lemma 4.2 plays an important role, not only in the analysis of the methods, but also in numerical calculations. Hence, we refer to [16] where numerical techniques for estimating constants like C_I have been considered. ■

The stability will be formulated using the following mesh dependent seminorm and norm:

$$|(v,\boldsymbol{\eta})|_h = \Big(\sum_{K\in\mathcal{C}_h} (t^2+\alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta})\|_{0,K}^2 \Big)^{1/2}, \tag{4.5}$$

$$|||(v,\boldsymbol{\eta})|||_h = \|v\|_1 + \|\boldsymbol{\eta}\|_1 + |(v,\boldsymbol{\eta})|_h. \tag{4.6}$$

We also define

$$\|\boldsymbol{q}\|_{-1,h} = (\sum_{K\in\mathcal{C}_h} h_K^2 \|\boldsymbol{q}\|_{0,K}^2)^{1/2}, \tag{4.7}$$

and note that the following equivalence holds.

Lemma 4.3 *There exists a positive constant C such that*

$$C|||(v,\boldsymbol{\eta})|||_h \le \|\boldsymbol{\eta}\|_1 + |(v,\boldsymbol{\eta})|_h \le |||(v,\boldsymbol{\eta})|||_h, \quad \forall (v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h.$$

Proof: The Poincaré inequality, Remark 3.6, Lemma 4.1 (with $m = 1$), and the inequality $(t^2 + \alpha h_K^2) \le C$ give

$$\begin{aligned}
\|v\|_1^2 &\le C\|\nabla v\|_0^2 = C\|\boldsymbol{R}_h \nabla v\|_0^2 \\
&\le C(\|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta})\|_0^2 + \|\boldsymbol{R}_h\boldsymbol{\eta}\|_0^2) \\
&\le C(\|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta})\|_0^2 + \|\boldsymbol{\eta}\|_1^2) \\
&\le C(\sum_{K\in\mathcal{C}_h} (t^2+\alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta})\|_{0,K}^2 + \|\boldsymbol{\eta}\|_1^2),
\end{aligned}$$

which proves the claim. ∎

With the aid of the previous auxiliary results we are now ready to prove that the methods are stable with respect to the norm $|||\cdot|||_h$.

Lemma 4.4 *There exists a constant $C > 0$ such that for $0 < \alpha < C_I$ it holds*

$$\mathcal{S}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) \ge C|||(v,\boldsymbol{\eta})|||_h^2, \quad \forall (v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h.$$

Proof: Using the inverse estimate of Lemma 4.2 and the Korn inequality we get

$$\begin{aligned}
\mathcal{S}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) &= a(\boldsymbol{\eta},\boldsymbol{\eta}) - \alpha \sum_{K\in\mathcal{C}_h} h_K^2 \|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 \\
&\quad + \sum_{K\in\mathcal{C}_h} (t^2+\alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 \\
&\ge (1 - \alpha C_I^{-1}) a(\boldsymbol{\eta},\boldsymbol{\eta}) + \sum_{K\in\mathcal{C}_h} (t^2+\alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 \\
&\ge C(\|\boldsymbol{\eta}\|_1^2 + \sum_{K\in\mathcal{C}_h} (t^2+\alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2).
\end{aligned} \tag{4.8}$$

The same inverse estimate and the boundedness of the bilinear form a also give

$$
\begin{aligned}
|(v,\boldsymbol{\eta})|_h^2 &= \sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}\|\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})\|_{0,K}^2 \qquad (4.9)\\
&\le C\Big(\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}\|\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+\alpha\sum_{K\in\mathcal{C}_h}h_K^2\|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2\Big)\\
&\le C\Big(\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}\|\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+a(\boldsymbol{\eta},\boldsymbol{\eta})\Big)\\
&\le C\Big(\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}\|\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+\|\boldsymbol{\eta}\|_1^2\Big).
\end{aligned}
$$

Combining (4.8), (4.9), and using Lemma 4.3 gives the desired result. ∎

Next, we note that in the bilinear form $\mathcal{S}_h$ is not consistent with the exact energy. In order to characterize the consistency error we define

$$
\begin{aligned}
\mathcal{E}_h(\boldsymbol{s};v,\boldsymbol{\eta}) &= (\boldsymbol{s},(\boldsymbol{R}_h-\boldsymbol{I})(\nabla v-\boldsymbol{\eta})) \qquad (4.10)\\
&\quad +t^2\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}(\boldsymbol{R}_h\boldsymbol{s}-\boldsymbol{s},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K.
\end{aligned}
$$

We then have

Lemma 4.5 *The solution* $(w,\boldsymbol{\beta})$ *to* (2.5) *satisfies*

$$
\mathcal{S}_h(w,\boldsymbol{\beta};v,\boldsymbol{\eta})=(g,v)+\mathcal{E}_h(\boldsymbol{q};v,\boldsymbol{\eta}),\quad \forall(v,\boldsymbol{\eta})\in H_0^1(\Omega)\times[H_0^1(\Omega)]^2.
$$

Proof: Using the constitutive relation (2.10) and the equilibrium equation (2.8), we get

$$
\begin{aligned}
\mathcal{S}_h(w,\boldsymbol{\beta};v,\boldsymbol{\eta}) &= a(\boldsymbol{\beta},\boldsymbol{\eta})-\alpha\sum_{K\in\mathcal{C}_h}h_K^2(\boldsymbol{L}\boldsymbol{\beta},\boldsymbol{L}\boldsymbol{\eta})_K\\
&\quad+\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}(\boldsymbol{R}_h(\nabla w-\boldsymbol{\beta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\beta},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K\\
&= a(\boldsymbol{\beta},\boldsymbol{\eta})+\alpha\sum_{K\in\mathcal{C}_h}h_K^2(\boldsymbol{q},\boldsymbol{L}\boldsymbol{\eta})_K\\
&\quad+\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}(t^2\boldsymbol{R}_h\boldsymbol{q}+\alpha h_K^2\boldsymbol{q},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K\\
&= a(\boldsymbol{\beta},\boldsymbol{\eta})+\alpha\sum_{K\in\mathcal{C}_h}h_K^2(\boldsymbol{q},\boldsymbol{L}\boldsymbol{\eta})_K+\sum_{K\in\mathcal{C}_h}(\boldsymbol{q},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K\\
&\quad+t^2\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}(\boldsymbol{R}_h\boldsymbol{q}-\boldsymbol{q},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K\\
&= a(\boldsymbol{\beta},\boldsymbol{\eta})+\sum_{K\in\mathcal{C}_h}(\boldsymbol{q},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta}))_K\\
&\quad+t^2\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}(\boldsymbol{R}_h\boldsymbol{q}-\boldsymbol{q},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K\\
&= a(\boldsymbol{\beta},\boldsymbol{\eta})+(\boldsymbol{q},\nabla v-\boldsymbol{\eta})+(\boldsymbol{q},(\boldsymbol{R}_h-\boldsymbol{I})(\nabla v-\boldsymbol{\eta}))\\
&\quad+t^2\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)^{-1}(\boldsymbol{R}_h\boldsymbol{q}-\boldsymbol{q},\boldsymbol{R}_h(\nabla v-\boldsymbol{\eta})-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K\\
&= (g,v)+\mathcal{E}_h(\boldsymbol{q};v,\boldsymbol{\eta}).\ \blacksquare
\end{aligned}
$$

Remark 4.2 Note that if we choose $\boldsymbol{R}_h = \boldsymbol{I}$, we get a pure consistent formulation:

$$\mathcal{S}_h(w, \boldsymbol{\beta}; v, \boldsymbol{\eta}) = (g, v), \quad \forall (v, \boldsymbol{\eta}) \in H_0^1(\Omega) \times [H_0^1(\Omega)]^2. \tag{4.11}$$

A family of methods of this kind has been introduced and analyzed in [29]. The only drawback of these consistent methods is that higher degree (i.e. $k+1$) shape functions must be used for the deflection in order to obtain the right balance between the approximation properties of W_h and $\boldsymbol{V}_h$. ■

The following auxiliary result is needed in estimating the consistency error.

Lemma 4.6 *For $\boldsymbol{s} \in [H^{k-1}(\Omega)]^2$ it holds*

$$|(\boldsymbol{s}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta})| \leq Ch^k \|\boldsymbol{s}\|_{k-1} \|\boldsymbol{\eta}\|_1, \quad \forall \boldsymbol{\eta} \in [H^1(\Omega)]^2.$$

Proof: If $k = 1$ the result follow directly from the Schwartz inequality and Lemma 4.1.

For $k \geq 2$, we let $\boldsymbol{P}_K : [L^2(\hat{K})]^2 \rightarrow [L^2(K)]^2$ be the Piola transformation defined through [9, page 97]

$$\boldsymbol{P}_K \hat{\boldsymbol{s}} = |\boldsymbol{J}_K|^{-1} \boldsymbol{J}_K \hat{\boldsymbol{s}}, \quad \hat{\boldsymbol{s}} \in [L^2(\hat{K})]^2, \tag{4.12}$$

and define the space $\boldsymbol{S}(\hat{K})$ by

$$\boldsymbol{S}(\hat{K}) = \begin{cases} [P_{k-2}(\hat{K})]^2 & \text{for Methods I and III,} \\ P_{k-1,k-2}(\hat{K}) \times P_{k-2,k-1}(\hat{K}) & \text{for Method II,} \end{cases} \tag{4.13}$$

Using the definition of the operator $\boldsymbol{R}_h$ and the properties (3.10), (3.18) and (3.21) we then get

$$\begin{aligned} &(\boldsymbol{P}_K \hat{\boldsymbol{s}}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta})_K \\ &= \int_K (\boldsymbol{P}_K \hat{\boldsymbol{s}})(x,y) \cdot (\boldsymbol{\eta}(x,y) - \boldsymbol{R}_h\boldsymbol{\eta}(x,y)) dxdy \\ &= \int_{\hat{K}} |\boldsymbol{J}_K|^{-1} \boldsymbol{J}_K \hat{\boldsymbol{s}}(\xi,\eta) \cdot (\hat{\boldsymbol{\eta}}(\xi,\eta) - \boldsymbol{J}_K^{-T} \boldsymbol{R}_{\hat{K}} \boldsymbol{J}_K^{-1} \hat{\boldsymbol{\eta}}(\xi,\eta)) |\boldsymbol{J}_K| d\xi d\eta \\ &= \int_{\hat{K}} \hat{\boldsymbol{s}}(\xi,\eta) \cdot (\boldsymbol{J}_K^{-1} \hat{\boldsymbol{\eta}}(\xi,\eta) - \boldsymbol{R}_{\hat{K}} \boldsymbol{J}_K^{-1} \hat{\boldsymbol{\eta}}(\xi,\eta)) d\xi d\eta \\ &= 0, \qquad \forall \hat{\boldsymbol{s}} \in \boldsymbol{S}(\hat{K}). \end{aligned} \tag{4.14}$$

Next, we let $\boldsymbol{\Pi}_{\hat{K}} : [L_2(\hat{K})]^2 \rightarrow \boldsymbol{S}(\hat{K})$ be the L_2-projection and define the mapping $\boldsymbol{\Pi}_K$ through

$$\boldsymbol{\Pi}_K = \boldsymbol{P}_K \boldsymbol{\Pi}_{\hat{K}} \boldsymbol{P}_K^{-1}. \tag{4.15}$$

By using standard techniques [13, 9] for deriving interpolation estimate we get

$$\|\boldsymbol{s} - \boldsymbol{\Pi}_K \boldsymbol{s}\|_{0,K} \leq Ch_K^k \|\boldsymbol{s}\|_{k-1,K}. \tag{4.16}$$

Using (4.14) we then have

$$\begin{aligned} (\boldsymbol{s}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta})_K &= (\boldsymbol{s} - \boldsymbol{\Pi}_K \boldsymbol{s}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta})_K \\ &\leq \|\boldsymbol{s} - \boldsymbol{\Pi}_K \boldsymbol{s}\|_{0,K} \|\boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta}\|_{0,K} \leq Ch_K^k \|\boldsymbol{s}\|_{k-1,K} \|\boldsymbol{\eta}\|_{1,K}. \end{aligned} \tag{4.17}$$

The desired result follows from (4.17) by summing over the elements $K \in \mathcal{C}_h$. ■

For the consistency error we now have the following result.

Lemma 4.7 *Suppose that the exact shear force satisfies* $\boldsymbol{q} \in [H^{k-1}(\Omega)]^2$ *and* $t\boldsymbol{q} \in [H^k(\Omega)]^2$. *Then it holds*

$$|\mathcal{E}_h(\boldsymbol{q}; v, \boldsymbol{\eta})| \le Ch^k(\|\boldsymbol{q}\|_{k-1} + t\|\boldsymbol{q}\|_k)\ |||(v, \boldsymbol{\eta})|||_h,\ \forall (v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h.$$

Proof: We first note that the boundedness of the bilinear form a together with Lemmas 4.2 and 4.1 imply

$$\begin{aligned}
&(\sum_{K\in\mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1}\|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2)^{1/2} \qquad (4.18)\\
&\quad \le C(\sum_{K\in\mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1}\|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta})\|_{0,K}^2 + \alpha \sum_{K\in\mathcal{C}_h} h_K^2\|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2)^{1/2}\\
&\quad \le C(|(v, \boldsymbol{\eta})|_h^2 + \alpha C_I^{-1} a(\boldsymbol{\eta}, \boldsymbol{\eta}))^{1/2} \le C(|(v, \boldsymbol{\eta})|_h^2 + \|\boldsymbol{\eta}\|_1^2)^{1/2}.
\end{aligned}$$

Hence, we get (using Lemmas 4.6 and 4.1)

$$\begin{aligned}
\mathcal{E}_h(\boldsymbol{q}; v, \boldsymbol{\eta}) &= (\boldsymbol{q}, (\boldsymbol{R}_h - \boldsymbol{I})(\nabla v - \boldsymbol{\eta})) \qquad (4.19)\\
&\quad + t^2 \sum_{K\in\mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1}(\boldsymbol{R}_h\boldsymbol{q} - \boldsymbol{q}, \boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K\\
&= (\boldsymbol{q}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta}) + t^2 \sum_{K\in\mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1}(\boldsymbol{R}_h\boldsymbol{q} - \boldsymbol{q}, \boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K\\
&\le (\boldsymbol{q}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta}) + Ct^2(\sum_{K\in\mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1}\|\boldsymbol{q} - \boldsymbol{R}_h\boldsymbol{q}\|_{0,K}^2)^{1/2}(|(v, \boldsymbol{\eta})|_h^2 + \|\boldsymbol{\eta}\|_1^2)^{1/2}\\
&\le (\boldsymbol{q}, \boldsymbol{\eta} - \boldsymbol{R}_h\boldsymbol{\eta}) + Ct\|\boldsymbol{q} - \boldsymbol{R}_h\boldsymbol{q}\|_0(|(v, \boldsymbol{\eta})|_h^2 + \|\boldsymbol{\eta}\|_1^2)^{1/2}\\
&\le Ch^k(\|\boldsymbol{q}\|_{k-1} + t\|\boldsymbol{q}\|_k)|||(v, \boldsymbol{\eta})|||_h. \quad \blacksquare
\end{aligned}$$

For the rest of the error analysis we will next define a special interpolation operator $I_h : H_0^1(\Omega) \to W_h$ through the following three conditions:

$$((v - I_h v) \circ \boldsymbol{F}_K)(\hat{p}) = 0,\ \forall \text{ vertices } \hat{p} \text{ of } \hat{K}, \qquad (4.20)$$

$$\int_{\hat{E}} ((v - I_h v) \circ \boldsymbol{F}_K)\hat{r}\ d\hat{s} = 0,\ \forall \hat{r} \in P_{k-2}(\hat{E}),\ \forall \text{ edges } \hat{E} \text{ of } \hat{K}, \qquad (4.21)$$

and

$$\int_{\hat{K}} ((v - I_h v) \circ \boldsymbol{F}_K)\hat{s}\ d\xi d\eta = 0,\ \begin{cases} \forall \hat{s} \in P_{k-3}(\hat{K}), & \text{for the Methods I and III,} \\ \forall \hat{s} \in Q_{k-2}(\hat{K}), & \text{for the Method II,} \end{cases} \qquad (4.22)$$

for every element $K \in \mathcal{C}_h$.

The operator I_h has optimal interpolation properties:

Lemma 4.8 *There exists a positive constant* C *such that for* $v \in H^m(\Omega)$ *and* $1 \le m \le k + 1$ *it holds*

$$\|v - I_h v\|_s \le Ch^{m-s}\|v\|_m,\ s = 0, 1.$$

Proof: Clearly I_h is a polynomial preserving operator in the sense that $I_h v = v$, $\forall v \in W_h$. Hence, we can deduce the asserted estimate from [13, Sections 15 and 16]. $\blacksquare$

The reason for introducing the operator I_h is the following technical result.

Lemma 4.9 *For $v \in H_0^1(\Omega)$ it holds*

$$\boldsymbol{R}_h \nabla (v - I_h v) = \boldsymbol{0}.$$

Proof: On each element $K \in \mathcal{C}_h$ we have (using (4.20) and (4.21))

$$\begin{aligned}\int_{\hat{E}} \hat{\nabla}((v - I_h v) \circ \boldsymbol{F}_K) \cdot \hat{\boldsymbol{\tau}}\ \hat{r}\ d\hat{s} &= \int_{\hat{E}} \frac{\partial}{\partial \hat{s}}((v - I_h v) \circ \boldsymbol{F}_K)\ \hat{r}\ d\hat{s} \\ &= \int_{\partial \hat{E}} ((v - I_h v) \circ \boldsymbol{F}_K)\hat{r} - \int_{\hat{E}} ((v - I_h v) \circ \boldsymbol{F}_K)\frac{\partial \hat{r}}{\partial \hat{s}}\ d\hat{s} = 0, \quad (4.23)\end{aligned}$$

for every edge $\hat{E}$ of $\hat{K}$ if $\hat{r} \in P_{k-1}(\hat{E})$ and (using (4.21) and (4.22))

$$\begin{aligned}\int_{\hat{K}} \hat{\nabla}((v - I_h v) \circ \boldsymbol{F}_K) \cdot \hat{\boldsymbol{s}}\ d\xi d\eta &= \int_{\partial \hat{K}} ((v - I_h v) \circ \boldsymbol{F}_K)\hat{\boldsymbol{s}} \cdot \hat{\boldsymbol{n}}\ d\hat{s} \\ &- \int_{\hat{K}} ((v - I_h v) \circ \boldsymbol{F}_K)\ \widehat{\operatorname{div}} \hat{\boldsymbol{s}}\ d\xi d\eta = 0, \quad (4.24)\end{aligned}$$

if $k \geq 2$ and $\hat{\boldsymbol{s}} \in \boldsymbol{S}(\hat{K})$. (See Lemma 4.6 for the definition of the space $\boldsymbol{S}(\hat{K})$). Here $\hat{\nabla}$ and $\widehat{\operatorname{div}}$ stand for the gradient and divergence operators with respect to the ξ and η variables of $\hat{K}$ and $\hat{\boldsymbol{n}}$ is the unit outward normal to $\partial \hat{K}$.

Hence, using (4.23), (4.24), and recalling the definition of the operator $\boldsymbol{R}_{\hat{K}}$, we get

$$\boldsymbol{R}_{\hat{K}} \hat{\nabla}((v - I_h v) \circ \boldsymbol{F}_K) = \boldsymbol{0}, \ \forall K \in \mathcal{C}_h, \quad (4.25)$$

and since it holds $\boldsymbol{J}_K^{-1}(\nabla v \circ \boldsymbol{F}_K) = \hat{\nabla}(v \circ \boldsymbol{F}_K), \ \forall K \in \mathcal{C}_h$, we conclude that

$$\begin{aligned}\boldsymbol{R}_h \nabla (v - I_h v)_{|K} &= \boldsymbol{J}_K^{-T} \boldsymbol{R}_{\hat{K}} \boldsymbol{J}_K^{-1} \nabla ((v - I_h v) \circ \boldsymbol{F}_K) \\ &= \boldsymbol{J}_K^{-T} \boldsymbol{R}_{\hat{K}} \hat{\nabla}((v - I_h v) \circ \boldsymbol{F}_K) = \boldsymbol{0}, \ \forall K \in \mathcal{C}_h. \ \blacksquare \quad (4.26)\end{aligned}$$

We will next state our main result.

Theorem 4.1 *Suppose that the solution to the problem (2.5) satisfies $w \in H^{k+1}(\Omega)$, $t\boldsymbol{\beta} \in [H^{k+2}(\Omega)]^2$ and $\boldsymbol{\beta} \in [H^{k+1}(\Omega)]^2$. For $0 < \alpha < C_I$ it then holds*

$$\|w - w_h\|_1 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_1 + \|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,h} + t\|\boldsymbol{q} - \boldsymbol{q}_h\|_0 \leq Ch^k(\|w\|_{k+1} + t\|\boldsymbol{\beta}\|_{k+2} + \|\boldsymbol{\beta}\|_{k+1}).$$

Proof: Let $\tilde{\boldsymbol{\beta}} \in \boldsymbol{V}_h$ be the usual Lagrange interpolant to $\boldsymbol{\beta}$ and $\tilde{w} = I_h w \in W_h$ the interpolant to w. From Lemmas 4.4 and 4.5, there exists a pair $(v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h$ such that

$$|||(v, \boldsymbol{\eta})|||_h \leq C, \quad (4.27)$$

and

$$\begin{aligned}|||(w_h - \tilde{w}, \boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}})|||_h &\leq \mathcal{S}_h(w_h - \tilde{w}, \boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}; v, \boldsymbol{\eta}) \\ &= \mathcal{S}_h(w - \tilde{w}, \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}; v, \boldsymbol{\eta}) - \mathcal{E}_h(\boldsymbol{q}; v, \boldsymbol{\eta}). \quad (4.28)\end{aligned}$$

For the consistency error term in (4.28) we directly obtain (using (4.27), (2.8) and Lemma 4.7)

$$|\mathcal{E}_h(\boldsymbol{q}; v, \boldsymbol{\eta})| \leq Ch^k(\|\boldsymbol{q}\|_{k-1} + t\|\boldsymbol{q}\|_k) \leq Ch^k(\|\boldsymbol{\beta}\|_{k+1} + t\|\boldsymbol{\beta}\|_{k+2}). \tag{4.29}$$

Next, let us write out the bilinear form $\mathcal{S}_h$ on the the right hand side of (4.28). Due to the definition of $\tilde{w}$ we have (using Lemma 4.9)

$$\begin{aligned}\mathcal{S}_h(w - \tilde{w}, \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}; v, \boldsymbol{\eta}) &= a(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}) - \alpha \sum_{K \in \mathcal{C}_h} h_K^2 (\boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}), \boldsymbol{L}\boldsymbol{\eta})_K \\ &+ \sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} (\boldsymbol{R}_h(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - \alpha h_K^2 \boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}), \boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K.\end{aligned} \tag{4.30}$$

From (4.27) and Lemma 4.2 it follows that

$$(\sum_{K \in \mathcal{C}_h} h_K^2 \|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2)^{1/2} \leq C, \tag{4.31}$$

and

$$(\sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2)^{1/2} \leq C. \tag{4.32}$$

Hence, for the first and second terms in (4.30) we get (using (4.31), Lemma 4.2 and continuity of the bilinear form a)

$$a(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}) \leq C\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_1 \leq Ch^k\|\boldsymbol{\beta}\|_{k+1}, \tag{4.33}$$

and

$$\sum_{K \in \mathcal{C}_h} h_K^2 (\boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}), \boldsymbol{L}\boldsymbol{\eta})_K \leq C(\sum_{K \in \mathcal{C}_h} h_K^2 \|\boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\|_{0,K}^2)^{1/2} \leq Ch^k\|\boldsymbol{\beta}\|_{k+1}. \tag{4.34}$$

Using the same estimates and (4.32) we obtain for the third term

$$\begin{aligned}&\sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} (\boldsymbol{R}_h(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - \alpha h_K^2 \boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}), \boldsymbol{R}_h(\nabla v - \boldsymbol{\eta}) - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K \\ &\leq C(\sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \alpha h_K^2 \boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\|_{0,K}^2)^{1/2} \\ &\leq C(\sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} \|\boldsymbol{R}_h(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\|_{0,K}^2 + \alpha \sum_{K \in \mathcal{C}_h} h_K^2 \|\boldsymbol{L}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\|_{0,K}^2)^{1/2} \\ &\leq C(\sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} (\|(\boldsymbol{I} - \boldsymbol{R}_h)(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\|_{0,K}^2 + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_{0,K}^2) + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_1^2)^{1/2} \\ &\leq C(\sum_{K \in \mathcal{C}_h} (t^2 + \alpha h_K^2)^{-1} (h_K^2 \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_{1,K}^2 + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_{0,K}^2) + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_1^2)^{1/2} \\ &\leq Ch^k\|\boldsymbol{\beta}\|_{k+1}.\end{aligned} \tag{4.35}$$

The estimate

$$|||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_h \leq Ch^k(\|w\|_{k+1} + t\|\boldsymbol{\beta}\|_{k+2} + \|\boldsymbol{\beta}\|_{k+1}) \tag{4.36}$$

follows now by combining (4.29), (4.33)-(4.35), using the triangle inequality and the interpolation estimate (here we need Lemma 4.9)

$$|||(w-\tilde{w},\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})|||_h \leq Ch^k(\|w\|_{k+1}+\|\boldsymbol{\beta}\|_{k+1}). \tag{4.37}$$

After this, the H^1-estimates for the errors $w-w_h$ and $\boldsymbol{\beta}-\boldsymbol{\beta}_h$ follow directly from (4.36) and from the definition of the norm $|||\cdot|||_h$.

Next, let us derive the asserted estimates for the shear. Recalling the definitions (2.6) and (4.3) of the quantities $\boldsymbol{q}$ and $\boldsymbol{q}_h$, we get

$$\begin{aligned}(\boldsymbol{q}-\boldsymbol{q}_h)_{|K} = (t^2+\alpha h_K^2)^{-1}(&\boldsymbol{R}_h(\nabla(w-w_h)-(\boldsymbol{\beta}-\boldsymbol{\beta}_h))\\ &-\alpha h_K^2\boldsymbol{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_h)+t^2(\boldsymbol{q}-\boldsymbol{R}_h\boldsymbol{q}))_{|K},\ \forall K\in\mathcal{C}_h.\end{aligned} \tag{4.38}$$

From this it follows that

$$\begin{aligned}&\Big(\sum_{K\in\mathcal{C}_h}(t^2+\alpha h_K^2)\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,K}^2\Big)^{1/2}\\ &\quad\leq C(|(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|_h+\|\boldsymbol{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_h)\|_{-1,h}+t\|\boldsymbol{q}-\boldsymbol{R}_h\boldsymbol{q}\|_0).\end{aligned} \tag{4.39}$$

Now, since an inverse estimate, an interpolation estimate and the estimate for $|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_h$ imply that

$$\|\boldsymbol{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_h)\|_{-1,h}\leq Ch^k(\|w\|_{k+1}+t\|\boldsymbol{\beta}\|_{k+2}+\|\boldsymbol{\beta}\|_{k+1}), \tag{4.40}$$

both estimates for the shear follow from (4.39) and Lemma 4.1. ∎

For a quasiuniform mesh we get the following estimates for the shear approximations.

Corollary 4.1 *Suppose that the mesh is quasiuniform, i.e. such that $h_K\geq Ch$, $\forall K\in\mathcal{C}_h$. Then it follows from Theorem 4.1 that*

$$\|\boldsymbol{q}-\boldsymbol{q}_h\|_0+\|\boldsymbol{q}-\boldsymbol{q}_h^e\|_0\leq Ch^{k-1}(\|w\|_{k+1}+t\|\boldsymbol{\beta}\|_{k+2}+\|\boldsymbol{\beta}\|_{k+1}).\ \blacksquare \tag{4.41}$$

Let us close the paper with a final comment. The regularity assumptions stated in our theorems are almost never satisfied in practice as it is well known that the Reissner-Mindlin model give rise to strong boundary layers in the solution, cf. [1, 2, 3]. Our estimates show, however, that the methods converge optimally in the sense that the error in the finite element solution is of the same order of magnitude as the interpolation error.

References

[1] D.N. Arnold and R.S. Falk. A uniformly accurate finite element method for the Reissner-Mindlin plate. *SIAM J. Num. Anal.*, 26:1276–1290, 1989.

[2] D.N. Arnold and R.S. Falk. The boundary layer for the Reissner-Mindlin plate model. *SIAM J. Math. Anal.*, 21:10–40, 1990.

[3] D.N. Arnold and R.S. Falk. Asymptotic analysis of the boundary layer for the Reissner-Mindlin plate model. *SIAM J. Math. Anal.*, 27:486–514, 1996.

[4] F. Aurichhio and R.L. Taylor. A triangular thick plate finite element with an exact thin limit. *Finite Elements in Analysis and Design*, 19:57–68, 1995.

[5] K.J. Bathe. *Finite Element Procedures.* Prentice Hall, Engelwood Cliffs, NJ, 1996.

[6] K.J. Bathe, F. Brezzi, and S.W. Cho. The MITC7 and MITC9 plate elements. *Comput. Struct.*, 32:797–814, 1989.

[7] K.J. Bathe, F. Brezzi, and M. Fortin. Mixed-interpolated elements for Reissner-Mindlin plates. *Int. J. Num. Meths. Eng.*, 28:1787–1801, 1989.

[8] K.J. Bathe and E. Dvorkin. A four node plate bending element based on Mindlin-Reissner plate theory and mixed interpolation. *Int. J. Num. Meths. Eng.*, 21:367–383, 1985.

[9] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, 1991.

[10] F. Brezzi, M. Fortin, and R. Stenberg. Error analysis of mixed-interpolated elements for Reissner-Mindlin plates. *Mathematical Models and Methods in Applied Sciences*, 1:125–151, 1991.

[11] F. Brezzi, J. Douglas Jr., M. Fortin, and L.D. Marini. Efficient rectangular mixed finite elements in two and three space variables. M^2AN, 21:237–250, 1987.

[12] D. Chapelle and R. Stenberg. Stabilized finite element formulations for shells in a bending dominated state. INRIA Rapport de Recherche 2941, Juillet 1996. http://www.inria.fr/RRRT/RR-2941.html.

[13] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North - Holland, 1978.

[14] I. Fried and S.K. Yang. Triangular, nine-degrees-of-freedom, C^0 plate bending element of quadratic accuracy. *Quart. Appl. Math.*, 31:303–312, 1973.

[15] V. Girault and P.A. Raviart. *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms.* Springer-Verlag, 1986.

[16] T.J.R. Hughes. *The Finite Element Method. Linear Static and Dynamic Analysis.* Prentice-Hall, 1987.

[17] T.J.R. Hughes and L.P. Franca. A mixed finite element formulation for Reissner-Mindlin plate theory: Uniform convergence of all higher-order spaces. *Comp. Meths. Appl. Mech. Engrg.*, 67:223–240, 1988.

[18] T.J.R. Hughes and R. L. Taylor. The linear triangular plate bending element. In J. R. Whiteman, editor, *The Mathematics of Finite Elements and Applications IV. MAFELAP 1981*, pages 127–142. Academic Press, 1982.

[19] C. Lovadina. Analysis of a mixed finite element method for the Reissner-Mindlin plate problem. *Comp. Meths. Appl. Mech. Engrg.*, To appear.

[20] M. Lyly. On the connection between some linear triangular Reissner-Mindlin plate bending elements. *Numer. Math.*, To appear.

[21] M. Lyly and R. Stenberg. New three and four noded thick plate bending elements. *Rakenteiden Mekaniikka*, 27:3–29, 1994.

[22] M. Lyly and R. Stenberg. Stabilized MITC plate bending elements. In M. Papadrakakis and B.H.V. Topping, editors, *Advances in Finite Element Techniques*, pages 11– 16. Civil Comp Press, 1994. Also available from: http://www.solid.hut.fi/reports/index.html.

[23] M. Lyly and R. Stenberg. Theory manual of the stabilized MITC plate and flat shell elements of SHIPFEM. Technical Report VTT VAL B 81, Technical Research Centre of Finland, 1995.

[24] M. Lyly, R. Stenberg, and T. Vihinen. A stable bilinear element for Reissner-Mindlin plates. *Comp. Meths. in Appl. Mech. Engrg.*, 110:343–357, 1993.

[25] M.Lyly and R.Stenberg. Some NAFEMS benchmarks applied to the stabilized MITC plate and flat shell elements. Technical Report VTT VAL B 82, Technical Research Centre of Finland, 1995.

[26] E. Oñate, O.C. Zienkiewicz, B. Suarez, and R.L. Taylor. A general methodology for deriving shear constrained Reissner-Minldlin plate elements. *Int. J. Num. Meths. Eng.*, 33:345–367, 1992.

[27] J. Pitkäranta. Analysis of some low-order finite element schemes for Mindlin-Reissner and Kirchhoff plates. *Numer. Math.*, 53:237–254, 1988.

[28] P.A. Raviart and J.M. Thomas. A mixed finite element method for second order elliptic problems. In *Mathematical Aspects of the Finite Element Method. Lecture Notes in Math. 606*, pages 292–315. Springer-Verlag, 1977.

[29] R. Stenberg. A new finite element formulation for the plate bending problem. In P. Ciarlet, L. Trabucho, and J.M. Viano, editors, *Asymptotic Methods for Elastic Structures*, pages 209–221. Walter de Gruyter & Co., Berlin - New York, 1995.

[30] R.L. Taylor and F. Aurichhio. Linked interpolation for Reissner-Mindlin plate elements. *Int. J. Num. Meths. Engng.*, 36:3057–3066, 1993.

[31] A. Tessler and T.J.R. Hughes. A three-node Mindlin plate element with improved transverse shear. *Comp. Meths. Appl. Mech. Engng.*, 50:71–101, 1985.

[32] J.M. Thomas. Thèse d'Etat, Université Pierre et Marie Curie 1977.

[33] Z. Xu. A thick-thin triangular thick plate element. *Int. J. Num. Meths. Engng.*, 33:963–973, 1992.

A LAYERED FLAKING MODEL FOR ICE LOAD DETERMINATION

By

T. KÄRNÄ,
VTT Building Technology
P.O. Box 18071, FIN-02044 VTT, Finland
Fax: +358-9-456 7006;
Email: tuomo.karna@vtt.fi

ABSTRACT

This paper addresses problems of ice load determination in conditions where an ice floe acts on an offshore structure. The ice sheet is assumed to fail by crushing and flaking. Results of several test series are first used to get a new insight to the flaking failure of an ice sheet. Horizontal cracks emanating form the ice edge appear to be a central phenomenon to be considered. Tests have also revealed that a large amount of energy dissipates at the ice edge during a short period within the loading process. This phenomenon explains the rapid decay of the transient vibration, which has been seen both in laboratory and in full scale. To consider this phenomenon, a new layered flaking model is derived.

1. INTRODUCTION AND OBJECTIVES

The dynamic forces due to level ice action pose two kind of problems for the design of offshore structures. First, a steady state vibration of the structure may arise. In this case the effects of the exiting force are magnified within the structure. Second, experimental data shows that the exiting global ice force increases with the structural compliance. This phenomenon is associated with changes in the ice failure mode and can be seen as an increase in the correlation between the local forces.

Simple design rules do not provide reliable predictions of the ice effects in dynamic interaction conditions. Therefore, Määttänen [22] made theoretical and experimental

studies and presented a numerical interaction model for narrow structures. Based on further measurements of steady state vibrations, Kärnä & Turunen developed another model for narrow structures [12]. They proved [13] that if a steady state vibration arises, the velocity amplitude of the structure at the waterline is approximately the same as the ice velocity. Eranti [2] and Kärnä et al. [15, 18] developed a nonlinear numerical interaction model "PSSII" for compliant wide structures. Kajaste-Rudnitski [8] used a linear stochastic approach to study how the global load is influenced by the correlation of the local forces.

New series of laboratory and field tests have been conducted during the last ten years. The new experimental information is used in this report to derive an updated formulation for the PSSII program. The paper considers wide and compliant structures with a vertical wall against the ice edge. Ice failure mode at any cross section of the ice edge is assumed to be flaking with ice crushing. This failure mode yields the largest global ice loads. Experiments have shown that ice failure occurs sometimes nonsimultaneously and sometimes simultaneously at different points of the ice edge. The change in the failure mode seems to depend on the velocity of the ice floe and on the compliance of the structure. Our objective is to show that an appropriate simulation of the ice failure process yields an interaction model, which predicts the observed change from nonsimultaneous to simultaneous ice failure.

2. EXPERIMENTAL BACKGROUND

2.1 Ice failure

Määttänen [22, 23] measured ice-induced vibrations of narrow structures both in the field and in laboratory. Further field data on this phenomenon was collected by Nordlund et al. [24]. Jeffereys & Wright [5] reported on severe dynamic interactions between a wide offshore structure and drifting ice floes. In this case a typical interaction cycle contained a quasi-static loading phase that was followed by transient vibration of the structure. This is the most common dynamic interaction phenomenon and it has been tested extensively in laboratory [9, 14, 21, 26].

Dynamic ice-structure interaction phenomena occur in conditions where the rate of interaction is sufficiently high to cause a brittle ice failure by crushing and flaking. In this failure mode the ice edge has a stepwise wedge shape. Horizontal splits emanating from the ice edge have an essential effect on the ice failure process (Fig. 1). Taylor [27] conducted indentation tests on lake ice with a thickness of 0.5 m. The rate of indentation varied from 1 mm/s to 7 mm/s. This is the transitional region for ductile and brittle ice fracture. Under these conditions, series of horizontal splits developed in the ice sheet. As a result of the ice failure, the ice edge had a stepwise wedge shape. Field tests reported by Fransson [4] and Kawamura et al. [11] yielded similar failure modes at higher rates of indentation.

Indentation test made in laboratory [3, 6, 10, 14, 17, 21] have shown that the horizontal splits tend to grow parallel to the ice surface. An analysis by Tuhkuri [28] suggests that this is an essential feature of the crack growth in the condition considered here. The tests have shown also that most of the force between the structure and the wedge shaped ice edge is transmitted in the central layer of the ice sheet. Furthermore, the ice edge has a tendency to maintain the wedge shape during the continuous flaking process.

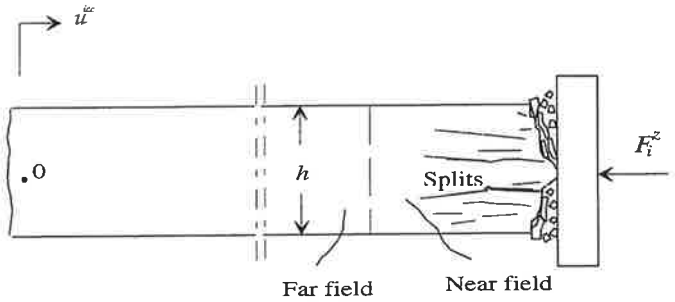


FIGURE 1: Near field zone with a wedge-shaped edge and horizontal splits.

2.2 Response of the structure

When a drifting ice flow acts on a structure, a continuous ice failure process may produce different kind of dynamic response modes within the structure. The main categories of structural response are: a *quasi static response* followed by transient vibration, *steady state vibration* and the *random response* to wide-band stochastic excitation.

Quasi-static response with transient vibration

Fig. 2 shows a typical result of an indentation test where a compliant and narrow structure was pushed into an ice sheet [18, 21]. A detail of the same test record is shown in Fig. 3. This test result shows that the acceleration and the associated mass forces are very small when the global ice force approaches its peak value during a major *loading phase*. A nearly static equilibrium exists between the internal and external forces acting on the structures at the events of maximum ice force. Therefore, we characterize the response as quasi-static. When the ice fails, the structure moves forward against the ice edge. This occurs as a transient vibration termed also as "spring-back". The first part of this period with a monotonously decreasing ice force is termed the *unloading phase*. If the structure is stiff the spring-back phase is often practically the same as the unloading phase. However, a compliant structure may continue its transient motion after a minimum load has been reached. In this case the global ice force remains at a lowe level.

An important feature of the transient is that it seems to stop abruptly after one or two cycles. This basic feature can be seen also in other reported measurements on compliant structures [5, 26]. Numerical simulations with a linear and proportional damping can not predict this feature [2, 12, 15, 18]. Therefore, we can anticipate that the ice edge provides an additional damping effect during the transient vibration. Indeed, Fig. 3 shows that a hysteretic loop appears in the force vs. displacement function. A detailed study shows that this hysteresis appears at the initial stage of a loading phase when the structure moves a little away from the ice edge. Therefore, we conclude that the additional damping effect is not related to the ice extrusion process, which takes place mainly during unloading. Another explanation will be proposed subsequently.

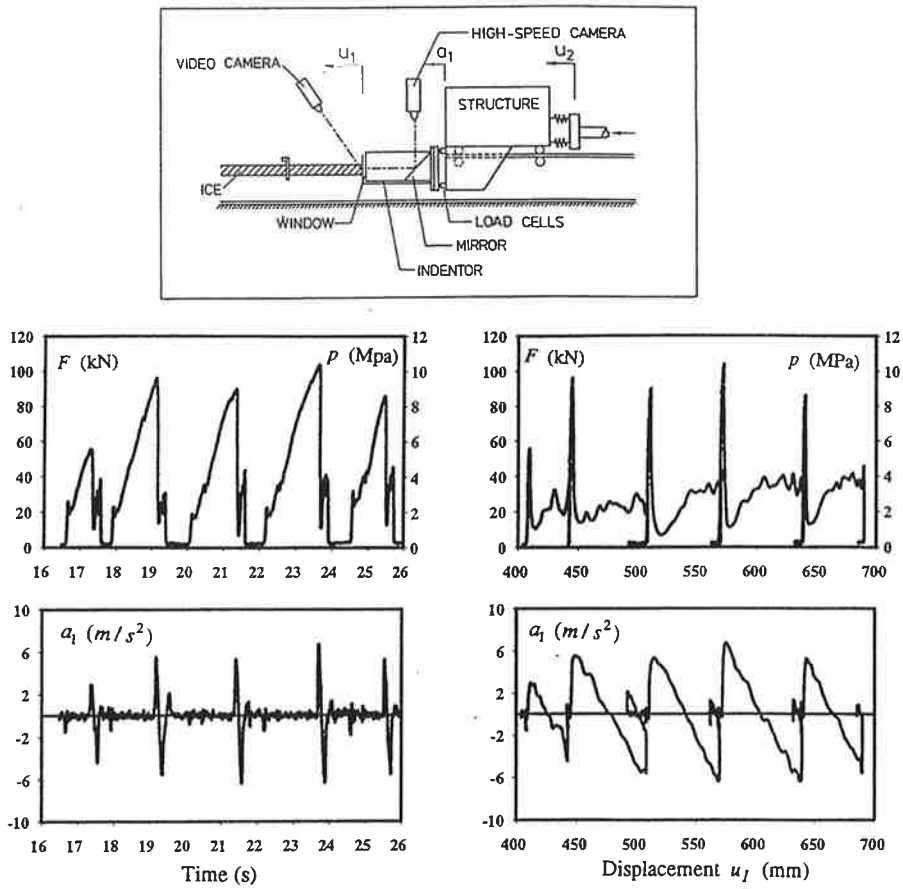


FIGURE 2: Quasi-static response with transient vibration [18,21].

Structural mass	M = 15 000 kg	Ice thickness	h = 100 mm
Spring stiffness	Ks = 2.4 kN/mm	Rate of indentation	v = 30 mm/s
Natural frequency	f = 2.0 Hz	Indentor width	D = 100 mm

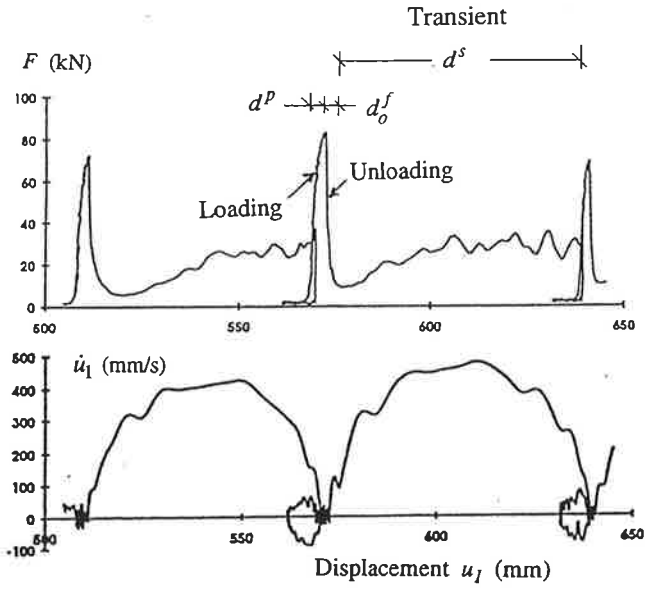


FIGURE 3: Details of the force and velocity signals for the test records shown in Fig. 2.

Other modes of response

The drifting ice field may occasionally induce *steady state vibration* in the structure. This kind of response was measured frequently on two channel markers in the Baltic Sea [12, 24]. The measured acceleration data shows that the time signal is smooth and almost sinusoidal. The effects of the ice force are magnified by the dynamics of the structure..

Kärnä et al. [17] analysed a test condition where a very stiff concrete cylinder was pushed against sea ice at high velocity. The measured time signal of the global force had a random character. The amplitudes of the force fluctuation were small compared to the mean level of the force. Similar results obtained by Sodhi [26] and Kamesaki et al, [9] suggest that this kind of force pattern is typical in conditions where the relative velocity between the structure and the ice edge is high. Tests with segmented indentors show that *random global forces* occur in conditions where the local forces fluctuate nonsimultaneously in front of the structure [4, 11, 26].

3. MODEL OF THE NEAR FIELD

3.1 Contact parameters

As depicted in Fig. 1, most of the damage at the ice edge occurs in the vicinity of the ice-structure contact surface. Therefore, the ice sheet is divided into a far field and a near field area (Fig. 4). The deformations of the far field are obtained from linear relationships [13,

16] and only the nonlinear behavior of the near field is discussed here. The ice failure process is controlled by the relative displacement and velocity between the ice sheet and the structure. To consider this, the near field is divided into a set of elements E_i, i = 1, ..., NS with a width B, length L and thickness h. The displacement vector of the structure's boundary is defined as

$$\mathbf{u}^z = (u_i^z), \quad i = 1, \ldots, NS \tag{1}$$

Similar definition applies for the displacements w_i^e at the boundary between the far field and the near field. The displacement of the ice field is denoted u^{ice} and φ_i is the direction of the ice motion relative to the normal direction of the ice edge. The vector of the compressive contact force is defined as $\mathbf{F}^z = \left(F_i^z\right)$.

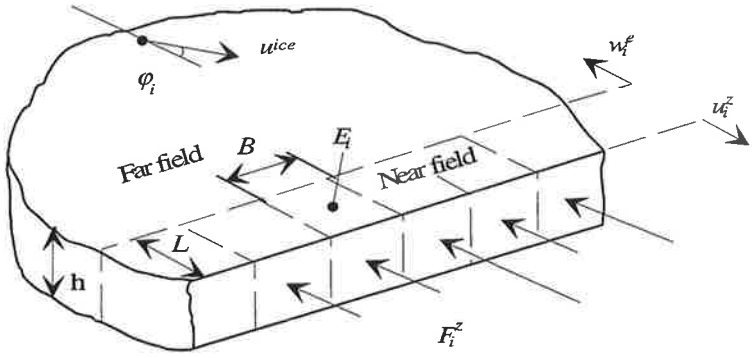


FIGURE 4: Near field and far field areas of the ice sheet.

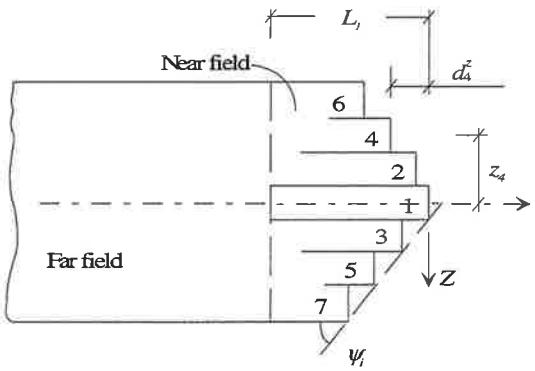


FIGURE 5: Model of the layered structure of the near field.

The preceding review of experimental data showed that brittle failure of the ice edge occurs frequently by flaking, which is associated with horizontal splits in the ice sheet. Therefore,

we will adopt a near field model described in Fig. 5. The near field elements E_i depicted in Fig. 4 are divided into horizontal layers ℓ_{ik} (i = 1,...,NS; k = 1,..., NDZ), which are bounded by the horizontal splits.

The geometry of the wedge shaped ice edge is defined by the wedge angel ψ_i and the cavities d^z between the structure and the layer. The wedge angle depends on the rate of interaction. Experimental records (Fig. 1) show that the longest horizontal splits appear close to the central plane of the ice sheet. Therefore, we assume that the length of the near field element E_i is the same as the length of the central layer ℓ_{ik} and is given by

$$L_1 = h \cot \psi_i \tag{2}$$

The other layers are assumed to be shorter as illustrated in Fig. 5. Details on the determination of the parameters ψ_I and d^z are given in [20].

3.2 Major and secondary flaking

Ice flaking can be either symmetric or asymmetric. These two flaking modes are termed here also as major and secondary flaking. We will now refer to the results of Saeki et al. [25] and Jones et al. [7] to give a plausible physical explanation for the differences between these flaking modes. These studies show that the coefficient of kinetic friction between ice and other materials is small at high sliding velocities (v > 10 mm/s) and high at low velocities (v < 0.1 mm/s).

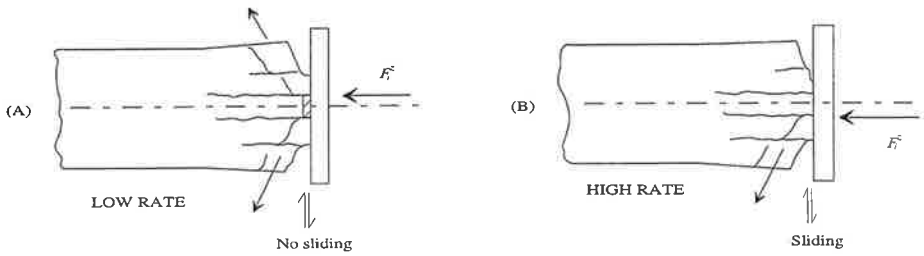


FIGURE 6: (A) Major flaking at low rate of interaction
(B) Secondary flaking at high rate.

The condition of an incipient major flaking event is shown in Fig. 6A. The force acting on the ice edge is usually eccentric. Due to the induced bending moment the ice sheet has a tendency to slide vertically along the contact surface. The actual amount of sliding depends on the friction forces at the ice-structure interface. At low rate of interaction the vertical sliding velocity is also small. Therefore, the friction force at the contact surface is large and

can prevent all sliding at the contact surface. High-speed photography of laboratory tests [21] shows that the failure starts as a rapid expansion in the middle level of the ice sheet. The flaking failure occurs symmetrically up and down.

At high rate of interaction the vertical sliding velocity is also high. Therefore, the frictional forces at the contact surface remain low and the structure does not provide confining forces to the ice edge and symmetric flaking is not likely to occur. As suggested in [1, 3, 6 and 28] we assume that asymmetric flaking depicted in Fig. 6B takes place at high rate [20].

3.3 Strength distribution

In the present model we assume that each layer fails due to the compressive force applied on it. The failure force of the layer ℓ_{ik} is determined as

$$F_{ik}^{p} = c(z_k)\,\varsigma_i\,\frac{h\,B}{NDZ}\,p_i^{cr} \tag{3}$$

where p_i^{cr} is the ice failure pressure at the near field element E_i. This parameter is obtained by adopting first a constant base value p^{cro} in accordance with the nominal contact area $A_i = h\,B$. A random number generator is then used to obtain log-normal distributed values for the failure pressure p_i^{cr}. The parameters ζ_i and $c(z_k)$ are used to consider two kind of strength distributions. First, ζ_i takes account for the differences between the local ice failure forces. Experiments show [18] that the failure strength assumes its lowest value in the middle of the structure. Second, the parameter $c(z_k)$ is used to consider the vertical strength distribution within the near field element. The central layers can sustain higher pressures than the outer layers. A simple cosine strength distribution is assumed [20].

4. ICE FORCE

4.1 The interaction process

A central variable in the simulation of the dynamic ice-structure interaction is the compressive displacement vector

$$\mathbf{w}^{dz} = (w_{ik}^{dz}), \qquad \begin{cases} i = 1,\ldots,NS \\ k = 1,\ldots,NDZ \end{cases} \tag{4}$$

which incorporates the relative near field displacements of the layers. The incremental updating equation for these displacement is

$$^{t+\Delta t}w_{ik}^{dz} = {}^{t}w_{ik}^{dz} + \Delta w_{ik}^{dz} \tag{5}$$

$$\Delta w_{ik}^{dz} = \Delta u^{ice} \cos\varphi_i - \Delta u_i^z - \Delta w_i^e \tag{6}$$

The initial condition is given by

$$^0 w_{ik}^{dz} = -g_i^{ap} \tag{7}$$

where g_i^{ap}, i = 1,...,NS is an initial gap between the ice edge and the structure. The local forces acting on the layers are determined as a function of the relative displacements,

$$F_{ik}^{dz} = F_{ik}^{dz}(w_{ik}^{dz}) \tag{8}$$

The compressive local forces acting on the near field elements E_i are calculated as

$$F_i^z = \sum_{k=1}^{NDZ} F_{ik}^{dz} \tag{9}$$

and the global force is obtained as

$$F = \sum_{i=1}^{NS} F_i^z\Big|_n \tag{10}$$

where $F_i^z\Big|_n$ is the component of the local ice force in the direction of the ice motion.

4.2 Loading phases

An implicit time integration technique is used to evaluate the ice force as a function of the relative displacement. Corresponding to the experimental findings, the interaction process is simulated considering four different loading phases termed as *loading*, *unloading*, *pure crushing with extrusion* and the *hysteretic unloading*.

Loading

A loading phase is defined as a period of ice-structure interaction where the ice force increases. This condition prevails if the structure and the ice edge are in contact with each other and are moving in opposite directions. In the previous versions of the PSSII programme [15, 18] the ice force was assumed to increase as a linear function of the compressive strain at the near field element. A nonlinear relationship derived by Kärnä and Sippola [19] is used here with some modifications [20].

Unloading

The ordinary loading phase in a layer is followed by an unloading phase when a major or secondary flaking occurs at the ice edge. Laboratory data shows that the drop from a peak

load to the next minimum occurs within a time T_i^u that is typically 10% to 20% of the preceding loading time. The first version of the PSSII model [16] assumed that after an ice failure the ice force decreases as a linear function of time. A more accurate physical model would take account of the extrusion of the crushed ice [18]. Considering the difficulties posed by the edge geometry, a simplified version of the extrusion process is adopted here.

Referring to basic characteristics of the unloading process, Kärnä et al. [20] showed that the unloading after a major flaking event can be simulated by the incremental force function

$$\Delta F_i^z(t) = 6\left(F_i^{cr} - F_i^{ex}\right)\left\{\left(\frac{t}{T_i^u}\right)^2 - \left(\frac{t}{T_i^u}\right)\right\}\frac{\Delta t}{T_i^u} \tag{11}$$

where F_i^{cr} is the peak load at the preceding event of ice failure and F_i^{ex} is the minimum force level at the end of the unloading. A similar force vs. time function is used to describe the unloading in the case of secondary flaking [20].

Pure crushing with extrusion

At the event of major flaking the near field element E_i looses ice material from the whole contact area [16]. Therefore, all the relative displacements $w_{ik}^{dz}, k = 1,\ldots,NDZ$ become negative. This condition prevails during the unloading phase. Our simplified equation (11) for the unloading phase is a function of time and not of the relative displacement. Accordingly, the relative displacements w_{ik}^{dz} may remain negative after the minimum level F_i^{ex} has been reached. Therefore, we define an intermittent phase *of pure crushing with extrusion*, where the ice force remains at a constant level until the next loading phase.

To give a physical justification for this approach we recognize that crushed ice may exists at the ice edge after the unloading phase. The crushed ice is capable of transmitting forces between the structure and the ice edge and it is extruded from the contact area if the structure moves against the ice edge. Furthermore, pure crushing at a low force level is likely to occur at the uneven ice edge before a new loading phase begins.

Hysteretic unloading

A new loading phase at a layer begins when w_{ik}^{dz} becomes positive. The loading phase will continue as long as the incremental relative displacement Δw_{ik}^{dz} remains positive but can be interrupted before an ice failure. This happens if the relative displacement increment Δw_{ik}^{dz} becomes negative. An intermittent drop in the contact force will occur in this situation. This condition is related to the hysteretic phenomenon that was discussed in Sect. 2.2.

A plausible physical explanations for this hysteretic damping effect can be given by considering the frictional forces associated with the layered near field model. Fig. 7 depicts a condition where the central layer is loaded. Due to the compressive force the central layer expands in vertical direction causing compressive vertical stresses on the adjacent layers. At a load reversal the structure departures from the ice edge. Therefore, the compressive force at the central layer is released. The friction acting on the layer surfaces may prevent the sliding at the surfaces for a while. When the force is sufficiently low, sliding occurs at the boundaries and the central layer rebounds towards the structure. After a while the structure moves again against the ice edge due to its transient vibration. The frictional forces on the boundaries of the central layer are also now opposing the sliding at the horizontal boundaries of the layer.

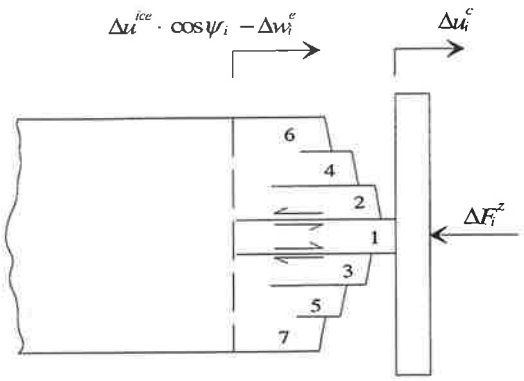


FIGURE 7: Damping mechanism at the near field.

Accordingly, we propose that the ice edge provides Coulomb damping which is sufficient to cause a rapid stop of the transient vibration of the structure. Fig. 7 illustrated this damping mechanisms referring to the central layer of the near field. The other layers are likely to provide damping in a similar way. In the numerical model we simulate the Coulomb damping by a viscose model [20].

5. EXAMPLES

5.1 Quasi-static response with transient vibration

To verify the present formulation, some of the test conditions discussed in Sect. 2 were simulated. The results of the computation were compared with measured response of test structures. The test condition shown in Fig. 2 is considered first. Detailed information on the test is given in the report [21].

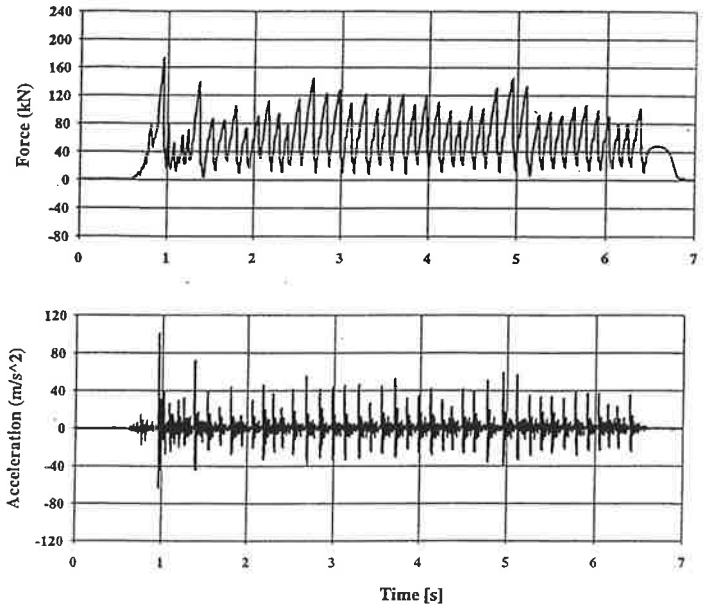


FIGURE 8. Measured ice force and structural response in an indentation test No 17 [21].

Structural mass	$M = 2\,000$ kg	Ice thickness	$h = 115$ mm
Spring stiffness	$K^s = 65$ kN/mm	Rate of indentation	$v = 45$ mm/s
Natural frequency	$f = 27$ Hz	Indentor width	$D = 300$ mm

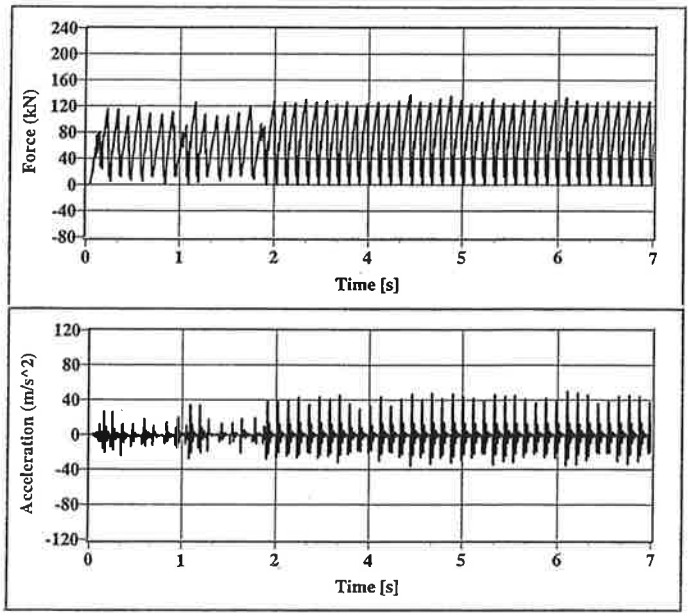


FIGURE 9: Simulated ice force and structural response in Test No 17 [Compare Fig. 9].

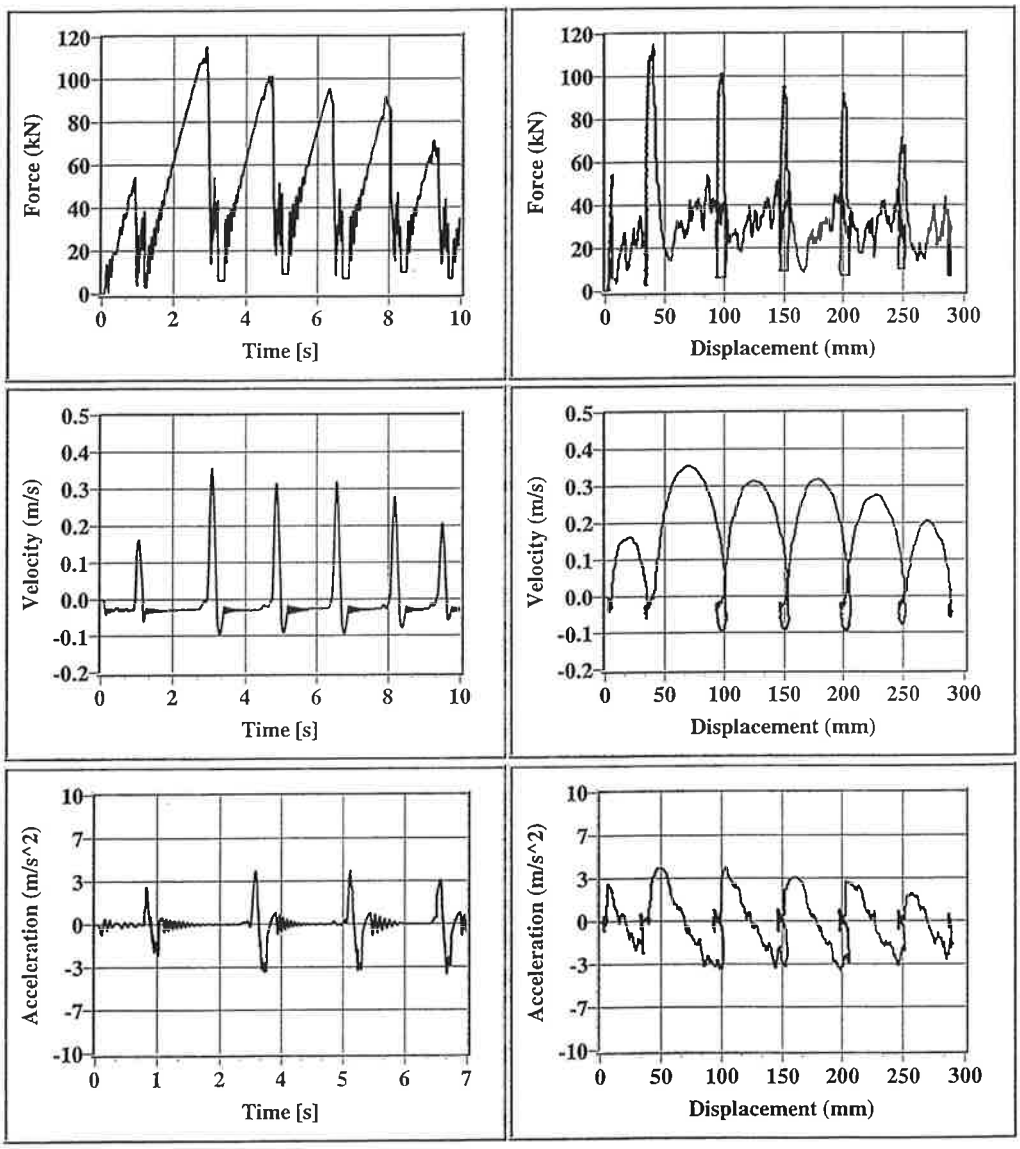


FIGURE 10: Simulated ice force and structural response in Test No 69 [Compare Fig. 2].

Figs. 8 and 9 show the measured and simulated ice force and response records for a test where the model structure was compliant but very stiff. A comparison shows that the simulated force and response signals are in good agreement with the measurements. Correspondingly, Fig. 10 shows the simulated response of Test No 69 where a very compliant structure was pushed against the ice sheet. The corresponding test results were shown earlier in Fig. 2. Also here the simulation gives good results.

It should be noticed that the earlier numerical models [2, 12,14, 18] did not predict two interesting details that have been seen in many tests [5, 10, 11, 14, 21, 26]. First, the transient vibration after a major peak force decays to a low level after one cycle. The hysteretic unloading phase of the present model accounts for this effect. The second interesting feature of this simulation is that the force signal $F(u)$ shows a growing trend during the transient spring-back phase (Figs. 2, 3 and 10). The secondary flaking process on the wedge-shaped ice edge accounts for this effect.

Figs. 2, 3 and 10 show that the global ice force remains at a relatively low level during the "spring-back" periods where the relative velocity is high. Nonsimultaneus secondary flaking failure occurs at the near field elements during these periods. The events of large peak forces occur in quasi-static conditions as simultaneous major flaking.

5.2 Steady state vibration

The channel marker described by Kärnä and Turunen [12] was used to simulate the ice-induced vibration. An ice field with a thickness of 200 mm was assumed to drift at the velocity of 100 mm/s against the structure. Fig. 11 shows the results of this simulation.

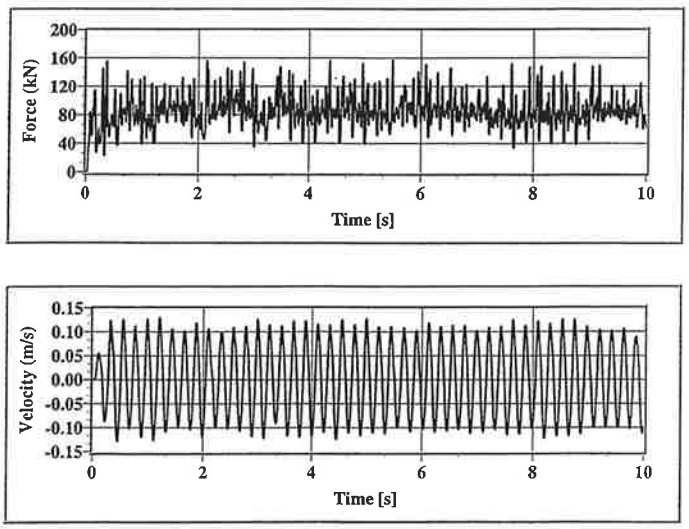


FIGURE 11: Simulated ice force and structural response on a channel marker.
v = 100 mm/s; h = 200 mm.

The simulated ice force has a similarity with the measured ice force in a tests condition where ice-induced vibration occurred in laboratory [14: Test 48]. The structural response at the water level is shown in Fig. 11. The velocity amplitude at the water level is approximately the same as the ice velocity, as predicted in [13].

6. CONCLUSIONS

A new model was developed for the near field area of an ice sheet that experiences dynamic interaction with an offshore structure. In this model, the ice volume close to the structure is divided first into a set of near field elements. These elements are then divided further into layers that are bounded by the horizontal cracks emanating form the ice edge during the ice-structure interaction. Depending on the rate of interaction, the ice failure at the near field elements occurs either by major flaking by secondary flaking. A hysteretic damping phenomenon that has been seen in experiments is included in the model.

The new model is incorporated in an existing computer program for wide and compliant structures. Simulation of a few test cases shows that the new model can predict the observed changes from a nonsimultaneous ice failure into simultaneous failure.

ACKNOWLEDGMENTS

This work was partially funded by the Technology Development Centre of Finland (TEKES) and by the European Union S&T Grant Programme in Japan.

REFERENCES

1. Daley C., Ice edge contact. A brittle failure process model. *Acta Polytechnica Scandinavica. Mechanical Engineering Series* No. 100. Helsinki University of Technology. Helsinki 1991. 92 p.
2. Eranti, E., Dynamic ice structure interaction - Theory and applications. *VTT Publications No. 90.* 81 p.
3. Fransson, L., Olofsson, T. & Sandkvist, J., Observations of the failure process in ice blocks crushed by a flat indentor. *Proc. 11th Int. Conf. Port & Ocean Eng. under Arctic Cond.* (POAC-91). St. John´s, Canada, Vol. 1, pp. 501-514.
4. Fransson, L., Ice load on Structures. Ice testing in Luleå Harbour 1993. Cold Tech Report 1995:06, 21 p + Apps.

5. Jefferies, M.G. & Wright, W.H., Dynamic response of "Molikpaq" to ice-structure interaction. *Proc. 7th Int. Conf. Offshore Mech. and Arctic Eng.,* (OMAE-88). Houston. Vol. IV, pp. 201-220.

6. Joensuu, A. & Riska, K., Contact between ice and structures. *Helsinki Univ. Techn., Faculty Mech. Eng., Ship Lab., Report M-88* (in Finnish).

7. Jones, D.E., Kennedy, F.E. & Schulson, E.M., The kinetic friction of saline ice against itself at low sliding velocities. *Annals of glaciology* Vol 15, p 242-246.

8. Kajaste-Rudnitski, J., On dynamics of ice-structure interaction. *VTT Publications* No. 257. 104 p. + Apps.

9. Kamesaki, K., Yamauchi, Y., Horikawa, T, Kawasaki, T., Ishikawa, S. & Tozawa, S., Experimental Study on independent failure zone of ice crushing. *Proc. 10th Int. Symp. Okhotsk Sea Ice & Peoples,* pp. 146-151.

10. Kamesaki, K., Tsukuda, H. & Yamauchi, Y., Indentation tests with vertically placed ice sheet. *Proc. 13th Int. Conf. Port & Ocean Eng. under Arctic Cond.* (POAC-97). Yokohama, Japan.

11. Kawamura, M., Takeuchi T., Akagawa, S., Terashima T., Nakazawa, N., Matsushita, H., Aoshima, M, Katsui, H. & Sakai, M., Ice-structure Interacions in medium scale field indentation tests. *Proc. 13th IAHR Ice Symp.* (IAHR-96). Beijing. Vol. 1, pp. 228-236.

12. Kärnä, T. & Turunen, R., Dynamic response of narrow structues to ice crushing. *Cold Regions Science and Technology.* 17(1989) 173-187.

13. Kärnä, T. & Turunen, R., A straightforward technique for analysing structural response to dynamic ice action. *Proc. 9th Int. Conf. Offshore Mech. and Arctic Eng.,* (OMAE-90). Houston. Vol. IV, pp. 135-142.

14. Kärnä, T. & Muhonen, A., Preliminary results from ice indentation using flexible and rigid indentors. *Proc. 10th IAHR Ice Symp.* (IAHR-90). Espoo, Vol. 3, pp. 261-275.

15. Kärnä, T., A procedure for dynamic soil-structure-ice interaction. *Proc. 2nd Int. Offshore and Polar Eng. Conf.* (ISOPE-92). San Francisco, Vol II, pp. 764 - 770

16. Kärnä, T., Finite ice failure depth in penetration of a vertical indentor into an ice edge. *Annals of Glaciology,* Vol 19. pp. 114-120.

17. Kärnä, T., Nyman, T., Vuorio, J. & Järvinen, E., Results from indentation tests in sea ice. *Proc. 12th Int. Conf. Offshore Mech. and Arctic Eng.,* (OMAE-93). Glasgow. Vol. IV, pp. 177-185.

18. Kärnä, T. & Järvinen, E., Dynamic unloading accross the face of a wide structure. *Proc. 12th IAHR Ice Symp.* (IAHR-94). Trondheim, Vol. 3, pp. 1018-1039.

19. Kärnä, T. & Sippola, M., Nonlinear loading phase in ice indentation. *Proc. 6th Int. Offshore and Polar Eng. Conf.* (ISOPE-96). Los Angeles. Vol. II, pp. 349-353..

20. Kärnä, T., Kamesaki, K. & Tsukuda H., A flaking model of dynamic ice-structure interaction. VTT Building Technology. *Internal Report* RTE38-IR-7/1997, 51 p.

21. Muhonen, A., Kärnä, T., Eranti, E., Riska, K., Järvinen, E & Lehmus E. Laboratory indentation tests with thick freshwater ice. Vol I. *VTT Research Notes 1370.* Espoo, 92 p.

22. Määttänen, M., On conditions for the rise of self-excited ice-induced autonomous oscillations in slender marine pile structures. *Winter Nav. Board, Finland, Res. Rep., 25.*

23 Määttänen, M., Laboratory tests for dynamic ice-structure interaction. *Eng. Struct., 3(1982),* pp. 111 - 116.

24. Nordlund, O.P., Kärnä. T. & Järvinen E., Measurements of ice-induced vibrations of channel markers. *Proc. 9th IAHR Ice Sym.* (IAHR-88). Sapporo. Vol. 2, pp. 537 - 548.

25. Saeki, H., Ono, T., Nakazawa, N., Sakai, M. & Tanaka, S., The coefficient of friction between sea ice and various materials used in offshore structures. *J. Energy Resources Technology.* Vol 108, March 1986. pp. 65-71.

26. Sodhi, D., Ice-structure interaction with segmented indentors. *Proc. 11th IAHR Ice Symp.* (IAHR-92). Banff, Alberta. Vol 2, pp. 909 - 929.

27. Taylor, T., An experimental investigation of the crushing strength of ice. *Proc. 6th Int. Conf. Port and Ocean Eng. under Arctic Cond.* (POAC-81). Québeck, Canada. Vol. 1, pp. 332-344.

28. Tuhkuri, J., Computational fracture mechanics analysis of flake formation in brittle ice. *Proc. 13th IAHR Ice Symp.* (IAHR-96). Beijing, China. Vol. 1, pp. 54-61.

DETERMINATION OF THERMAL PROPERTIES USING REGULARISED OUTPUT LEAST-SQUARES METHOD

Jukka Myllymäki & Djebar Baroudi
VTT BUILDING TECHNOLOGY/Fire Technology
P.O. Box 1803, FIN-02044 VTT

ABSTRACT

The process of analysis in structural fire design comprises three main components; determination of the fire exposure, the thermal analysis and the structural analysis. The thermal analysis requires well-defined input information on thermal material properties for determining the transient temperature state of the fire-exposed structure.

Some applications of a systematic methodology to treat identification of temperature dependent thermal properties and of other relevent quatities from tests are presented. This method is known as the Regularized Output Least Squares Method (ROLS). Applications of the method to identification of thermal properties in different cases are presented. For each case, the Direct problem consists of a set of non-linear partial differential equations which are semi-discretized via the variational form of the heat conduction problem. The solution of the Direct problem is obtained by time-integrating the semi-discrete equations by mean of numerical quadrature. The problem of identification of the parameters appearing in the formulation of the Direct problem is know as an inverse problem.

INTRODUCTION

A common feature of inverse problems is the *instability,* that is, small changes in the data may give rise to large changes in the solution. Small finite dimensional problems are typically stable, however, as the discretization is refined, the number of variables increases and the instability of the original problem increases. Therefore regularization is needed. Both mesh coarsing and Tikhonov-regularization have been adopted in order to get a stabilised solution. The available *a priori* known physical constraints on the parameters are taken into account in the minimisation.

The distributed parameters are discretized, usually, the thermal properties are approximated as piece-wise linear functions of temperature. The unknowns are found by minimising a constrained and regularised functional which is the sum of the squares of residual norm of the errors (data - model) plus the square of the norm of the second derivatives of the properties with respect to the temperature. An appropriate balance between the need to describe the measurements well and the need to achieve a stable solution is reached by finding an optimal regularization parameter. Both Newton and conjugate gradient methods have been used in the minimisation. The Morozov discrepancy principle is used to find a reasonable value for the regularization parameter.

FORMULATION OF THE DIRECT PROBLEM (HEAT CONDUCTION PROBLEM)

The basic idea is to solve the temperature field $T_i(x,t)$ in a given material region. The field equation

$$\rho c\ \dot{T}(x,t) = \vec{\nabla}\cdot\left(\lambda\vec{\nabla}T(\vec{x},t)\right) + r(\vec{x},T) \tag{1}$$

is the diffusion equation with *r(x,T)* as an arbitrary source term. The Fourier heat conduction constitutive relation is assumed. This equations is complemented with the appropriate initial-boundary conditions to get a well-posed problem. The boundary conditions may be, for example, a Dirichlet type or Neumann type as for instance normal heat flux $q_n = h(T)(T - T_\infty) + \sigma\ \varepsilon\ (T^4 - T_\infty^4)$ with convection and radiation parts. The boundary terms as also the source terms if present will be included into the force vector of the discritized heat conduction equation. This will be a clear and systematic way to treat boundary and source terms.

SEMI-DISCRETIZATION OF THE FIELD EQUATIONS

Using the standard FE-approach ,[1], one obtain the variational form of the problem (1) as

$$\int_\Omega \rho c\dot{T}\, v\ d\Omega + \int_\Omega \lambda\ \vec{\nabla}T\cdot\vec{\nabla}v\ d\Omega = \int_\Omega r\ v\ d\Omega - \int_{\partial\Omega_q} \vec{q}\cdot\vec{n}\ v\ d\Gamma \tag{2}$$

with the temperature field approximated by $T^e(x,t) = \mathbf{N}^e(x)\mathbf{T}^e(t)$, where the test and the basis functions **N(x)** (Galerkine) linear. For instance in the case of 1-D for a linear element we have $N_1(\xi) = (1-\xi)/2$ and $N_2(\xi) = (1+\xi)/2$.

The semi-descretization of the heat conduction equation (2) produces the non-linear initial value problem

$$\mathbf{C}(t,\mathbf{T})\dot{\mathbf{T}}(t) = \mathbf{f}(t,\mathbf{T}) - \mathbf{K}(t,\mathbf{T})\mathbf{T}(t)\,,\ t > 0 \tag{3}$$

$$\mathbf{T}(0) = \overline{\mathbf{T}}_0\,,\quad t = 0\,,$$

where **T**(t) is the global vector of the unknown temperatures.

Equation (3) is a set of *n x 1*- non-linear ordinary differential equation. Notice that the right hand in the equation (2) corresponds to the force vector $\mathbf{f}(t,\mathbf{T})$, which contains the boundary terms as also all possible source terms. Equation (3) is be complemented with appropriate initial conditions. Natural boundary conditions are already included in the variational form (2). The essential boundary conditions will be taken into account during

the solution process of the initial value problem. The global matrices and vectors are assembled using standard FE-assembling techniques. The elementary conductivity matrix

$$K_{ij}^e = \int_{\Omega^e} \lambda(T(\mathbf{x}))\ \vec{\nabla}\mathrm{N_i}(\mathbf{x}) \cdot \vec{\nabla}\mathrm{N_j}(\mathbf{x})\ \ d\Omega\ , \tag{4}$$

the elementary capacity matrix

$$C_{ij}^e = \int_{\Omega^e} \rho(T(\mathbf{x}))\ c(T(\mathbf{x}))\ \mathrm{N_i}(\mathbf{x})\ \mathrm{N_j}(\mathbf{x})\ \ d\Omega \tag{5}$$

and the elementary force vector

$$f_i^e = \int_{\Omega^e} r(T(\mathbf{x}))\ \mathrm{N_i}(\mathbf{x})\ d\Omega - \int_{\partial\,\Omega_q} \vec{q}\cdot\vec{n}\ \mathrm{N_i}(\mathbf{x})\ d\Gamma \tag{6}$$

are obtained. For instance, in 1-D cases these matrices look like:

$K_{ij}^e = 2/l^e \int_{-1}^{1} \lambda(T(\xi))\ N_{i,\xi}(\xi) N_{j,\xi}(\xi)\ d\xi$, $C_{ij}^e = \frac{l^e}{2}\int_{-1}^{1} \rho(T(\xi))c(T(\xi))N_i(\xi)N_j(\xi)\ d\xi$ and

$f_i^e = \frac{l^e}{2}\int_{-1}^{+1} r(T(\xi))N_i(\xi)\ d\xi - \left[q_n N_i(\xi)\right]_{-1}^{+1}$, respectively.

The elementary matrices and vectors are integrated numerically using Gauss-Legend integration scheme. The elementary matrices and vectors may depend on the unknown temperature.

The above mentioned integration scheme leads to consistent capacity matrix, where the non-diagonal terms $C_{i,j}^e\ (i \neq j)$ are non-zero. In some cases it is practical to use diagonal capacity matrix ($C_{i,j}^e\ (i \neq j) = 0$) especially when we use Dirichlet type of boundary condition. Using Newton-Cote integration scheme where the nodal points are used as integration points and the weights of the integration are calculated as $w_i = \int_0^l N_i N_i dx$, we always get a diagonal capacity matrix.

TIME-INTEGRATION OF THE ODE-SYSTEM

Depending on the integration scheme used in equation (3) we get explicit or implicit methods. In the case of a heat conduction problem, the time-integration gives us the non-linear system of equations

$$\mathbf{A}(t_{n-1}, \mathbf{T}_{n-1})\mathbf{T}(t_n) - \mathbf{g}(t_{n-1}, \mathbf{T}_{n-1}) = \mathbf{0} \tag{7}$$

for which the solution **T**(t) is solved from equation (8) at each time step using the fixed point iteration procedure. The matrix in equation (7) is calculated as

$$\mathbf{A}(t_{n-1}, \mathbf{T}_{n-1}) = \mathbf{C}(t_{n-1}, \mathbf{T}_{n-1}) + \Delta t_n \mathbf{K}(t_{n-1}, \mathbf{T}_{n-1}) \tag{8}$$

and the vector

$$\mathbf{g}(t_{n-1}, \mathbf{T}_{n-1}) = \mathbf{C}(t_{n-1}, \mathbf{T}_{n-1})\mathbf{T}(t_{n-1}) + \Delta t_n \mathbf{f}(t_{n-1}, \mathbf{T}_{n-1}) \tag{9}$$

GENERAL FORMULATION OF THE INVERSE PROBLEM

Consider a coefficient determination problem, i.e. the problem of determining a non-constant coefficient $a(y)$ in an initial value problem (3) on the base of the existing data about the solution y (y is the temperature). In (3) the unknown parameters entering the capacity matrix **C(T)**, the heat conductivity matrix **K(T)** and the force vector **f(T)** are gathered into the vector **a**(**y**), where $\mathbf{y} \equiv \mathbf{T}$. The non-linear inverse problem (10) is solved using the *regularised output least squares method* (**ROLS**). We have to discretize the distributed unknown parameter $a(y)$ into a certain number of sub-intervals $[y_i, y_{i+1}]$ of arbitrary length $y_{i+1} - y_i$ using linear basis functions. The goal is to find out the regularised least square solution for the vector of the nodal values $\vec{a}_i^*$, [2]. Therefore one seeks unknowns $\vec{a}_i^*$ such that

$$\min_{\vec{a}_i^* \in D} \left\| \widetilde{\mathbf{F}}(\vec{a}_i) - \vec{y} \right\|^2 + \alpha \left\| \mathbf{L}\vec{a} \right\|^2 \tag{10}$$

where the constraints set D is the set of physically admissible parameters and the notation $\left\| \widetilde{\mathbf{F}}(\vec{a}_i) - \vec{y} \right\| \equiv \left\| \mathbf{T}_{FE}(\mathbf{a}) - \mathbf{T}_{data} \right\|$ is used. Unfortunately the data vector $\vec{y}$ is known (or measured) only within a certain tolerance δ. Only an approximation $\vec{y}^{\delta}$ satisfying the condition $\left\| \vec{y} - \vec{y}^{\delta} \right\| \le \delta$ is known (for example due to the scatter/data errors in the experimental measurements) and one therefore seeks an $\vec{a}^*$ minimising (10) using data infected with noise. Here $\vec{y}_k^{\delta}$ is the vector of measured data. The minimisation problem is non-linear. Here either Newton or Conjugate Gradient methods are used.

The overall procedure of determination of the thermal properties (and other relevant parameters) may be condensed schematically as follow:

Discretize the unknown vector **a** with respect to the temperature (if necessary) using piece-wise linear basis functions. Guess a 'realistic' initial value for **a** and choose a value for α. Then ***Solve*** **a** from the minimisation problem

- $\min_{\vec{a} \in D} \left\| T_{FE}(\vec{a}; \vec{x}; t) - T_{data}(\vec{x}, t) \right\|^2 + \alpha \left\| \mathbf{L}\, \vec{a} \right\|^2$ with respect to **a**,

where the model is $T_{FE}(\vec{a};\vec{x};t) = \mathbf{N}(\vec{x})\mathbf{T}(t)$

and the data is $T_{data}(\vec{x};t)$ with **T**(t) the solution of the initial value problem:

$\mathbf{C}(\vec{a}(\mathbf{T}))\ \dot{\mathbf{T}}(t) = \mathbf{f}(\vec{a}(\mathbf{T})) - \mathbf{K}(\vec{a}(\mathbf{T}))\ \mathbf{T}(t)$ & BC and IC

- The unknown parameter vector is: $\vec{a}(\mathbf{T}) = \left(\lambda(T) \quad c_p(T) \quad \vec{b}\right)$ with the vector

$\vec{b} = (\varepsilon \quad h(T) \quad Q_{loss} \quad etc...)$ containing the remaining relevant additional parameters we want to estimate. The norms (Euclidean) are taken with respect to the collocation points **x** at collocation time *t* as $\|f(\vec{x};t)\|^2 = \sum_i \sum_j \left|f(\vec{x}_i;t_j)\right|^2 \quad (i,j) \in \{\ collocation\ index\ set\}$. **L** is a regularization operator (depending on the degree of regularization we want), usually the identity **I** matrix or a discrete version of the Laplacian with respect to the temperature.

Since the *inverse problem is ill-posed* it has to be *regularised.* Here, in the regularised output least squares method (ROLS) the regularization of the problem is achieved by the use of the *penalised least squares* method.. The last one can be regarded as *Tikhonov regularization* of non-linear problems, [2], [3] and [4]. Mesh coarsing and the use of the available *a priori* known physical constraints on the parameters is also used as regularization. In penalised least squares method one seeks a minimum for the functional

$$\left\|\vec{F}(\vec{a}^*) - \vec{y}^\delta\right\|^2 + \alpha\left\|L\vec{a}^*\right\|^2 \tag{11}$$

where $\alpha(>0)$ is a small regularization parameter depending on the noise level of the data and $L = I$ or some other suitable differential operator (D^1 or D = Laplacian) depending on the needed *regularity* of the solution. The first term in Eqs. (10-11) enforces the consistency of the solution when the second term enforces its stability. An appropriate balance between the need to describe the measurements well and the need to achieve a stable solution is reached by finding an optimal regularization parameter.

The use of the Morozov´s discrepancy principle and the L-curve.

In the equations (10-11) parameter α controls how much weight is given to minimisation of $\left\|L\vec{a}^*\right\|^2$ relative to minimisation of the square of the residual norm $\left\|\vec{F}(\vec{a}^*) - \vec{y}^\delta\right\|^2$. The problem considered here is the appropriate choose of the parameter α so that we can distinguish the real signal from the measurement errors, the noise.

Perhaps the simplest and clearest rule to choose the regularization parameter is to set the residual norm equal to some upper bound for the norm $\left\|\vec{F}(\vec{a}^*) - \vec{y}^\delta\right\|$ of the errors, i.e. find α such that, [2]

$$\left\| \tilde{F}(\vec{a}^*) - \vec{y}^{\delta} \right\| \leq R\,\delta \qquad (12)$$

where δ is the measure of the error during the time considered $\delta^2 = \int_0^t |\,\partial\,|^2 dt$ and ∂ is the error at certain time. An appropriate value of coefficient is $R \approx 1{,}6 - 1{,}7$. In connection with discrete ill-posed problems this is called the *Morozov's discrepancy principle.*

Another, more recent alternative is to base the regularization parameter on so-called *L-curve* [3] and [5]. The *L*-curve is a parametric plot of the measure of the size of the regularised solution and the corresponding residual. The underlying idea is that a good method for choosing the regularization parameter for discrete ill-posed problems must incorporate information about the solution size in addition to using information about the residual size. Usually the *L*-curve has a distinct L-shaped corner located where the solution changes in nature from being dominated by regularization errors to being dominated by the errors in the right side. Corner of the *L*-curve corresponds to a good balance between minimisation of the sizes, and the corresponding regularization parameter α is a good one, [4] and [5]. In the calculation examples presented here, this corner is seldom seen, therefore the Morozov discrepancy principle was preferred.

APPLICATION TO HEAT CONDUCTION PROBLEMS OF FIRE ENGINEERING

Thermal diffusivity of a nickel wire

The thermal diffusivity κ (= $\lambda/\rho c$) of a Nickel alloy (95 % Ni) was estimated using temperature measurements. The problem is formulated by the *1-D* heat conduction equation $\rho c\dot{T} - \lambda\Delta T = \rho r \equiv 2Q_{loss} / L_{12}$, where L_{12} is the total length of elements 1 and 2. In this example the FE-direct formulation was constructed using two linear elements. Diagonal capacity matrix was used and this leads to the initial value problem $\dot{T}_2 = -2\beta_2\kappa T_2 + 2\kappa / L_{12}\left(T_1 / L_1 + T_3 / L_2\right) + 2Q_{loss} / \rho c L_{12}$, where $\beta = (1/L_1 + 1/L_2)/L_{12}$. The index k in T_k represents the number of the thermocouple, Fig. 1b). T_1 and T_3 are the measured temperatures (Direchlet BC).The energy conservation equation includes an unknown loss term Q_{loss}, which is part of the parameters to be estimated. Q_{loss}, takes into account heat loses from the wire into the insulation. Thus the unknowns are $\mathbf{a} = (Q_{loss}, \kappa)$. The cylindrical wire is of diameter 0.5 mm and heated at one free end (TC1), Fig. 1b). The wire was thermally insulated by a thick layer of mineral wool in order to get a 1-D problem. The wire was initially at ambient temperature. The temperatures were recorded using thermocouples at specified points, Fig. 1b). The whole temperature history was used in the inverse problem.

The parameters were estimated as $\mathbf{a} = (Q_{loss} = V\rho\, r, \kappa) = (-1.08\ \text{W},\ 2.133x10^{-5}\ \text{m}^2/\text{s})$. The heat loses are over a length of the wire L_{12} = 4.7 cm. The thermal conductivity was then

deduced as λ = 80 W/m K. The density of the Ni-alloy was 8447 kg/m^3 and its thermal capacity c_p was fixed to 444 J/kg K. The values presented in literature for the next Ni-alloy (Ni 80 % and Cr 20%) are: λ = 15 - 104 W/m K and c_p = 460 - 500 J/kg K. These values agree well with the value λ = 80 W/m K calculated although the Ni-alloy we used in the experiment was 95 % Ni 3 % Al and 2 % Mn.

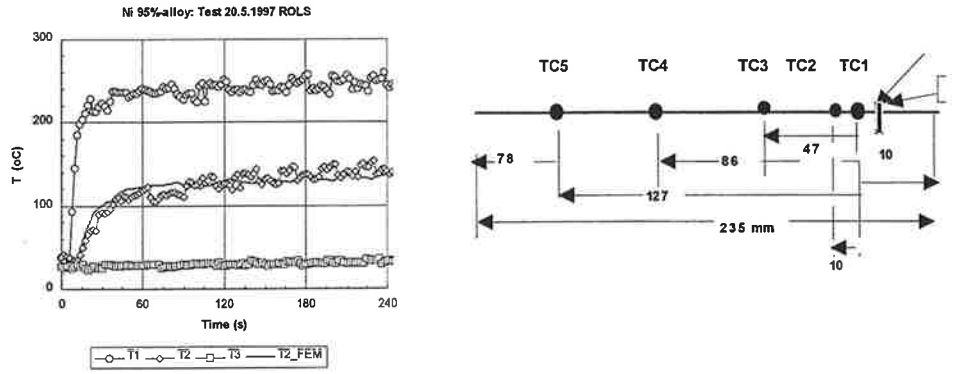


Figure 1. a) Measured (markers) and calculated (continuous line) temperatures at locations TC1, TC2 and TC3. b) A schematic view the Ni-wire and configuration of the thermocouples TC1,TC2...TC5.

Heat capacity and thermal conductivity of gypsum board.

Cone calorimeter tests in horizontal configuration at a heat flux level of q_{cone} = 25 kW/m^2 were performed. The test specimen consisted of a 13 mm thick gypsum board (density 721 kg/m^3) laying on a 30 mm thick layer of mineral wool (Fig. 2). The surface area exposed to the heat flux was A_1=100 mm x 100 mm. There was a 10 mm thick aluminium plate under the gypsum board in one test. In an other test there was no aluminium plate present. The first gypsum example presented uses data from the test without the aluminium plate. The temperature of the upper surface of the gypsum board was measured using an infra-red temperature measuring device The temperature profile inside the specimen as a function of time was measured using thermocouples.

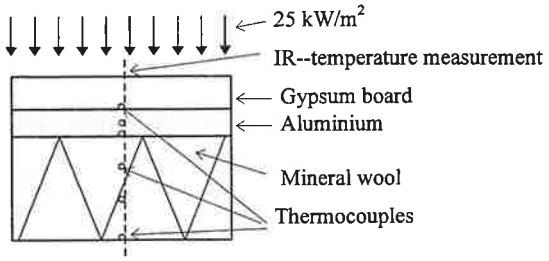


Figure 2. A schematic representation of test arrangement in cone calorimeter test for gypsum board.

The problem is to estimate the heat capacity $c_p(T)$ and the thermal conductivity $\lambda(T)$ of a given gypsum board from temperature measurement tests. Two tests was performed. In this first example the specific heat of the gypsum board was calculated. The specific heat is discretized using piece-wise linear basis function with respect to the temperature $c(T) = \sum N_j a_j(T)$. The conservation of energy in the gypsum board can be written as

$$\int_V \rho c \dot{T} dV = -\int_{\partial V} \vec{q} \cdot \vec{n} d\Gamma + \int_V \rho r dV \tag{13}$$

In the present example the source term in the equation is incorporated into the effective specific heat. The equation (13) after semi-discretization reads as

$$\dot{\tilde{T}}(t) = +q_{cone} \frac{A_1}{\rho c(\tilde{T}) V} \equiv f(T(x;t),t), \tag{14}$$

where $\tilde{T}(t) = 1/d \int_d T(x;t)\, dx$. The ODE (14) is integrated numerically using the explicit Euler scheme

$$\tilde{T}(t_{k+1}) = \tilde{T}(t_k) + \int_{t_k}^{t_{k+1}} f(\tilde{T}(\tau),\tau) d\tau \approx \tilde{T}(t_k) + f(\tilde{T}(t_k),t_k)\Delta t_k \tag{15}$$

The specific heat $c_p(T)$ was the regularised solution of the constrained minimisation problem

$$\min\left(\left\| \tilde{T}_{test}(t) - \tilde{T}_{calc.}(\vec{a};t) \right\|^2 + \alpha \left\| \mathbf{L}\vec{a} \right\|^2 \right), \quad with\ a_j \in D(\vec{a}) \tag{16}$$

The solution of (16) is found from the domain of physically admissible functions which takes into account the possible range of the unknown parameters a_j The equation (16) is non-linear for which a solution is found using Newton method. A reasonable degree of regularization (the value of the regularization parameter α) is found using the Morozov's discrepancy principle $\left\| \tilde{T}^{\delta}(t) - \tilde{T}_{calc.}(\vec{a}^{\delta}_{\alpha(n)};t) \right\| \approx R\delta$ (R = 1.6 and α = 0.00001). The accuracy of the temperature measurements in these tests was estimated to be $\left\| \tilde{T}^{\delta}(t) - \tilde{T}(t) \right\| \le \delta \approx \int_{t_\infty} 2^o C\, dt \approx 98^o Cs$.

The results of the calculations are shown on Fig. 3. The relative amount of humidity (mass of water / total mass of gypsum) in the gypsum board was calculated from the peak of the calculated specific heat was found to be equal to 21 %. Experiments show when drying the gypsum boards that the water contents was about 18 % in the temperature range < 200 °C. This is with agreement with the results of the calculation.

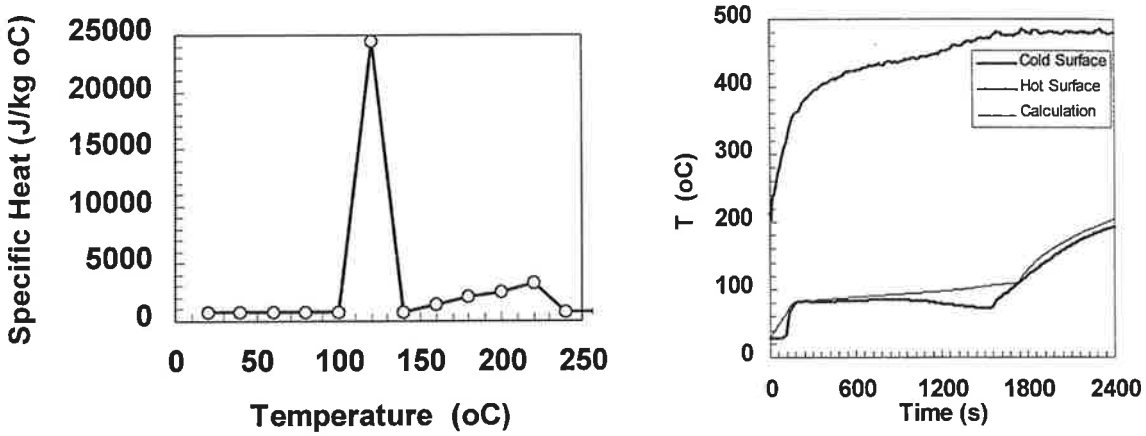


Figure 3. a) calculated specific heat of the gypsum board from cone calorimeter experiment At 25 kW/m^2. b) Measured surface temperature, bold lines, the calculated in thin line.

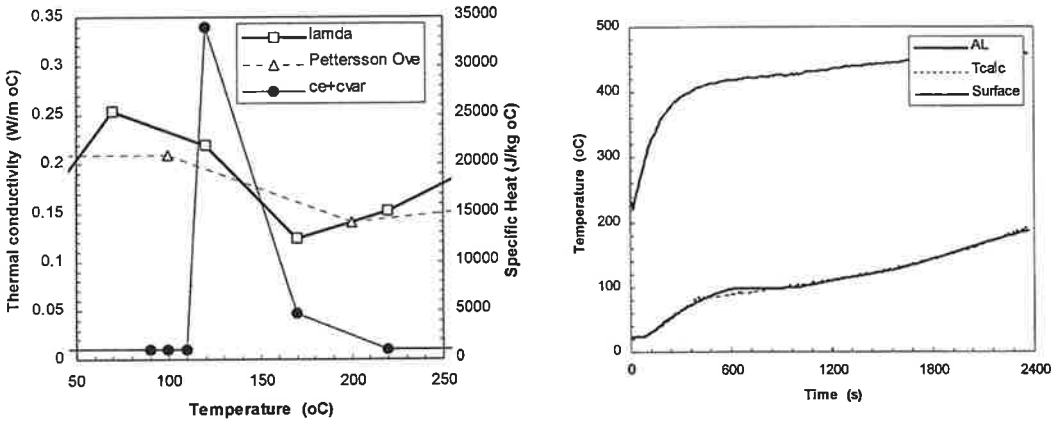


Figure 4. a) Calculated specific heat and thermal conductivity of the gypsum board from cone calorimeter experiment at 25 kW/m^2. Test data for the thermal conductivity after Pettersson [5] *is shown for comparison. b) Measured surface temperatures and the temperature of the aluminium plate, bold lines, the calculated in dotted thin line.*

The second test the was similar to the first test except for the aluminium plate. Directly under the gypsum board an aluminium 10 mm thick aluminium plate was present. The temperature of the test specimen was recorded as in the first example. The temperature distribution inside the gypsum board was approximated linearly and the aluminium temperature as constant in the space dimension. Here both the specific heat $c_p(T)$ and the thermal conductivity $\lambda(T)$ were the unknowns $\mathbf{a} = (.c_p(T)\ ,\ \lambda(T))$. They were both discritized using piece-wise linear basis functions in respect to the temperature. Figure 4 shows the results. Also the thermal conductivity of gypsum after Pettersson [5] is shown for comparison.

Heat capacity of an aluminium bar

Let us consider a case were a uninsulated aluminium specimen, a bar, (2700 kg/m^3) is placed in an oven in order to be tested at high temperatures. The specimen is surrounded by the oven, but the ends of the specimen are clamped into steel rods so part of the heat flow escapes through the ends of the specimen. The heat conduction problem is dealt as one dimensional where it has been assumed that heat loss $Q_{loss} = q_n A_{\Gamma_0} / c_{Al}\rho_{Al} A_{Al}\, d_{Al}$ is constant.

The inverse problem is - given the measured temperatures - to find the effective heat losses Q_{loss} (from the aluminium specimen to the surrounding), the effective heat convection coefficient h (between the specimen and the heater), the resultant emissivity ε of the aluminium alloy and the thermal capacity $c_p(T)$ of the aluminium alloy. This last one is temperature depended (discretized with respect to the temperature by linear basis functions). Therefore the unknown vector of parameters is $\mathbf{a} = (c_p(T)\ ,\ h,\ \varepsilon,\ Q_{loss}\ )$., where $A_{\Gamma 0}$ is the sum of the areas of the ends of the specimen through which the heat losses, to the surrounding, happen.

Tests have been conducted at Helsinki University of Technology in a project dealing with the high temperature properties of aluminium [7]. Temperature of the oven is controlled in certain way to obtain a constant temperature of the specimen during the test or specimen temperature that has a certain rate, Fig. 5b) and 6.

The unknown parameter vector **a** have been solved fitting all the three tests (with temperatures shown on figures 5a) and 6) were used simultaneously as data. It was found that h = 10.4 W/m^2 K, ε = 0.11 and Q_{loss} = -0.74 W. The calculated specific heat capacity of the aluminium allow is shown in Fig 5a) as a function of the specimen temperature. These calculated values seem to be reasonable (for pure Al values of c_p = 900 J/kg K at 20 °C are given in literature). The measured temperatures of the oven and specimen and also the calculated temperatures of the aluminium specimens are shown in the Figures 5a) - 6.

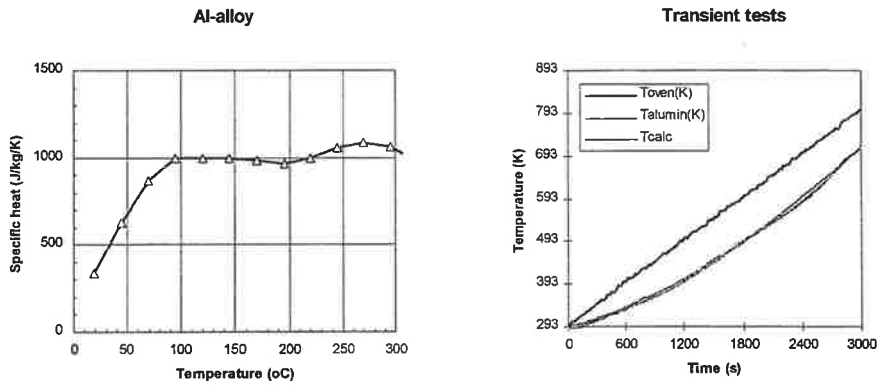


Figure 5 a) Calculated heat capacity of the aluminium alloy. b) Measured and calculated temperatures of aluminium specimen in heating oven, transient test. The upper curve represents the temperature of the oven.

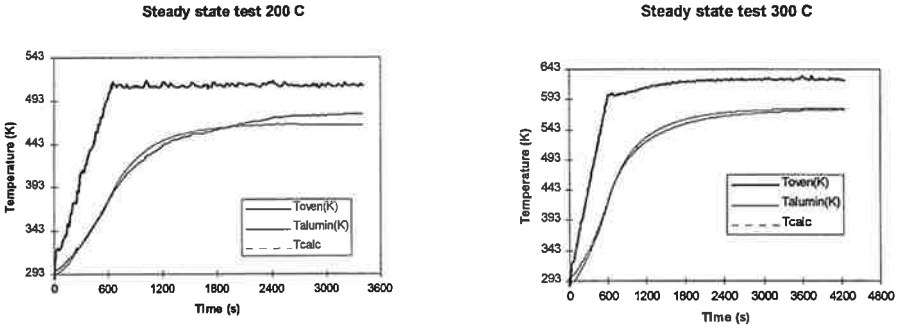


Figure 6 a) and b) Measured and calculated temperatures of aluminium specimen in heating oven in two different tests # 2 and # 3. The upper curves are the temperatures of the oven. The lower curves, those of the aluminium specimen (calculated and measured).

Thermal conductivity of the mineral wool

The thermal conductivity as a function of the temperature is determined from 'full scale' fire tests. The heat loses are also estimated. Thus the problem is to estimate the thermal conductivity and the 'heat loses ' terms appearing in the energy conservation equation. The heat loses , in the direction perpendicular to the cross section, are the terms Q_1 (through steel) and Q_2 (Wool). These terms are due to the treatment of the original 3-D heat conduction problem as of 1-D problem.

Therefore, the inverse problem consists of estimating the thermal conductivity of the thermal insulation (rock wool) as a function of the temperature using data (measured temperatures) from 'full scale' fire tests of insulated steel structure, i.e. $\mathbf{a} = (\lambda(T)\ Q_1\ Q_2)$.

Consider the case of an insulated steel structure, Figure 7. The direct problem is now dealt as one dimensional problem using three elements. One element is used for the steel part and two linear elements for the insulation part. For the steel part it is assumed that the temperature is uniform (one basis function $N_1 = 1$).

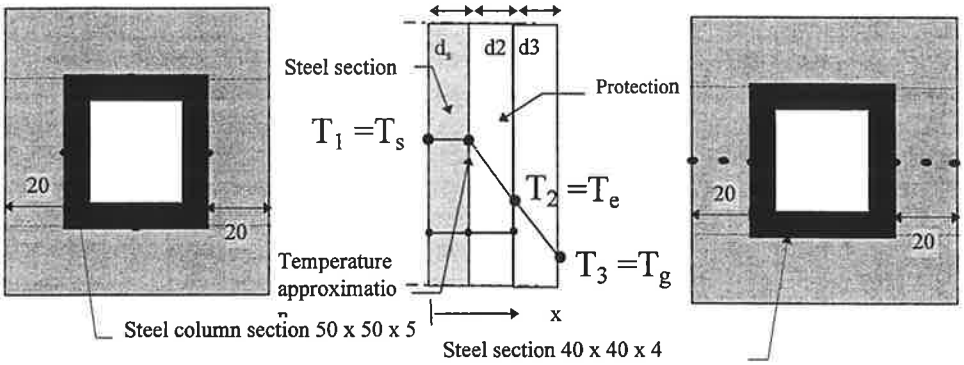


Figure 7. The cross section of the test columns and one dimensional idealisation of insulated steel structure in fire, linear elements at the fire protection.

Fire resistance tests on steel columns were performed at the VTT Fire [8]. The columns were rectangular hollow sections with the cross-sectional dimension (RHS 40x40x4 and RHS 50x50x5). The length of the columns was 900 mm with the end plates. The columns were protected with 20 mm thick rock wool (density 220 kg/m^3). The specimen was applied to the fire exposure of 15 °C/min during 65 min, Fig. 8b.

The temperatures of each column and of the furnace gas were measured at three cross-sections using a total around 24 thermocouples.

In the first case only the temperature of the steel section was measured and in the second case also temperature of the fire protection at the centre was measured (Fig. 7b). In both cases the direct problem was formulated using two linear elements in the fire protection as

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{pmatrix} \dot{T}_1 \\ \dot{T}_2 \end{pmatrix} + \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = - \begin{pmatrix} K_{13} T_g + C_{13} \dot{T}_g \\ K_{23} T_g + C_{23} \dot{T}_g \end{pmatrix}. \tag{17}$$

The system (17) of ODE´s was integrated using explicit Euler scheme.

Integration of the heat conductivity matrix $\mathbf{K}^e$ and of the force vector $\mathbf{f}^e$ are performed using one Gauss integration point at the centre of the elements. The element capacity matrix $\mathbf{C}^e$ was integrated using two-points Newton-Cotes integration (at the nodal points of the element, $\xi = -1$ and $\xi = +1$ in Eq. (17)) in order to get a diagonal matrix ($\mathbf{C}ij = 0$, when i different from j, i.e. C_{12}, C_{13}, C_{23} *and* C_{21} in Eq. (17) are all zero). In this way we avoid unstable numerical differentiation of the T_g in Eq. (17), since the gas temperature is very noisy as seen on Fig. 8b (the upper curve). With Dirichlet boundary condition $T_3 = T_g$ (surface temperature of the fire protection is same as gas temperature).

Following values of parameters were assumed in the calculation: density of the fire protection $\rho_p = 220$ kg/m^3, specific heat of the protection c_p=1000 J/kgK, density of steel $\rho_s = 7850$ kg/m^3 and specific heat of the steel c_p=540 J/kgK. The mean of the measured temperatures of the steel section and of the insulation (at the midpoint) at the centre of the column were used as collocation points.

Thus, the effective thermal conductivity $\lambda(T)$ was found as the regularised solution of the constrained minimisation problem

$$\min\left(\left\| \tilde{T}_{test}^{\ \delta}(t) - \tilde{T}_{calc.}(\vec{a};t) \right\|^2 + \alpha \left\| \mathbf{L}\vec{\lambda} \right\|^2 \right), \quad with\ \lambda_j \in D(\vec{\lambda}) \tag{18}$$

where the operator $\mathbf{L}$ is the central difference discretization of the second derivatives $\partial^2\lambda\ (T)/\partial\ T^2$ of the thermal conductivity. The accuracy measurements of the temperature was estimated to be $\left\| \tilde{T}^{\delta}(t) - \tilde{T}(t) \right\| \le \delta \approx \int_{t_\infty} 10\ ^oC\, dt$ when applying the Morozov discrepancy principle.

The solution of the problem is presented in Fig. 8a) and compared with the values provided by the producer of rock wool (the dashed line). For the heat loss one get (Q_1,Q_2) = (-2.2, -0.4) W. The calculated temperature history (for one test) is compared to the experimental one in Fig. 8b. Legends for Figure 8a:

- The balls (column: 40x40x4), calculated using two elements and two collocation points (the steel and the mineral wool)
- The triangles (column: 40x40x4), shows the same as the balls but only one collocation point was used (the steel temperature)
- The squares represent the case where two data tests were used simultaneously (column 40x40x4 with two collocation points: the steel and the wool at the midpoint and column 50x50x5 with one collocation point: the steel). The temperature calculations were made using two elements for both columns.
- The dashed line shows values provide by the producer.

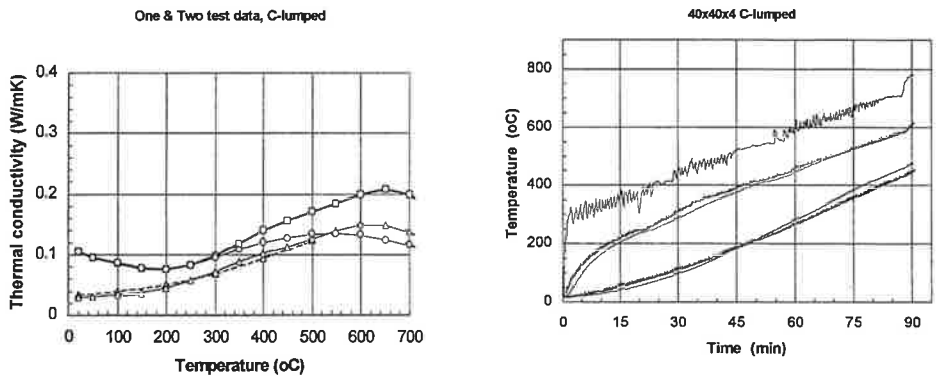


Figure 8. a) Calculated thermal conductivity of mineral wool using different number of elements and collocation points with test data (columns: 40x40x4 and 50x50x5). b) The calculated temperatures (thick) and the measured ones (thin) for the case of 40x40x4 column (the temperature of the protection at the midpoint and of the steel). The upper curve represents the gas temperature.

CONCLUSIONS AND ACKNOWLEDGEMENTS

Several applications of the regularised output least square method (ROLS) to parameter identification of heat transfer in structures were presented. Although the examples shown consist only of a one or two ODE, the method is directly applicable to system of ODE obtained by semi-discretization of the variational formulation of the general heat conduction problem using a finer FE-mesh. With a finer FE-mesh (in the solution of the direct problem) the calculated material properties are assumed to converge to the 'right' ones when the distributed parameters are discritized using an optimal mesh.

The applications presented in this paper were developed in "*VTT/STEEL*" programme funded by VTT Building Technology and " *Properties of aluminium at high temperatures*" programme at Helsinki University of Technology funded by the Technology Development Centre of Finland (TEKES).

REFERENCES

[1] Eriksson, K. Estep, D. Hansbo P. Johnson C. 1996. *Computational Differential Equations*. Studentlitteratur, Lund, Sweden. pp. 530.

[2] C.W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Wieweg Mathematics for Scientists and Engineers, Vieweg, 1993 pp.151.

[3] Tikhonov, A.N. and Arsenin V.Y. 1977. *Solutions of Ill-Posed Problems*. Wiley, New York, 1977.

[4] Hansen, P.C. 1990. *Analysis of discrete ill-posed problems by means of the L-curve*. Technical Report MCS-p157-0690. Mathematics and Computer Science Division, Argonne National Laboratory.

[5] Hansen, P.C. and O'Leary D.P. 1991. *The use of the L-curve in the regularization of discrete ill-posed problems*. Technical Report UMIACS-TR-91-142. University of Maryland. Maryland. p. 23.

[6] O. Pettersson, K. Ödeen, *Brandteknisk dimensionering*, Liberförlag, Stockholm 1978, pp.181

[7] Myllymäki, J. 1997. *Mechanical behaviour of aluminium alloys at elevated temperatures*. Helsinki University of Technology (to be published)

[8] Ala-Outinen, T. and Oksanen, T. 1997. *Stainless steel compression members at fire*, VTT Research Notes (to be published).

MERIJÄÄN PURISTUSLUJUUDEN RIIPPUVUUS KUORMITUSSUUNNASTA JA C-AKSELISTA

Eila Lehmus
VTT Rakennustekniikka
PL 18071, 02044 VTT

TIIVISTELMÄ

Tässä kirjoituksessa esitetään osa vuonna 1995 VTT Rakennustekniikassa tehtyjen merijään puristuslujuuskokeiden tuloksista. Koejärjestelyihin ja koekappaleiden valmistukseen ja säilytykseen kiinnitettiin erityistä huomiota. Koesarjan tarkoituksena oli selvittää mm. jääkiteiden suuntautumisen (c-akselin) vaikutus puristuslujuuteen. Kokeet tehtiin vakiomuodonmuutosnopeudella 10^{-3} 1/s lämpötilassa - 10 °C. Tulosten lukumäärä on liian pieni tarkan tilastollisen tarkastelun tekemiseksi. Tulosten perusteella voidaan kuitenkin todeta, että sekä kuormitussuunnalla että c-akselilla on merkitystä jään puristuslujuuteen.

1. JOHDANTO

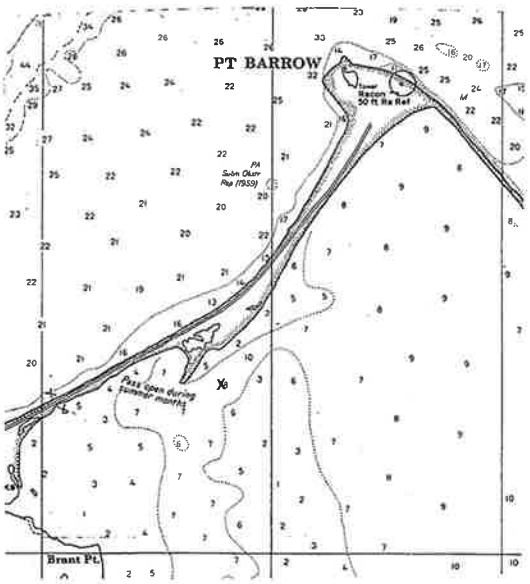
Jään lujuusominaisuuksia on tutkittu paljon ja tiedetään, että ne riippuvat monesta tekijästä. Eräitä merkittävimpiä tekijöitä ovat lämpötila, kuormitusnopeus ja jään rakenne. Lisäksi on jo aikaisemmin tutkittu kuormitussuunnan vaikutusta jään lujuuteen. Tällöin kokeet on kuitenkin tehty joko jään kasvusuunnassa tai kohtisuorassa jään kasvusuuntaa vastaan. Jäässä olevat kiteet voivat tietyissä olosuhteissa suuntautua siten, että niiden c-akselit ovat samansuuntaisia, jolloin kuormitussuuntaa tarkasteltaessa ei enää riitä näkökulmaksi jään kasvusuunta, vaan uutena asiana on otettava huomioon myös jääkiteiden suuntautuminen.

Rakenteita vastaan liikkuessaan jääkenttä voi murtua monella tavalla. Kuormituksen kasvaessa osa jääkentästä on puristuksen alaisena, jolloin kuormien arviointia ja mallintamista ajatellen on hyvä tietää puristuslujuudet kolmessa suunnassa. Jääkentän alemmissa kerroksissa jää on yleensä anisotrooppista ja pystykiteistä.

Kirjallisuudessa on esitetty melko vähän vastaavan tyyppisiä koetuloksia, koska suuri osa tehdyistä kokeista on suoritettu laboratoriojäälle. Laboratorio-olosuhteissa jäädytetty jää ei yleensä ole rakenteeltaan suuntautunutta. Joitakin tuloksia on kuitenkin saatavissa [1, 2 ja 3] ja vertailut osoittavat joitakin yhdenmukaisuuksia. Koemenetelmillä sekä jään säilytyslämpötiloilla näyttäisi olevan merkitystä tuloksiin.

2. KOEMENETELMÄT

Jäänäytteet kairattiin maaliskuussa 1994 Elson Lagoonista Alaskasta läheltä Barrowia (kuva 1.) Näytteet otettiin alueelta, jossa hieman syvempi kanava saa aikaan virtauksen. Jäänalainen virtaus puolestaan aiheuttaa jääkiteiden suuntautumista [4].



Kuva 1. Jäänäytteiden kairauspaikka Elson Lagoonilla, lähellä Barrowia Alaskassa.

Näytteitä kairattiin sekä jään kasvusuunnassa että kasvusuuntaa vastaan kohtisuoraan. Näytteiden otto tapahtui -25 °C:een lämpötilassa, jolloin suolaveden valumista pois näytteestä ei päässyt tapahtumaan. Näytteitä säilytettiin alle -23 °C:een lämpötilassa. Näytteet kuljetettiin lentokoneella Alaskasta Suomeen. Myös kuljetuksen aikana lämpötila pysyi huolellisen pakkaamisen ja hiilihappojään avulla selvästi alle -23 °C:een.

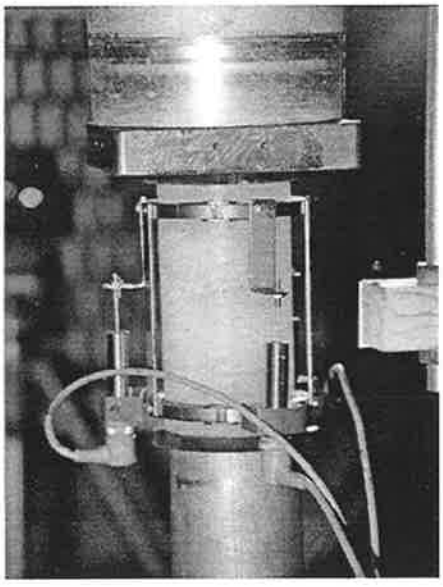
Jään suolapitoisuuden ja kiderakenteen sekä kiteiden suuntautumisen selvittämiseksi jokaisen näytteen päästä sahattiin noin senttimetrin paksuinen kappale. Siitä hiottiin mikrotomilla ohuthie (kuva 2.), josta polarisaatiolevyjen ja universaalipöydän avulla määritettiin sekä kiteiden koko että niiden c-akselien suunnat. Kidenäytteet valokuvattiin myöhempää tarkastelua varten.

Kuva 2. Jään ohuthie kiderakenteen ja c-akselin määrittämistä varten.

Puristuskoekappaleet valmistettiin -25 °C:een lämpötilassa ja niiden halkaisija oli 70 mm ja korkeus 150 mm. Koekappaleiden annettiin tasaantua koelämpötilassa juuri sen verran, että kappaleiden lämpötila oli tasaantunut. Tämän varmistamiseksi mitattiin vertailukappaleen lämpötilaa. Lämpötilan muutos -10 °C:een koelämpötilaan vei 2,5 tuntia.

Lämpötilan tarkan seurannan tarkoituksena oli estää suolaveden aiheuttamat muutokset jään rakenteessa. Suolaveden aiheuttamia muutoksia jäähän (mm. mikrohalkeamien syntymistä) on kuvattu tarkemmin lähteessä [5].

Puristuskokeet tehtiin VTT Rakennustekniikan pakkashuoneessa. Tasaisen puristuksen varmistamiseksi koekappaleiden päät hiottiin yhdensuuntaisiksi. Päiden tasaisuutta ja yhdensuuntaisuutta mitattiin profilometrillä. Sallittu poikkeama oli ±0,15 mm. Lisäksi kappaleen päiden ja kuormitustelineen väliin asennettiin 8 mm:n kumilevyt ja kuormitustelineen toinen pää nivelöitiin. Kokeissa pyrittiin saavuttamaan vakio muodonmuutosnopeus 10^{-3} 1/s ohjaamalla kuormituslaitteen siirtymää. Siirtymäohjaukseen käytettiin kolmen siirtymäanturin keskiarvoa. Kuvassa 3 on esitetty koejärjestelyt. Kaikki kokeet kuvattiin videolle, josta voitiin jälkikäteen tarkastaa ensimmäisen murtuman luonne ja ajankohta.



Kuva 3. Puristuskokeen koejärjestelyt.

3. KOETULOKSET

3.1. Kuormitus jään kasvusuunnassa

Jään kasvusuunnassa otettujen näytteiden ylälaita alkoi 42 cm:n syvyydeltä jään pinnasta, jossa jään rakenne oli selvästi pystykiteistä. Kuormitussuunta oli kohtisuorassa jään c-akselia vastaan. Näytteiden suolapitoisuus vaihteli välillä 4 - 5,5 ‰. Koesarja käsitti viisi koekappaletta, jotka murtuivat halkeamalla lähinnä kiderajoiltaan. Murtumista vastaavat puristuslujuudet on esitetty taulukossa 1. Taulukon arvoista laskettu keskiarvo on 7,97 MPa. Halkeilun jälkeen kappale kesti yleensä vielä kuormitusta kunnes yksittäiset kiteet murtuivat. Vaihtelu kiteiden lukumäärässä, koossa ja pituudessa aiheutti tuloksiin suuren hajonnan.

Taulukko 1. Jään kasvusuunnassa kuormitettujen koekappaleiden puristuslujuudet.

Koekappaleen numero	Suolapitoisuus [‰]	Lämpötila [°C]	Puristuslujuus [Mpa]
2	4	-10	6.75
4	5.5	-10	7.27
5	5	-10	10.39
6	5	-10	9.61
23	5	-10	5.84

3.2. Kuormitus horisontaalisesti jään kasvusuuntaa vastaan kohtisuoraan

Vertailun vuoksi tehtiin vastaavia kokeita myös jään kasvusuuntaa vastaan kohtisuorassa suunnassa. Koekappaleet otettiin 65 cm:n syvyydeltä jään pinnasta. Jään kasvusuunnassa kuormitettujen koekappaleiden keskikohta oli suunnilleen samalta syvyydeltä. Vaakasuuntaiset koekappaleet valmistettiin siten, että kuormitussuunta oli joko c-akselin suuntainen tai sitä vastaan kohtisuora. Tuloksia tarkasteltaessa puristuslujuudeksi valittiin ensimmäiseen murtumaan liittyvä arvo, vaikka se ei kaikissa tapauksissa ollut suurin.

C-akselin suuntaisesti kuormitettiin viisi koekappaletta, joiden keskiarvoksi saatiin 7,27 MPa. Tulos vastaa hyvin lähteessä [5] esitettyjen 42 cm:n syvyydeltä otettujen koekappaleiden tulosten keskiarvoa, joka on 7,23 MPa. C-akselia vastaan kohtisuorassa suunnassa tehtiin neljä koetta, joiden tulosten keskiarvo oli 5,84 MPa. Yksittäisten kokeiden tulokset on esitetty taulukossa 2.

Wang [3] on saanut samantyyppiselle merijäälle arvoiksi 8 MPa c-akselin suuntaisella kuormituksella ja 6,5 MPa c-akselia vastaan kohtisuorassa kuormitussuunnassa. Lämpötila oli molemmissa koejärjestelyissä sama.

Taulukko 2. Jään kasvusuuntaa vastaan kohtisuorassa suunnassa kuormitettujen koekappaleiden puristuslujuudet.

Koekappaleen numero	Suolapitoisuus [‰]	Kuormitus-suunta	Puristuslujuus [Mpa]
68	5	c-akselin suuntainen	8.83
71	5	c-akselin suuntainen	7.27
77	4.5	c-akselin suuntainen	8.05
80	5	c-akselin suuntainen	5.45
82	5	c-akselin suuntainen	6.75
86	5	kohtisuorassa c-akselia vastaan	5.97
93	4.5	kohtisuorassa c-akselia vastaan	4.42
96	5.5	kohtisuorassa c-akselia vastaan	7.27
98	5	kohtisuorassa c-akselia vastaan	5.71

4. JOHTOPÄÄTÖKSET

Kuormitettaessa jään kasvusuuntaan, c-akselia vastaan kohtisuorassa suunnassa, puristuslujuuden keskiarvoksi saatiin 7,97 MPa, mikä on paljon suurempi kuin vastaava arvo jään kasvusuuntaa vastaan kohtisuorassa, 5,84 MPa. C-akselin suunnalla ei tällä tavoin pääteltynä näyttäisi olevan merkitystä jään puristuslujuuteen, koska pystysuuntainen arvo on 36 % suurempi kuin vaakasuuntainen arvo. Jään anisotrooppisuudesta johtuen puristuslujuuden suuriin eroihin on todennäköisesti olemassa muita syitä, joista ehkä selkein on suolataskujen sijainti ja muoto. Lähteessä [5] on kuvattu suolataskujen muodon muuttumista lämmitys-/jäädytysvaiheiden aikana, jolloin suolataskujen reunoille saattaa muodostua jännityskeskittymiä, jotka vaikuttavat puristuslujuuden arvoihin enemmän kuin c-akselin suunta. Vaakasuuntaisessa kuormituksessa ero on kuitenkin selvä, c-akselin suuntainen kuormitus antaa arvoksi 7,27 MPa ja vastaava c-akselia vastaan kohtisuora vain 5,84. Eroa on siis 24 %.

Vaikka tehtyjen kokeiden määrä on pieni voidaan näiden koetulosten perusteella todeta, että jään puristuslujuuden arvo riippuu paitsi kuormitussuunnasta myös jään c-akselista. Merkitystä tällä on erityisesti silloin, kun tarkastellaan merirakenteille kohdistuvia jääkuormia alueilla, joissa c-akselin suuntautumista on tapahtunut. Laboratoriojäälle tehdyissä kokeissa c-akselin suuntautumista on ollut vaikea ottaa huomioon.

LÄHTEET

1. Peyton, H.R., (1966), "Sea Ice Strength", Geophysical Institute, University of Alaska, Report UAG R-182, Final Report, Office of Naval Research, Contract No. Nonr 2601(01), December 1966.

2. Saeki, H., Nomura, T., and Ozaki, A., (1978), "Experimental Study on the Testing Methods of Strength and Mechanical Properties for Sea Ice",

Proceedings IAHR Symposium of Ice Problems. Part 1. Luleå Sweden, August 7-9, 1978, pp. 135 - 149.

3. Wang, Y.S., (1979), "Sea Ice Properties", Technical Seminar on Alaskan Beaufort Sea Gravel Island Design, Exxon U.S.A., Anchorage, Alaska, October 15, 1979.

4. Weeks, W.F. and Gow, A.J., (1978), "Crystal Alignments in the Fast Ice Arctic Alaska", Journal of Geophysical Research 85(C2), pp. 1137 - 1146.

5. Lehmus, E., Kärnä, T., Tanabe, A., Yoshizawa, M., Ishibashi, Y. and Sackinger, W., (1996), "Compressive strength of natural sea ice in horizontal loading", Proc. of the sixth international offshore and polar engineering conference. Los Angeles, USA, May 26-31, 1996. Vol.II, pp. 285 -290.

ANALYSIS OF SANDWICH PLATES UNDERGOING LARGE DEFLECTIONS FOR TAILORING

MARKKU LAITINEN, MIKA JURVAKAINEN, ANTTI PRAMILA
University of Oulu
Department of Mechanical Engineering
Engineering Mechanics Laboratory

ABSTRACT

The aim of this study was to develop a suitable solution algorithm for the bending problem of a laterally loaded geometrically nonlinear sandwich plate for future implementation into the LAMINV-program system which solves the problem by inverse technique. The sandwich plate analysis includes the Reissner-Mindlin hypothesis with the von Karman displacement field. The solution method for the nonlinear deflection of the sandwich plate is the Ritz method with direct minimization of the total potential energy. Preliminary comparison with existing FEM-solution indicates reasonable agreement

INTRODUCTION

The purpose of this study is to develop a solution algorithm to be used in the inverse method in the design of sandwich plates undergoing large deflections. The geometrical nonlinearities become significant, when deflection of the plate exceeds the half of the thickness of the plate. The favorable stiffness and weight properties of composite materials are not fully utilized, when a linear deflection theory is used.

The use of composite materials offers the designer many possibilities to tailor the response of the structure. By changing the design variables, e.g., the ply orientations and/or thicknesses, and the thickness of the core, different structural properties can be created for a sandwich structure. When specific properties are to be created, the trial and error method is likely to fail. In the solution of the inverse problem, the designer wishes to find the design variables so, that the structure deforms in a specific way for given loads. The solution of the inverse problem often leads to a minimization problem, where one minimizes the difference between the desired value and the calculated value. Generally the solution of the inverse problem is not unique; several design variable combinations can fulfill the demands.

Various plate theories are suggested for the analysis of sandwich plates. Among the most widely used are the first order shear deformation theory (Reissner-Mindlin), higher order theories and the discrete layer theory. When geometrical nonlinearities are to be taken into account, each of these theories can be modified to include the von Karman type large deflection assumptions. From the computational point of view, the Reissner-Mindlin theory is of course the most easily applicable. The higher order theories and especially the discrete layer theory are computationally expensive in practical cases.

For a sandwich plate analytical closed form solutions are found only in few special cases and numerical methods are to be used. The designer must choose the most efficient discretization method in order to perform the solution of the inverse problem with resonable computational effort. The most suitable discretization methods for sandwich structures are the finite element method (FEM) and the traditional Ritz method. The advantage of the finite element method is that more complex geometries can be analyzed and the disadvantage is, that the CPU-time increses considerably, when more elements (more d.o.f.) are used. The advantage of the Ritz method is that fewer degrees of freedom are needed, and the disadvantage is, that the method is suitable only for simple geometries. Since the geometry of the structure investigated here is very simple and the solution is required several times during the minimization, the Ritz method is chosen for the discretization.

LAMINATE THEORY

Let us take a brief look at the classical laminate theory in order to illustrate the macro-mechanical behaviour of some important types of laminates. The force-deformation equation of a general laminate is

$$\begin{Bmatrix} N_x \\ N_y \\ N_{xy} \\ M_x \\ M_y \\ M_{xy} \end{Bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{16} & B_{11} & B_{12} & B_{16} \\ A_{12} & A_{22} & A_{26} & B_{12} & B_{22} & B_{26} \\ A_{16} & A_{26} & A_{66} & B_{16} & B_{26} & B_{66} \\ B_{11} & B_{12} & B_{16} & D_{11} & D_{12} & D_{16} \\ B_{12} & B_{22} & B_{26} & D_{12} & D_{22} & D_{26} \\ B_{16} & B_{26} & B_{66} & D_{16} & D_{26} & D_{66} \end{bmatrix} \begin{Bmatrix} \varepsilon_x^0 \\ \varepsilon_y^0 \\ \gamma_{xy}^0 \\ \kappa_x \\ \kappa_y \\ \kappa_{xy} \end{Bmatrix}, \tag{1}$$

where the **A** is the extensional stiffness matrix, **B** is bending-extension coupling and **D** is the bending stiffness matrix. The presence of the **B**-matrix implies coupling between bending and extension of a laminate. It is impossible to pull on a such laminate without at the same time bending and/or twisting it. The stacking sequence of these laminates is unsymmetric with respect the geometric mid-plane of the laminate. Usually, the symmetric stacking sequencies are preferred for obvious reasons. For these laminates, the **B**-matrix vanishes and extension and bending are decoupled. If the laminate has got an angle-ply stacking sequency, there still exists at least the bending-twisting coupling, namely the terms D_{16} and D_{26} (transversely isotropic layers). This coupling still prevents the analytical closed form solution.

The simpliest case for orthotropic laminates is the specially orthotropic case. In that case, each layer is stacked symmetrically with respect to the geometric mid-plane with 0° and 90° orientation angles (additionally, there can also be isotropic layers in the laminate, typically the core, provided, that the stacking sequence is still a symmetric one). For this type of laminate the **B**-matrix and the coupling terms A_{16}, A_{26} D_{16}, and D_{26} vanish. This type of laminate is often used, since it is easily manufactured. The drawback is, that the designer has to give up the possibility to use ply orientations as design variables.

When Reissner-Mindlin theory is used, shear force-strain relations are needed

$$\begin{Bmatrix} Q_y \\ Q_x \end{Bmatrix} = \begin{bmatrix} k_1^2 A_{44} & k_1 k_2 A_{45} \\ k_1 k_2 A_{45} & k_2^2 A_{55} \end{bmatrix} \begin{Bmatrix} \gamma_{yz} \\ \gamma_{xz} \end{Bmatrix}, \tag{2}$$

where k_1 and k_2 are the shear correction factors. These factors can be derived using energy principles for constant and parabolic transverse stress distributions and imposing the equality of these two, as is done in /1/, and /2/. With proper correction factors, sandwich plates can be analyzed quite accurately.

PLATE THEORY USED

With the present day computer capabilities, the suitable sandwich plate theory is the Reissner-Mindlin (or YNS) plate theory. In this theory, the planar sections remain straight, but are allowed to rotate during deformation. Since this study is focused on the large deflection analysis of composite plates, the following deformation kinematic relations include the von Karman -type nonlinearity.

The Reissner-Mindlin-type (YNS) displacements in the plate can be written as, /3/,

$$\begin{aligned} u(x,y,z) &= u^0(x,y) + z\,\psi_x(x,y) \\ v(x,y,z) &= v^0(x,y) + z\,\psi_y(x,y) \\ w(x,y,z) &= w(x,y), \end{aligned} \tag{3}$$

where u, v, and w are the displacements anywhere in the plate; u^0 and v^0 are the displacements of the mid-plane, and ψ_x and ψ_y are the rotations.

Since we investigate the large deflection of laminated plates, the nonlinear strain-displacement relations are used, /4/,

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} + \frac{\partial u_k}{\partial x_i}\frac{\partial u_k}{\partial x_j} \right). \tag{4}$$

For brevity we have used in the above equation the indicial notation in a standard way, i.e. x_1=x, x_2=y, ... , u_3=w. If the in-plane displacement remain small and the transverse displacements (deflection) moderate, the in-plane rotations can be neglected and the usual von Karman equations for large deflections are obtained.

Using equation (3), the final form of the strain-displacement relation is

$$\varepsilon_x = \frac{\partial u^0}{\partial x} + \frac{1}{2}\left(\frac{\partial w}{\partial x}\right)^2 + \frac{\partial w_0}{\partial x}\frac{\partial w}{\partial x} + z\frac{\partial \psi_x}{\partial x}$$

$$\varepsilon_y = \frac{\partial v^0}{\partial y} + \frac{1}{2}\left(\frac{\partial w}{\partial y}\right)^2 + \frac{\partial w_0}{\partial y}\frac{\partial w}{\partial y} + z\frac{\partial \psi_y}{\partial y}$$

$$\gamma_{xy} = \frac{\partial u^0}{\partial y} + \frac{\partial v^0}{\partial x} + \frac{\partial w}{\partial x}\frac{\partial w}{\partial y} + \frac{\partial w_0}{\partial x}\frac{\partial w}{\partial y} + \frac{\partial w_0}{\partial y}\frac{\partial w}{\partial x} + z\left(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x}\right), \tag{5}$$

$$\gamma_{xz} = \psi_x + \frac{\partial w}{\partial x}$$

$$\gamma_{yz} = \psi_y + \frac{\partial w}{\partial y}$$

where engineering notation is used and w_0 is the initial deflection of the plate.

RITZ DISCRETIZATION

The principle of the minimum potential energy states: "Of all the admissible displacement fields, the one that minimizes the total potential energy of the structure quarantees equilibrium." Mathematically the problem is formulated as follows

$$\min_{\mathbf{u}} \Pi = \min_{\mathbf{u}} \left(\frac{1}{2}\int_V C_{ij}\,\varepsilon_i\,\varepsilon_j\,dV - \int_S f_i\,u_i\,dS\right), \tag{6}$$

where **u** denotes the displacement field and the first integral is the strain energy, the second integral being the potential of the external load set.

Using all assumptions above, the strain energy of the plate can be expressed as a function of the five unknowns u^0, v^0, w, ψ_x and ψ_y. The strain energy of plate is then

$$U = \frac{1}{2}\int_\Omega \left(\varepsilon^T \mathbf{A}\,\varepsilon + 2\varepsilon^T \mathbf{B}\,\varkappa + \varkappa^T \mathbf{D}\,\varkappa + \gamma^T \overline{\mathbf{A}}\,\gamma\right) d\Omega, \tag{7}$$

where Ω denotes the mid-plane of the plate, the vector ε contains the midplane strains, the κ contains the curvatures, the vector γ contains the transverse shear strains and the matrices have their standard meanings (see eqs. (1) and (2)).

The mid-plane strains can be separated into linear and nonlinear parts ε_1 and ε_{N1}. Then, the strain energy can be broken into following components, /5/, each having a specific meaning. The in-plane stretching of the plate (linear membrane behavior)

$$U_1 = \frac{1}{2}\int_\Omega \left(\varepsilon_1^T \mathbf{A}\,\varepsilon_1\right) d\Omega, \tag{8}$$

the geometric coupling between in-plane deformations and the deflection

$$U_2 = \frac{1}{2}\int_\Omega \left(2\varepsilon_1^T \mathbf{A}\,\varepsilon_{N1}\right) d\Omega, \tag{9}$$

the fourth order terms in the deflection w

$$U_3 = \frac{1}{2}\int_\Omega \left(\varepsilon_{N1}^T \mathbf{A}\,\varepsilon_{N1}\right) d\Omega, \tag{10}$$

the material coupling between in-plane deformations and the deflection

$$U_4 = \frac{1}{2}\int_\Omega \left(2\varepsilon_1^T \mathbf{B}\,\varkappa\right) d\Omega, \tag{11}$$

the material/geometric coupling between in-plane deformations and the deflection

$$U_5 = \frac{1}{2}\int_\Omega \left(2\varepsilon_{N1}^T \mathbf{B}\,\varkappa\right) d\Omega, \tag{12}$$

the bending stiffness of the plate

$$U_6 = \frac{1}{2}\int_\Omega \left(\varkappa^T \mathbf{D}\,\varkappa\right) d\Omega, \tag{13}$$

and the transverse shear stiffness

$$U_7 = \frac{1}{2}\int_\Omega \left(\gamma^T \overline{\mathbf{A}}\,\gamma\right) d\Omega. \tag{14}$$

These strain energy equations are greatly simplified, when the stacking sequence of the covering laminates is restricted to be symmetric with recpect to the midplane of the plate and even more, when specially orthotropic laminates are considered.

The potential of the external load set is

$$W = \int_\Omega p\,w\,d\Omega + \int_S \mathbf{N}^T \cdot \mathbf{u}\,dS, \tag{15}$$

where p is the transverse pressure, and $\mathbf{N}^T$ is the in-plane loads.

Additionally several other terms are introduced from the initial imperfection w_0. These terms are obtained by substituting w_0 for w in appropriate places.

BOUNDARY CONDITIONS

The potential energy approach along the Ritz method is used here, so only the geometric (essential) boundary conditions must be satisfied, although, if natural boundary conditions are present, the convergence is more rapid if all the boundary conditions are satisfied. The boundary conditions can be separated into two categories (n denotes normal, t denotes tangential)

1) Out-of-plane (w, ψ_n and ψ_t). Simply supported; $w = 0$; no restrictions on ψ_n and ψ_n . Clamped; $w = \psi_n = w_{,n} = 0$.
2) In-plane (u_n, u_t). Straight edge; u_n constant or zero. Stress-free edge; u_n and u_t can take arbitrary values. Uniformly loaded edge; u_n and u_t can take arbitrary values.

Further on the in-plane boundary conditions can be divided into three categories:

1) All edges straight (the stiffest case).
2) The lateral edges are free to deform and the loaded edges are kept straight.
3) All edges are free to deform.

DIRECT MINIMIZATION OF THE POTENTIAL ENERGY

Usually one would take the first variation of the total potential energy and equaling it to zero to obtain a set of nonlinear partial-differential equations. Here we use the direct minimization of the potential energy. First we use generalized coordinates q_i to obtain a discrete expression of the total potential energy. The expression is a fourth-order polynomial as follows, /5/,

$$
\begin{aligned}
F(q_i) &= F(q_i^{uv}, q_i^{w}, q_i^{r}) \\
&= \sum_i f_i q_i^{w} + \sum_i \sum_j K_{ij}^{e} q_i^{uv} q_j^{uv} \\
&+ \sum_i \sum_j \sum_k K_{ijk}^{s} q_i^{uv} q_j^{w} q_k^{w} + \sum_i \sum_j \sum_k \sum_l K_{ijkl}^{q} q_i^{w} q_j^{w} q_k^{w} q_l^{w} \\
&+ \sum_i \sum_j K_{ij}^{f} q_i^{r} q_j^{r} + \sum_i \sum_j K_{ij}^{grr} q_i^{r} q_j^{r} \\
&+ 2\sum_i \sum_j K_{ij}^{grw} q_i^{r} q_j^{w} + + \sum_i \sum_j K_{ij}^{gww} q_i^{w} q_j^{w},
\end{aligned}
\tag{16}
$$

where

$\mathbf{f}$ is the load vector due to lateral uniform pressure load,
$\mathbf{K}^e$ is linear membrane stiffness matrix,
$\mathbf{K}^s$ is nonlinear membrane stiffness matrix,
$\mathbf{K}^q$ is nonlinear bending stiffness matrix,
$\mathbf{K}^f$ is linear bending stiffness matrix,

$\mathbf{K}^{grr}$ is linear rotation stiffness matrix (rotational d.o.f. only),
$\mathbf{K}^{grw}$ is linear rotation stiffness matrix (rotational/bending coupling),
$\mathbf{K}^{gww}$ is linear rotation stiffness matrix (bending d.o.f. only)

The last three matrices are related to the transverse shear stiffness of the plate.

This function is then minimized using unconstrained minimization algorithm. This minimization algorithm should not be confused with the constrained nonlinear minimization algorithm used in the inverse problem solution. Here we wish find the unconstrained minimum of the potential energy, which gives us the nonlinear displacements of the plate. The minimum of the potential energy is found in three steps, /4/:

1) Find the descent direction for $F(\mathbf{q})$

$$\nabla F_n \cdot \mathbf{s}_n < 0, \tag{17}$$

where ∇F_n is the gradient of F at $\mathbf{q}_n$.

2) Along the line defined by $\mathbf{s}_n$, find the absolute minimum of $F(\mathbf{q})$. This operation is called the line search.

3) Increment the unknowns to that new point and start over again until convergence is reached.

There exists many suitable algorithms for this procedure, for example, the conjugate gradient method, the Newton-Raphson method and the variable metric method (quasi-Newton method). Here we choose the Broyden-Fletcher-Goldfarb-Shanno (BFGS) version of the variable metric method, /6/. The one obvious advantage of the direct minimization of the potential energy is that no incremental loading is needed as would be the case with the finite element method.

RITZ DISPLACEMENT MODES

Next we have to decide the which shape functions we use over the plate domain Ω in order to represent the displacements. With the Ritz method, the shape functions are usually taken as double trigonometric series in which the variables (x and y) are separable. Here we choose

$$\begin{aligned}
u^0(x,y) &= \sum_i \sum_j u_{ij} E_i F_j \\
v^0(x,y) &= \sum_i \sum_j v_{ij} G_i H_j \\
w(x,y) &= \sum_i \sum_j w_{ij} E_i H_j \\
\psi_x(x,y) &= \sum_i \sum_j x_{ij} T_i H_j \\
\psi_y(x,y) &= \sum_i \sum_j y_{ij} E_i Z_j ,
\end{aligned} \tag{18}$$

where the functions E_i, G_i, and T_i are trigonometric functions (series) depending on the x-coordinate and F_j, H_j, and Z_j are trigonometric functions (series) depending on the y-coordinate. These shape functions must at least satisfy the geometric boundary conditions.

NUMERICAL EXAMPLE

As a first numerical test example we have used the one found from reference /7/. The size of the plate was 2 m * 3m. The material properties are the same as in the above reference.

The transverse displacement of the sandwich plate due to uniform transverse pressure load is shown in figure 1. The reference solution of Hildebrandt and our solution have rather good overall agreement. Our solution seems to be monotonic and quite fair up to 75 kPa loading. The reference solution has more undulations and they start earlier. At first sight there is a peculiarity that the nonlinear analysis with the direct minimization gives larger deflections than the linear analysis with small loads. There is, however a physical explanation for it.

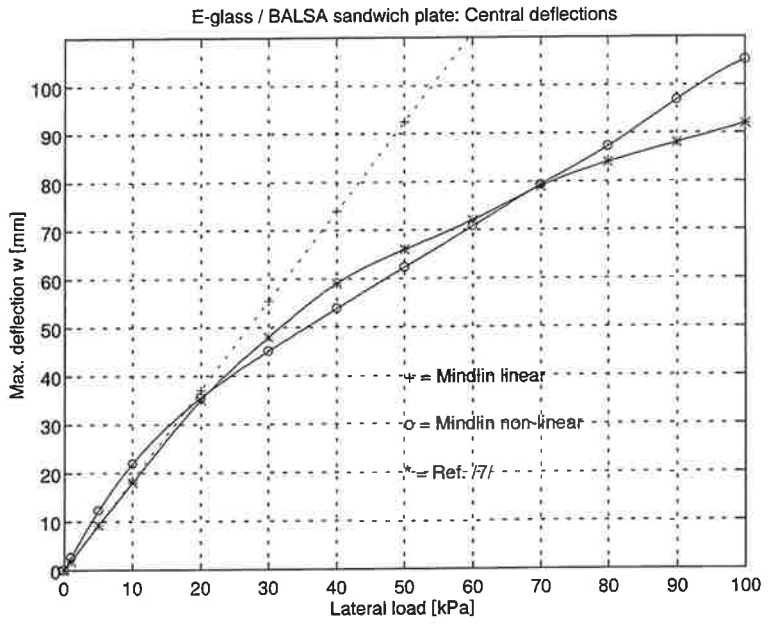


Fig. 1. Simply supported sandwich plate: linear and non-linear central deflections.

CONCLUSIONS

Our aim was to develop an algorithm for analysing sandwich plates undergoing large deflections by using the Ritz method and direct minimization of the total potential energy. The algorithm implemented seems to work rather well.

REFERENCES

1. Vlachoutsis, S. Shear correction factors for plates and shells. *International Journal for Numerical Methods in Engineering,* **33**, pp. 1537-1552 (1992).

2. Laitinen, M., Lahtinen, H., Sjölind, S-G. Transverse shear correction factors for laminates in cylindrical bending. *Communications in Numerical Methods in Engineering*, **11**, pp. 41-47 (1995).

3. Reddy, J.N. A penalty plate bending element for the analysis of laminated anisotropic composite plates. *International Journal for Numerical Methods in Engineering,* **15**, pp. 1187-1206 (1980).

4. Minguet, P.J., Dugundji, J., Lagace, P. Postbuckling behavior of laminated plates using a direct energy-minimization technique. *AIAA Journal*, **27**(12), pp. 1785-1792 (1989).

5. Minguet, P.J., *Buckling of graphite/epoxy sandwich plates*. M.S. Thesis, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology (MIT), Cambridge, MA (1986).

6. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. *Numerical Recipes*. Cambridge University Press, Cambridge, MA (1986).

7. Hildebrand, M., Visuri, M. The non-linear behaviour of stiffened FRP-sandwich structures for marine applications. Espoo, Technical Research Centre of Finland, VTT Techical Report VTT VALB 155. 53 p.

HARUSTETUN MASTORAKENTEEN OPTIMOINNISTA

Timo Turkkila
Tampereen teknillinen korkeakoulu/ Teknillinen mekaniikka
PL 589
33101 Tampere

TIIVISTELMÄ

Artikkelissa esitellään harustetun mastorakenteen optimointia. Muodostettava optimointitehtävä on monitavoitteinen ja diskreetti. Monitavoitteinen tehtävä ratkaistaan rajoitusmenetelmällä ja siinä tarvittavat diskreetit tehtävät branch and bound tai geneettisellä algoritmilla. Rakenne analysoidaan elementtimenetelmällä käyttäen lineaarista laskentamallia.

JOHDANTO

Nykyaikainen tietoliikenne tarvitsee runsaasti erilaisia mastorakenteita. Niiden analysointi ja optimointi ovat haastavia mekaniikan ja matematiikan ongelmia. Tämä artikkeli perustuu lisensiaatintyöhön [1] ja tässä yhteydessä keskitytään suhteellisen matalien, putkirunkoisten harustettujen mastorakenteiden optimointiin. Esitettäviä optimointimenetelmiä voi soveltaa myös korkeampiin ristikkorakenteisiin mastoihin, jos maston runkorakenne voidaan korvata esimerkiksi artikkelissa [2] esitetyllä ekvivalenttisella palkkielementillä.

Mastorakennetta optimoitaessa tehtävän suunnittelumuuttujia voivat olla esimerkiksi harustasojen lukumäärä, harusköysien halkaisijat, haruksien kiinnityspisteiden paikat sekä mastoputken valinta. Suunnittelumuuttujana oleva harustasojen lukumäärä tekee

optimointitehtävästä monimutkaisen, sillä muiden suunnittelumuuttujien ja rajoitusehtojen lukumäärät riippuvat harustasojen lukumäärästä. Tässä artikkelissa optimointialgoritmeina käytetään branch and bound ja geneettistä algoritmia.

Mastorakenteesta muodostetaan monitavoitteinen optimointitehtävä, joka ratkaistaan rajoitusmenetelmää käyttäen. Siinä yksi kriteeri valitaan minimoitavaksi ja loput kriteerit parametrisoidaan rajoitusehdoiksi. Rajoitusmenetelmää käytetään, koska se on osoittautunut tehokkaammaksi kuin yleisesti käytetty painokerroinmenetelmä, jossa kohdefunktio muodostetaan eri kriteerien painotettuna summana. Monitavoitteista optimointia ja sen eri ratkaisumenetelmiä on esitelty tarkemmin esimerkiksi lähteessä [3].

DISKREETIT OPTIMOINTIALGORITMIT

Tässä artikkelissa kokeillaan kahta optimointialgoritmia: branch and bound ja geneettistä algoritmia, joiden toimintaa esitellään lyhyesti tässä luvussa. Branch and bound algoritmissa käytetään jatkuvia osatehtäviä, jotka ratkaistaan jollakin jatkuvan optimointitehtävän algoritmilla. Jos osatehtävästä ei saada diskreettiä ratkaisua, siitä muodostetaan kaksi uutta osatehtävää. Tässä yhteydessä edellisen osatehtävän ei-diskreetti optimipiste jää seuraavien osatehtävien käypien alueiden ulkopuolelle. Käypiä alueita pilkotaan edelleen, kunnes osatehtävästä viimein saadaan diskreetti ratkaisu, käypä alue kutistuu pois tai osatehtävän kohdefunktion arvo on annettua ylärajaa suurempi. Ylärajana käytetään parhaan tunnetun diskreetin ratkaisun kohdefunktion arvoa.

Koska suunnittelumuuttujien ja rajoitusehtojen lukumäärä ei ole vakio, branch and bound algoritmia ei voi käyttää suoraan. Toisaalta diskreetti optimointitehtävä voidaan ratkaista täysin normaalisti, jos harustasojen lukumäärä on vakio. Tämä onnistuu, jos harustasojen lukumäärän ilmaiseva topologiamuuttuja ja muut suunnittelumuuttujat ratkaistaan erikseen. Koska topologiamuuttujia on vain yksi, eri arvojen järjestelmällinen kokeileminen on käyttökelpoinen menetelmä. Tehdyssä ohjelmassa harustasojen lukumäärälle annetaan ala- ja yläraja. Tämän jälkeen harustasojen lukumäärää kasvatetaan alarajalta lähtien. Jokaisella lukumäärämuuttujan arvolla ratkaistaan diskreetti optimointitehtävä, ja lukumäärämuuttujan arvoa suurennetaan niin kauan, kun kohdefunktion arvo paranee tai tullaan lukumäärämuuttujan ylärajalle.

Geneettisen algoritmin idea on saatu luonnon evoluutioteoriasta. Siinä seurataan tietyn eliölajin kehitystä sukupolvien kuluessa. Jokaisessa sukupolvessa parhailla yksilöillä on heikompia yksilöitä suurempi todennäköisyys saada jälkeläisiä, jolloin laji vähitellen jalostuu. Yksilöiden perimä on geeneissä, jotka muodostavat kromosomin. Uutta sukupolvea luotaessa vanhempien kromosomit katkeavat ja liittyvät uuteen järjestykseen. Samalla geeneille voi tapahtua myös mutaatioita.

Matemaattisessa optimointialgoritmissa suunnittelumuuttujavektori koodataan kromosomiksi. Yleensä käytetään binäärikoodausta, jolloin jokaista kromosomipaikkaa tavoittelee kaksi kilpailevaa geeniä 0 ja 1. Yksinkertaisuuden vuoksi yleensä käytetään vain yhtä kromosomia yksilön perimän tallentamiseen. Risteytettävät yksilöt valitaan tavallisesti käyttäen fitness-lukua, joka muodostetaan kohdefunktion ja rajoitusehtojen arvojen avulla. Risteytyksessä yksilöiden kromosomit katkaistaan yhdestä tai useammasta kohdasta ja jollakin ennalta määrättävällä todennäköisyydellä kromosomipalojen paikkoja vaihdetaan. Tällöin esimerkiksi kromosomeista 11111 ja 10000 voi tulla uudet kromosomit 11100 ja 10011. Mutaatiossa ennalta määrättävällä pienellä todennäköisyydellä geenin 1 paikalle vaihdetaan geeni 0 ja päinvastoin.

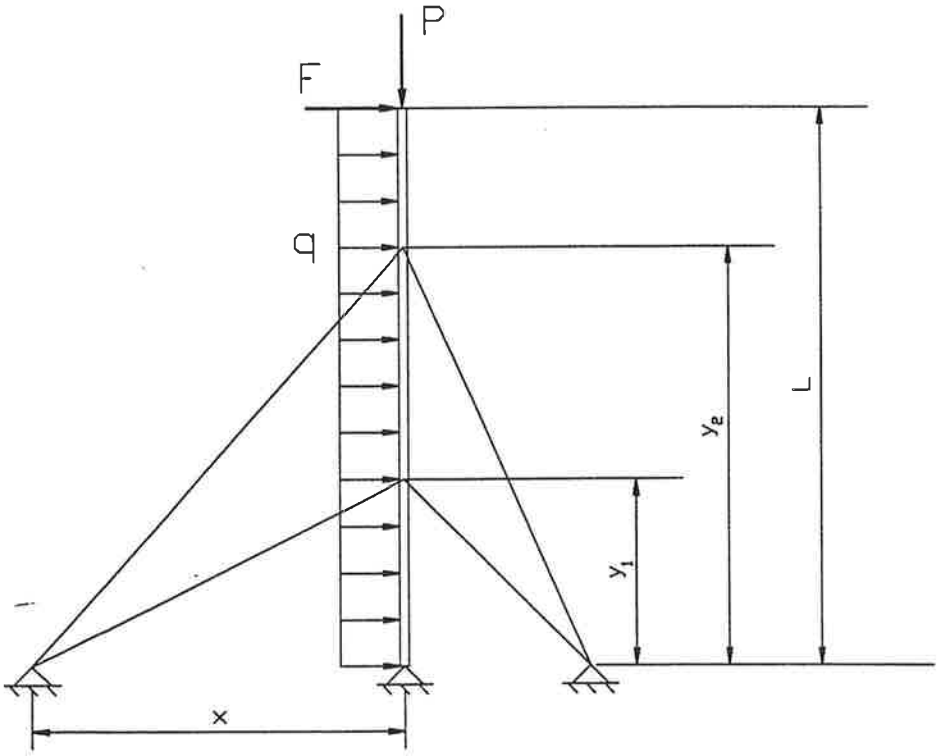
Topologiamuuttujat voidaan käsitellä geneettisessä algoritmissa varaamalla riittävän pitkä kromosomi käyttäen topologiamuuttujien arvoina niiden ylärajoja. Jos topologiamuuttujan arvo on pienempi, ylimääräiset geenit jäävät merkityksettömiksi. Tämän menetelmän käytöstä saattaa syntyä ongelmia, jos kahdella risteytettävällä yksilöllä ei ole sama harustasojen lukumäärä, jolloin jotkut geenit eivät ole kaikille yksilöille merkityksellisiä.

Maston laskentamalli muodostetaan virittämällä harustasojen väliin elementtiverkko. Tästä syystä kahta harustasoa ei saa kiinnittää samaan kohtaan ja harustasot on oltava kiinnityskorkeuksien mukaisessa järjestyksessä. Risteytykset ja mutaatiot sekoittavat hyvin tehokkaasti tätä korkeusjärjestystä, joten sen säilymisestä on huolehdittava. Korkeusjärjestys säilytetään vaihtamalla väärässä järjestyksessä olevien harustasojen paikkoja yksilön kromosomivektorissa.

Branch and bound algoritmia on esitelty tarkemmin esimerkiksi lähteessä [4] ja geneettistä algoritmia lähteissä [5] ja [6].

ESIMERKKI

Esimerkkinä on kuvassa 1 esitetty 30 m korkea mastorakenne, jonka yläpäässä on antenni. Maston alapää on niveltuettu ja harukset on kiinnitetty kolmeen maakiinnikkeeseen, jotka sijaitsevat mastosta katsottuna 120 asteen välein. Tuulikuormitus synnyttää mastoputkeen viivakuorman q ja antenniin pistevoiman F. Pistevoima on vakio, mutta viivakuormitus riippuu tuulenpaineesta p ja mastoputken ulkohalkaisijasta D kaavalla $q = pD$. Tässä yhteydessä tuulenpaine oletetaan vakioksi. Lisäksi antennin painosta aiheutuu pistevoima P.



Kuva 1. Kuva optimoitavasta mastosta

Tehtävän suunnittelumuuttujia ovat harustasojen lukumäärä r, maakiinnikkeiden etäisyydet x, harustasojen kiinnityskorkeudet y_i, haruksien halkaisijat d_i ja mastoputken

ulkohalkaisija D. Näistä harustasojen kiinnityskorkeudet ja maakiinnikkeiden etäisyydet ovat jatkuvia suunnittelumuuttujia ja harustasojen lukumäärä, mastoputken sekä haruksien halkaisijat ovat diskreettejä. Suunnittelumuuttujien lukumäärä on $2r+3$.

Optimointitehtävän kohdefunktioina ovat materiaalikustannukset c, suurin vaakasiirtymä max δ, alin ominaiskulmataajuus ω_1 ja maston vääntöjäykkyys k_v. Materiaalikustannuksissa huomioidaan harusvaijereiden, mastoputken ja harustasokiinnikkeiden kustannukset. Alin ominaiskulmataajuus saadaan taivutusvärähtelyistä. Vääntöjäykkyys kuvaa antennin tuulikuormituksesta ja epäkeskeisestä kiinnityksestä mastoon syntyvän vääntömomentin ja siitä maston yläpäähän aiheutuvan kiertymän välistä suhdetta. Ominaiskulmataajuutta ja vääntöjäykkyyttä maksimoidaan, kun muita kriteerejä minimoidaan.

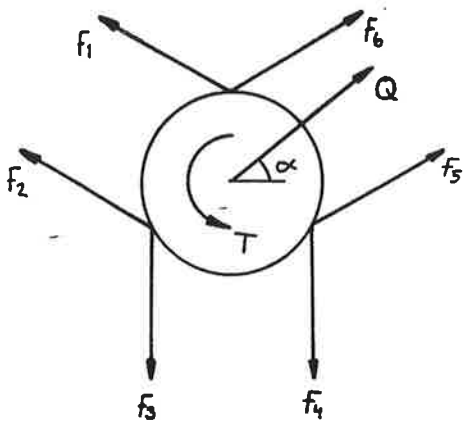
Tehtävän rajoitusehdoissa tutkitaan mastoputkeen taivutuksesta syntyviä normaalijännityksiä σ_p, haruksien normaalijännityksiä σ_{hi}, nurjahduskuormituskerrointa λ sekä harustasojen keskinäistä etäisyyttä ja järjestystä $y_{i+1} - y_i$. Optimointitehtävä on siis

$$\begin{aligned} \min \; & \left[\; c \quad \max \, \delta \quad -\omega_1 \quad -k_v \; \right]^T \\ & \max \, \sigma_p \le \sigma_{psall} \\ & \sigma_{hi} \le \sigma_{hsall} \qquad i = 1, 2, \ldots, r \\ & \lambda \ge \lambda_{sall} \\ & y_{i+1} - y_i \ge y_0 \\ & r \in \{ 1, 2, 3, 4 \} \\ & d_i \in D_1 \\ & D \in D_2 \\ & x_L \le x \le x_U \end{aligned} \tag{1}$$

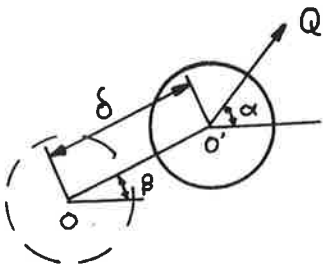
Tehtävän vakioarvot ovat: maston pituus L = 30 m, tuulenpaine p = 750 Pa, antennin tuulikuorma F = 300 N, antennin paino P = 500 N, mastoputken kimmomoduuli E_p = 210 GPa, haruksen kimmomoduuli E_h = 105 GPa, putken ja haruksien tiheys ρ = 7800 kg/m^3, Poissonin vakio ν = 0,3, mastoputken suurin sallittu normaalijännitys $\sigma_{p\,sall}$ = 140 MPa, harusvaijerien suurin sallittu jännitys $\sigma_{h\,sall}$ = 210 MPa, pienin sallittu nurjahduskuormituskerroin λ_{sall} = 3,5, vaakaetäisyyden alaraja x_L = 5 m, vaakaetäisyyden yläraja x_U = 25 m, tuentakorkeuden alaraja y_L = 3 m, tuentakorkeuden yläraja y_U = 27 m, harustasokiinnikkeen hinta h_k = 500 mk, mastoputken kilohinta h_p = 10 mk/kg ja

harusvaijerin kilohinta h_h = 32 mk/kg. Harushalkaisijoille on diskreetti joukko D_1 = {2 4 6 8 10} mm ja mastoputket valitaan joukosta D_2 = {(26,7 2,6) (33,7 2,6) (42,4 3,2) (48,3 3,2) (60,3 3,2) (76,1 4) (88,9 4) (101,6 4) (108 4) (114,3 4) (127 4) (139,7 5)} mm [7]. Jälkimmäisen joukon pisteissä ensimmäinen luku tarkoittaa putken ulkohalkaisijaa ja toinen seinämän paksuutta. Seinämän paksuuden ja ulkohalkaisijan välinen yhteys hoidetaan interpoloimalla splinikäyrällä sisä- ja ulkohalkaisijan välistä suhdetta.

Masto analysoidaan tasomallia käyttäen. Tasomallinnusta voidaan käyttää, sillä mastoputki on pyörähdyssymmetrinen ja lineaarista laskentaa käytettäessä harustason jäykkyys ei riipu kuormituksen suunnasta. Mastoputken mallinnukseen käytetään palkkielementtejä ja harustaso korvataan yhdellä jousielementillä, jolla on harustason jäykkyyttä vastaava jousivakio. Jokaisesta maakiinnikkeestä on harusvaijeri kahteen samalla korkeudella olevaan putkikiinnikkeeseen. Maakiinnikkeestä samalle harustasolle lähtevät harukset voidaan suurta virhettä tekemättä olettaa samansuuntaisiksi. Jos maston siirtymät ja kiertymät oletetaan pieniksi, putkikiinnitys voidaan mallintaa kuvan 2. mukaisella kiekolla. Kiekko on tuettu jousilla, joiden jousivoimia on merkitty symboleilla F_1, ... , F_6, ja sitä kuormitetaan ulkoisella voimalla Q sekä vääntömomentilla T.



Kuva 2. Kiekko, jolla mallinnetaan harustason kiinnitystä



Kuva 3. Kiekon siirtyminen voiman vaikutuksesta

Kiekon tuennassa käytettävät jousivakiot saadaan haruksen vaakasuuntaisesta jäykkyydestä

$$k_{xi} = \frac{E_h A_i}{l_i}\cos^2(\alpha_i), \tag{2}$$

missä A_i tarkoittaa harusvaijerin poikkipinta-alaa, l_i haruksen pituutta sekä α_i haruksen ja maan välistä kulmaa.

Kuvan 2. kiekon tasapainoehdot ovat

$$\begin{aligned} \rightarrow\ & \left[-(F_1+F_2) + (F_5+F_6)\right]\cos\left(\frac{\pi}{6}\right) + Q\cos(\alpha) = 0 \\ \uparrow\ & -(F_3+F_4) + (F_1+F_2+F_5+F_6)\sin\left(\frac{\pi}{6}\right) + Q\sin(\alpha) = 0 \\ & (F_1-F_2+F_3-F_4+F_5-F_6)\cos\left(\frac{\pi}{6}\right)r_0 + T = 0, \end{aligned} \tag{3}$$

missä r_0 on kiinnityksen säde. Jos kiekko siirtyy kuvan 3. mukaisesti akselinsa ympäri kiertymättä pisteestä O pisteeseen O', niin jousivoimien lausekkeiksi saadaan

$$\begin{aligned} F_1 &= F_2 = N + k_{xi}\delta\cos\left(\frac{\pi}{6}+\beta\right) \\ F_3 &= F_4 = N + k_{xi}\delta\sin(\beta) \\ F_5 &= F_6 = N - k_{xi}\delta\cos\left(\frac{\pi}{6}-\beta\right), \end{aligned} \tag{4}$$

missä N on esikiristysvoima, β siirtymän suuntakulma ja δ siirtymän arvo. Kun nämä sijoitetaan tasapainoehtoihin (3), tulokseksi saadaan

$$\begin{aligned} Q &= 3k_{xi}\delta \\ T &= 0 \\ \alpha &= \beta. \end{aligned} \tag{5}$$

Ulkoinen voima ei siis aiheuta momenttia, ja kiekon siirtymä on voiman suuntainen. Haruskiinnityksen ekvivalenttiseksi jousivakioksi saadaan $k_{eq} = 3k_{xi}$.

Vääntötapauksen ekvivalenttinen jousivakio saadaan, kun kiekkoa kierretään kulman θ verran aiheuttamatta sille siirtymää. Jos oletetaan, että kulma θ on pieni, niin jousivoimien lausekkeiksi saadaan

$$F_1 = F_3 = F_5 = N - k_{xi}\frac{\sqrt{3}}{2}r_0\theta$$
$$F_2 = F_4 = F_6 = N + k_{xi}\frac{\sqrt{3}}{2}r_0\theta. \quad (6)$$

Nämä sijoitetaan tasapainoehtoihin (3), jolloin saadaan

$$Q = 0$$
$$T = \frac{9}{2}k_{xi}r_0^2\theta. \quad (7)$$

Jos kiinnityksen säteeksi r_0 valitaan mastoputken ulkohalkaisija D, harustason ekvivalenttiseksi jousivakioksi väännössä saadaan $k_{eqv} = 4{,}5k_{xi}D^2$. Väännössä maston alapää on jäykästi tuettu eli maston alapäähän väännöstä ei synny kiertymiä.

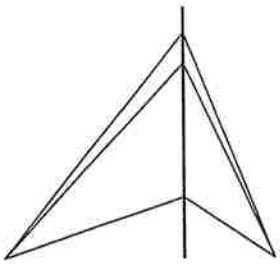
Nurjahduskuormituskerroin lasketaan lineaarisen ominaisarvotehtävän avulla. Maston normaalivoimakuormitus syntyy haruksien esikiristysvoimista ja mastoputken sekä antennin painosta. Haruksien esikiristysvoimat lasketaan siten, että haruksien normaalivoimat ovat koko ajan positiivisia, jolloin harus ei pääse löystymään. Tämä on välttämätöntä, jotta harusvaijeri toimisi jousen tavoin. Haruksien omaa painoa ei huomioida.

Rakenteen alin ominaiskulmataajuus lasketaan taivutusvärähtelyistä. Mastoputken massamatriisina käytetään konsistenttia massamatriisia. Harukset ja kiinnikkeet huomioidaan keskittämällä puolet harustason harusvaijereiden massasta pistemassaksi harustason kiinnityskohtaan. Myös antenni huomioidaan pistemassana. Alin vääntöominaiskulmataajuus todettiin olevan selvästi alinta taivutusominaiskulmataajuutta suurempi, jolloin sitä ei tarvinnut tässä yhteydessä huomioida.

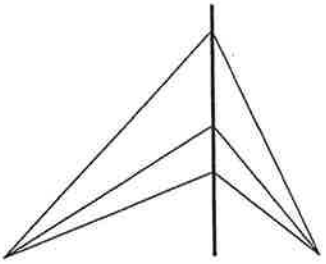
Tehtävästä laskettiin yksittäisten kriteerien optimipisteet sekä kaksi kompromissiratkaisua. Edullisena kompromissiratkaisuna lasketaan materiaalikustannusten minimi, kun rakenteen maksimitaipuma on enintään 10 cm ja vääntöjäykkyys on vähintään 15 kNm/rad. Vääntö- ja taivutusjäykkänä kompromissiratkaisussa maksimoidaan vääntöjäykkyyttä, kun materiaalikustannukset ovat enintään 15000 mk, maksimitaipuma on enintään 5 cm ja alin ominaiskulmataajuus on vähintään 15,708 rad/s. Ominaiskulmataajuuden raja vastaa ominaistaajuutta 2,5 Hz. Kriteerien ja suunnittelumuuttujien arvot on esitetty taulukossa 1 ja saadut mastorakenteet kuvissa 4 ... 9.

Taulukko 1. Kriteerien ja suunnittelumuuttujien arvot lasketuissa Pareto-optimeissa. Piste 1 on materiaalikustannusten minimi, piste 2 maksimitaipuman minimi, piste 3 vääntöjäykkyyden maksimi, piste 4 alimman ominaiskulmataajuuden maksimi, piste 5 on edullinen ja piste 6 jäykkä kompromissiratkaisu.

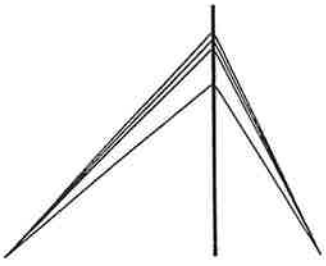
1	c = 4410 mk max δ = 13,2 cm k_v = 5,42 kNm/rad ω_1 = 6,07 rad/s	r = 3 x = 21,40 m D = 88,9 mm	d_1 = 2 mm d_2 = 2 mm d_3 = 2 mm	y_1 = 7,28 m y_2 = 23,36 m y_3 = 27,00 m
2	c = 17410 mk max δ = 0,88 cm k_v = 38,33 kNm/rad ω_1 = 16,28 rad/s	r = 3 x = 25,00 m D = 139,7 mm	d_1 = 10 mm d_2 = 10 mm d_3 = 10 mm	y_1 = 9,97 m y_2 = 15,58 m y_3 = 27,00 m
3	c = 23500 mk max δ = 13,7 cm k_v = 49,85 kNm/rad ω_1 = 7,78 rad/s	r = 4 x = 25,00 m D = 139,7 mm	d_1 = 10 mm d_2 = 10 mm d_3 = 10 mm d_4 = 10 mm	y_1 = 20,74 m y_2 = 25,00 m y_3 = 26,00 m y_4 = 27,00 m
4	c = 20380 mk max δ = 1,5 cm k_v = 44,61 kNm/rad ω_1 = 20,95 rad/s	r = 4 x = 25,00 m D = 139,7 mm	d_1 = 10 mm d_2 = 10 mm d_3 = 10 mm d_4 = 10 mm	y_1 = 10,66 m y_2 = 22,24 m y_3 = 23,24 m y_4 = 27,00 m
5	c = 5400 mk max δ = 10,0 cm k_v = 16,47 kNm/rad ω_1 = 6,25 rad/s	r = 2 x = 11,47 m D = 127 mm	d_1 = 4 mm d_2 = 4 mm	y_1 = 7,65 m y_2 = 25,29 m
6	c = 15000 mk max δ = 3,0 cm k_v = 38,44 kNm/rad ω_1 = 15,708 rad/s	r = 4 x = 24,01 m D = 139,7 mm	d_1 = 4 mm d_2 = 4 mm d_3 = 8 mm d_4 = 10 mm	y_1 = 13,04 m y_2 = 22,64 m y_3 = 26,00 m y_4 = 27,00 m



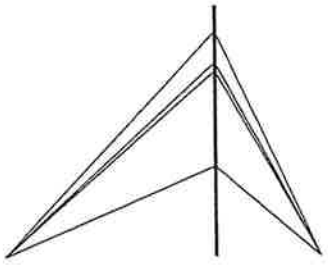
Kuva 4. Materiaalikustannusten minimi



Kuva 5. Maksimisiirtymän minimi

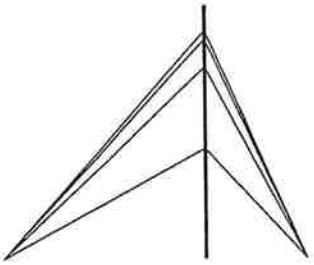


Kuva 6. Vääntöjäykkyyden maksimi



Kuva 7. Alimman ominaiskulmataajuuden maksimi

Kuva 8. Edullinen kompromissiratkaisu



Kuva 9. Jäykkä kompromissiratkaisu

Nämä Pareto-optimit on saatu branch and bound algoritmilla. Samat pisteet laskettiin käyttäen myös geneettistä algoritmia. Koska geneettinen algoritmi on stokastinen algoritmi, pisteet laskettiin kahteen kertaan. Populaation kokona käytettiin 100 yksilöä ja sukupolvien lukumääränä 800. Materiaalikustannusten minimiksi saatiin 4800 mk ja 5160 mk (4410 mk), maksimitaipuman minimiksi 0,88 cm ja 0,89 cm (0,88 cm), vääntöjäykkyyden maksimiksi 49,83 kNm/rad ja 49,84 kNm/rad (49,85 kNm/rad), alimmaksi

ominaiskulmataajuudeksi molemmilla kerroilla 20,95 rad/s (20,95 rad/s), halvemman kompromissiratkaisun materiaalikustannuksiksi 5780 mk ja 6330 mk (5400 mk) sekä jäykemmän kompromissiratkaisun vääntöjäykkyydeksi 38,27 kNm/rad ja 38,29 kNm/rad (38,44 kNm/rad). Suluissa olevat vertailuarvot on saatu branch and bound algoritmilla. Geneettinen algoritmi oli näissä pisteissä jonkin verran hitaampi kuin branch and bound algoritmi.

Geneettisellä algoritmilla oli vaikeuksia löytää materiaalikustannusten minimi ja halvempi kompromissiratkaisu. Muissa pisteissä saadut kohdefunktion arvot ovat varsin tarkkoja. Ongelmapisteissä algoritmilla oli harustasojen lukumäärän valinnassa ongelmia. Vaikuttaisi siltä, että harustasojen lukumäärä konvergoisi hyvin varhaisessa vaiheessa ja sen jälkeen lukumäärän vaihtaminen on hyvin epätodennäköistä. Käytetty geneettisen algoritmin koodi ei toimi parhaalla mahdollisella tavalla, jos tehtävässä on runsaasti aktiivisia rajoitusehtoja. Kun materiaalikustannuksia minimoidaan, lähes kaikki jännitysrajoitusehdot sekä nurjahdusrajoitusehto ovat aktiivisia. Tämä hankaloittaa materiaalikustannusten ja halvemman kompromissiratkaisun konvergointia.

Maksimitaipuman minimissä geneettinen algoritmi valitsi neljä harustasoa, kun branch and bound suosittelee kolmea. Molemmissa ratkaisuissa kaikkien kohdefunktioiden arvot olivat kuitenkin hyvin lähellä toisiaan. Vääntöjäykkyyden maksimoinnissa kolme maston huipussa olevaa harusta aiheuttivat geneettiselle algoritmille ongelmia. Nämä harukset ovat metrin välein toisistaan, jolloin pienikin muutos johonkin harustasokorkeuteen tekee rakenteesta epäkäyvän. Ominaiskulmataajuuden maksimointi onnistui sitä vastoin erinomaisesti, sillä suunnittelumuuttujien arvot olivat laskentatarkkuuden puitteissa samoja kuin branch and bound algoritmilla saadut. Jäykemmässä kompromissiratkaisussa haruksien halkaisijat ei ollet aivan samoja, mutta saatu kohdefunktion arvo oli silti todella hyvä.

YHTEENVETO

Artikkelissa on esitelty kahden eri optimointialgoritmin käyttöä harustetun maston optimoinnissa. Laskuesimerkkinä on käytetty matalahkoa putkirunkoista mastoa, joka

laskettiin monitavoitteisena optimointitehtävänä. Tehtävästä laskettiin kuusi erilaista Pareto-optimia, joissa harustasojen lukumäärä vaihtelee kahdesta neljään.

Molemmilla algoritmeilla on saatu samankaltaisia tuloksia. Geneettisellä algoritmilla oli vaikeuksia löytää materiaalikustannusten minimiä ja edullista kompromissiratkaisua. Näissä pisteissä harustasojen lukumäärän valinta ja usean aktiivisen rajoitusehdon samanaikainen käsittely näyttivät tuottavan hankaluuksia käytetylle ohjelmaversiolle. Muiden Pareto-optimien laskenta onnistui varsin hyvin.

LÄHTEET

1. Turkkila, T. *Kantavien rakenteiden diskreetti optimointi.* Tampere 1997. Lisensiaatintyö, Tampereen teknillinen korkeakoulu, konetekniikan osasto.

2. Syvänen, L. A. *An Equivalent Beam for calculating Slender Trusses.* 5th Finnish Mechanical Days, 26-27.5.1994. Jyväskylän Yliopisto. S. 111-118.

3. Koski, J. *Multicriterion Structural Optimization.* In: Adeli, H. Advances in Design Optimization. London 1994. Chapman & Hall. S. 194-224

4. Nehmhauser, G.L. and Wolsey, L.A. *Integer and Combinatorial Optimization.* New York 1988, John Wiley & Sons. 763 s.

5. Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading 1989. Addison-Wesley Publishing Company Inc. 412 s.

6. Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs.* 2. p. Berlin 1994. Springer-Verlag. 340 s.

7. Rautaruukki. *Suunnittelijan opas.* Raahe 1989, Rautaruukki Oy. 192 s.

THE FULLY STRESSED DESIGN AND THE MINIMUM OF VOLUME

PERTTI HOLOPAINEN
Department of Mechanical Engineering
Tampere University of Technology
P.O. Box 589, SF-33101 TAMPERE, FINLAND

ABSTRACT

It is known, on principle, that the fully stressed design of statically indeterminate trusses is not possible because of compatibility.
The fully stressed design method for statically indeterminate trusses using pre-stressing is proposed. The fully stressed design procedure to obtain the absolute minimum of volume, deleting suitable uneconomical members of statically indeterminate trusses, is presented for one loading. The combination of two different loadings is also considered. The design variables are the only reduntant forces.

Introductory example

Consider the three-bar system as shown in the Fig. 1. It is a symmetric simple plane truss with one degree of redundancy.

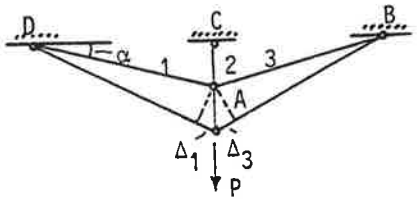


Fig. 1.

Member force $S_2 = X$ is chosen here as a reduntant force and it is assumed that member 2 is cutt at support C. When members 1 and 3 cross-sectional areas $A_1 = S_1 / \sigma_m$ and $A_3 = S_3 / \sigma_m$, when $|\sigma_m|$ is a maximum allowable value, members 1 and 3 are fully stressed. Therefore elongations $\Delta_i = (S_i l_i / EA_i)$, $(i = 1,3)$ are known in advance. It is assumed here, that Δ_i are small. In this case point A moves down distance $\Delta_A = \Delta_1 / \sin\alpha$. But member 2 is shorter, $l_2 = l_1 \sin\alpha$, and its fully stressed elongation $\Delta_2 = \Delta_1 \sin\alpha$. This means, that between support C and member 2 there remains a gap

$$\delta_2 = \frac{\Delta_1}{\sin\alpha} - \Delta_1 \sin\alpha = \Delta_1 \frac{\cos^2\alpha}{\sin\alpha}. \qquad (1)$$

Member 2 can be prepared a quantity δ_2 longer, than the length l_2 calculated from the initial geometry. Therefore the compatibility becomes true in the fully stressed state. But this means, that the whole structure is stressed, when the loading is absent. Member 2 is 'too long' setting up the structure and it must be forced between points A and C. So the structure becomes prestressed. Member 2 can be lengthened δ_2 in the structure after the mounting as is shown in principle in Fig. 2.

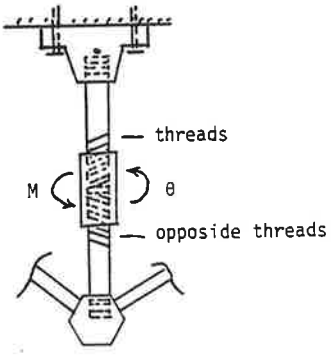


Fig. 2.

The elongation

$$\delta_2 = 2\theta h \tag{2}$$

and the pre-stressing force

$$Q = \frac{M}{2r\tan(\alpha+\phi)} \tag{3}$$

after /1/ for example. In Eq. (2) Θ=rotation angle and h=pitch. In Eq. (3) M=rotation moment, r =mean radius of the thread, α =helix angle (same at both ends) and the friction angle $\phi = \arctan\mu$ in square threaded tap and $\phi_1 = \arctan(\mu/\cos\beta)$ in V-threaded tap and 2β =ridge angle. Elongation δ_2 may not be too large, so that the pre-stressing force exceeds allowable limitations in the absence of loading. When elongation δ_2 is made, the fully loaded structure is also fully stressed. Here it is not necessary to solve statically indeterminate structure in conventional sense. The equilibrium conditions must always be true, this means, all the memberforces can be calculated as a function of force X and loading P. Varying X only all the possible variations of all member forces S_1, S_2, S_3 can be presented.

When the angle α (Fig. 1) is varied, it can be obtained

$$\lim_{\alpha\to\pi/2}\delta_2 = 0 \tag{4}$$

Thus, when $\alpha = \pi/2$ there is no gap at support C. But in this case members 1, 2 and 3 joins with each other. As a result become a statically determinate (and here also movable) structure.

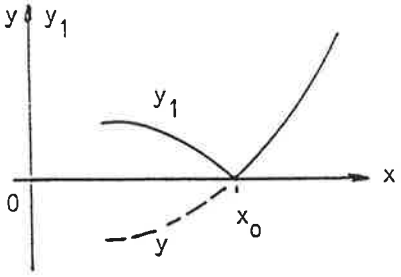
If support C lies higher, so that $l_2 = l_1/\sin\alpha$, the fully stressed elongation of member 2 $\Delta_2 = \Delta_1/\sin\alpha$. Thus there is no gap at support C in the fully stressed state.

Is the volume (or own weight) of the structure as shown in Fig. 1. in the absolute minimum when the structure is in the fully stressed state loaded by P? The whole volume of structure

$$V = \frac{1}{|\sigma_m|}[|S_1|l_1 + |S_2|l_2 + |S_3|l_3] \tag{5}$$

where the posible buckling of some members is not taken into account. When force X is varied, the V-X-line is brocken. The angle in theV-X-line is produced, when some member force changes its sign as X is varied. The minimum of V can occur only in some angle point (Fig. 4). The minimum of V is reached by that value of X, which produces some member force equal to zero.

Derivation of absolute value function. Let y = f(x) continuously differentiable $\forall x \in R$. Let $y_1 = |f(x)| \ \forall x \in R$.



Kuva 3.

Function y_1 has derivative from the point x_0 (Fig. 3.) on the left and on the right and the derivatives before have different values. But in point x_0 the funktion y_1 has no derivative. When it is denoted

$$|f(x)| = \sqrt{f(x)^2} \qquad \text{(positive square root)} \tag{6}$$

the derivatives $y_1^{(k)}$ $(k = 1,2,...)$ can be expressed by one formula in the both sides of x_0. For example

$$y_1' = \frac{f(x)}{\sqrt{f(x)^2}} f'(x) \quad \forall x \in R - A \tag{7}$$

where

$$A = \{x \in R: f(x) = 0\}.$$

So instead of Eq. 5. there is written

$$V = \frac{1}{|\sigma_m|}\left[\sqrt{S_1^2}\, l_1 + \sqrt{S_2^2}\, l_2 + \sqrt{S_3^2}\, l_3\right]. \tag{8}$$

Further

$$\frac{\partial V}{\partial X} = \frac{1}{|\sigma_m|}\left[\frac{S_1 l_1}{\sqrt{S_1^2}}\frac{\partial S_1}{\partial X} + \frac{S_2 l_2}{\sqrt{S_2^2}}\frac{\partial S_2}{\partial X} + \frac{S_3 l_3}{\sqrt{S_3^2}}\frac{\partial S_3}{\partial X}\right] =$$

$$= \frac{1}{|\sigma_m|}\left[\operatorname{sgn}(S_1) l_1 \frac{\partial S_1}{\partial X} + \operatorname{sgn}(S_2) l_2 \frac{\partial S_2}{\partial X} + \operatorname{sgn}(S_3) l_3 \frac{\partial S_3}{\partial X}\right]. \qquad (9)$$

If the member end loads S_i $(i = 1,2,3)$ are linear funcktions of X, the partial derivates $\partial S_i / \partial X$ are constants in Eq. (9). In this case $\partial V / \partial X$ is constant between the zero points of $S_i(x)$ and is discontinuous in the zero points of $S_i(x)$.

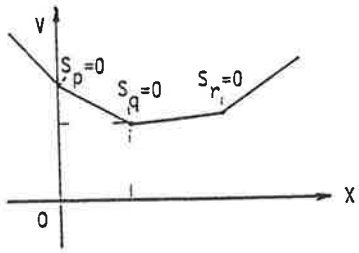


Fig. 4.

Numerical example.

Determine the absolute V_{min} of the three-bar truss as shown in Fig. 5a.

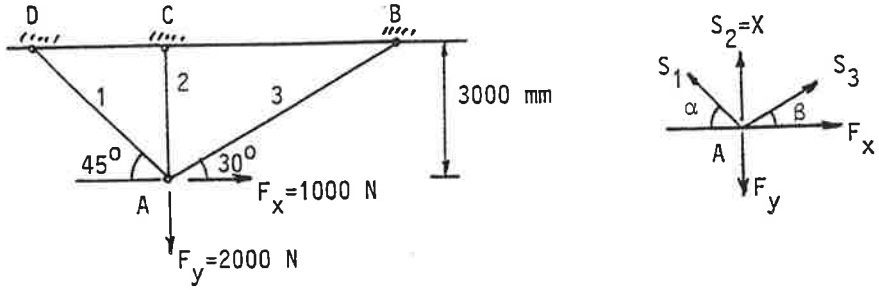


Fig. 5.

- V_{min} examining the derivative $\partial V / \partial X$.

It is assumed that all members are of the same material and the maximum allowed value $|\sigma_m| = 140 N / mm^2$. The truss under consideration being reduntant to the first degree. The member force $S_2 = X$ is chosen here as a reduntant force. From equilibrium conditions of joint A (Fig. 5b) it can be obtained

$$S_1 = \frac{\sin\beta}{D}F_x + \frac{\cos\beta}{D}F_y - \frac{\cos\beta}{D}X$$

$$S_3 = -\frac{\sin\alpha}{D}F_x + \frac{\cos\alpha}{D}F_y - \frac{\cos\alpha}{D}X \qquad (10)$$

$$(S_2 = X)$$

$$D = \cos\alpha\sin\beta + \sin\alpha\cos\beta \qquad (11)$$

When S_1 changes its sign

$$S_1 = 0 \;\&\; (10)_1 \Rightarrow X = 2577{,}35N\ (= S_2) \;\&\; S_3 = -1154{,}7N$$

$$V = \frac{1}{|\sigma_m|}\left[|S_2|l_2 + |S_3|l_3\right] = 104716{,}07\,mm^3. \qquad (12)$$

When S_2 changes its sign

$$S_2 = X = 0 \;\&\; (10) \Rightarrow S_1 = 2310{,}80N \;\&\; S_3 = 732{,}07N.$$

$$V = \frac{1}{|\sigma_m|}\left[|S_1|l_1 + |S_3|l_3\right] = 101402\,mm^3. \qquad (13)$$

When S_3 changes its sign:

$$S_3 = 0 \;\&\; (10)_2 \Rightarrow X = 1000N\ (= S_2) \;\&\; S_1 = 1414{,}2N$$

$$V = \frac{1}{|\sigma_m|}\left[|S_1|l_1 + |S_2|l_2\right] = 64285{,}3\,mm^3. \qquad (14)$$

$(12)\&(13)\&(14) \Rightarrow V_{\min} = 64259\,mm^3$.

The structure coresponding to $V_{\min}$ is presented in Fig. 6.

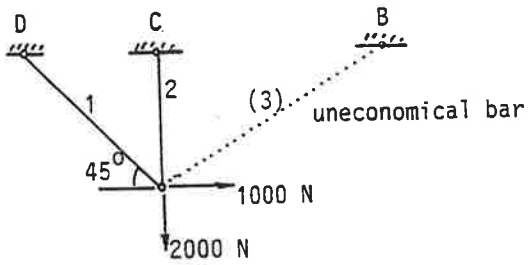


Fig. 6.

$-V_{min}$ calculating V(X) by different values of X.

S_1	$S_2 = X$	S_3	V	ΔV	S_1	$S_2 = X$	S_3	V	ΔV
2759N	-500N	1098N	141387mm^3		1683	700	220	75420	-3712
2669	-400	1025	133390	-7997	1594	800	146	71709	-3711
2580	-300	951	125393	-7997	1504	900	73	67997	-3712
2490	-200	878	117396	-7997	1414	1000	0	64285,7	-3711
2400	-100	805	109398	-7998	1325	1100	-73	66849	+2563
2311	0	732	101401	-7997	1235	1200	-146	69412	2563
2221	100	659	97690	-3711	1145	1300	-220	71975	2563
2131	200	586	93978	-3712	1056	1400	-293	74538	2563
2042	300	512	90267	-3711	966	1500	-366	77102	2563
1951	400	439	86555	-3712	876	1600	-439	79665	2563
1863	500	366	82843	-3712	787	1700	-512	82228	2563
1773	600	293	79132	-3711	697	1800	-586	84791	2563
607	1900	-659	87354	2563	-200	2800	-1318	122522	7997
518	2000	-732	89918	2564	-289	2900	-1391	130519	7997
428	2100	-805	92481	2563	-378	3000	-1464	138517	7998
338	2200	-878	95044	2563	-469	3100	-1537	146514	7997
249	2300	-952	97607	2563	-558	3200	-1611	154511	7997
159	2400	-1025	100170	2563	-648	3300	-1684	162508	7997
69,4	2500	-1098	102733,5	2563,5					
-20,3	2600	-1171	106527,5	(3794)					
-110	2700	-1244	114525	7997,5					

It can be seen, that V_{min}= 64285.7 mm^3 , which is nearly the same as that (64259 mm^3) obtained examining the derivative $\partial V / \partial X$ before.

If the loading as shown in Fig. 5 is changed so that F_x is to the left, then $F_x = -1000\,N$ *and* $F_y = 2000\,N$ as before, the following calculations for V_{min} can be obtained.

109N	1300N	1244N	84525mm^3	-3712
20,3	1400	1171	80813	
-69,4	1500	1098	81305	
-159	1600	1025	83027	1722

Using linear interpolation V_{min}=80835 mm^3 can be obtained. The structure corresponding to V_{min} is presented in Fig. 7.

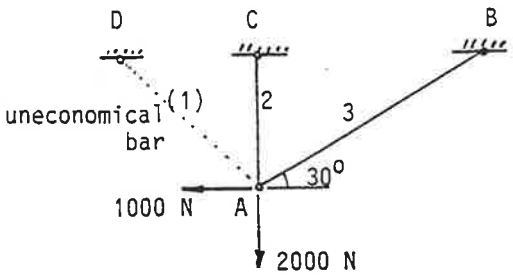


Fig. 7.

If as a loading are the same as both previous cases, either $F_x = 1000N$, $F_y = 2000N$ or $F_x = -1000N, F_y = 2000N$ the final structure can be combined from the trusses of Fig.s 6 and 7. Therefore members 1, 2 and 3 remains so as the statically indeterminancy. To obtain the absolute V_{min} by fixed truss geometry the prestressing and the fully stressed design must be used as described in before mentioned introductory example .

General consideration

It is assumed that the truss is reduntant to the n degree. The truss, plane or space truss, contains m members, m>n. The member forces $S_1 \ldots S_m$ can be determined as a function of loads and redundant forces $X_1 \ldots X_n$

$$\begin{aligned} S_1 &= S_{0_1} + S_{11}X_1 + S_{12}X_2 + \ \ldots \ S_{1n}X_n \\ S_2 &= S_{0_2} + S_{21}X_1 + S_{22}X_2 + \ \ldots \ S_{2n}X_n \\ &\vdots \\ S_m &= S_{0_m} + S_{m1}X_1 + S_{m2}X_2 + \ \ldots \ S_{mn}X_n \end{aligned} \tag{15}$$

The Equations (15) can be derived either by the method of joints /1/ or by the unit force method /2/.The reduntants can be both member forces and support reactions. The volume of members material

$$V = \frac{1}{|\sigma_m|}\left[\sqrt{S_1^2}l_1 + \sqrt{S_2^2}l_2 + \ \ldots \ \sqrt{S_m^2}l_m\right] \tag{16}$$

where it is assumed that all members are made of the same material and the buckling is not taken into account. The partial derivatives

$$\begin{aligned}
\frac{\partial V}{\partial X_1} &= \frac{1}{|\sigma_m|}\left[\operatorname{sgn}(S_1)l_1S_{11} + \operatorname{sgn}(S_2)l_2S_{21} + \ \ldots\ \operatorname{sgn}(S_m)l_mS_{m1}\right] \\
\frac{\partial V}{\partial X_2} &= \frac{1}{|\sigma_m|}\left[\operatorname{sgn}(S_1)l_1S_{12} + \operatorname{sgn}(S_2)l_2S_{22} + \ \ldots\ \operatorname{sgn}(S_m)l_mS_{m2}\right] \\
&\vdots \\
\frac{\partial V}{\partial X_n} &= \frac{1}{|\sigma_m|}\left[\operatorname{sgn}(S_1)l_1S_{1n} + \operatorname{sgn}(S_2)l_2S_{2n} + \ \ldots\ \operatorname{sgn}(S_m)l_mS_{mn}\right]
\end{aligned} \tag{17}$$

How find the $V_{\min}$ and uneconomical members? (Cf. numerical example before)

It must chose an abitrary combination (ν) of n member forces $S^{(\nu)}$ and set them equal to zero. Thus the gradV makes a jump and on the surface V is an edge. So

$$\{0\} = \{S_0^{(\nu)}\} + [S^{(\nu)}]\{X\} \ \&\ (15) \Rightarrow \{X\} = [S^{(\nu)}]^{-1}\{S_0^{(\nu)}\}. \tag{18}$$

Then calculates the remaining memberforces

$$\{S^{(n-\nu)}\} = \{S_o^{(n-\nu)}\} + [S^{(n-\nu)}]\{X\}. \tag{19}$$

Then calculates $(0^{(\nu)}, |S^{(n-\nu)}|)$.

This procedure must be repeated by all possible combinations of $\{S^{(\nu)}\}$. Then notes the $V_{\min}$ and the correspondent combination $\{S^{(\nu)}\} = \{0\}$ Then delete the correspondent members from the truss. The remaining system is a statically determinate truss containing member combination $(m-\nu)$. The deleted members can be replaced by the inexpensive members with sufficient slack in fixing.

Conclusion.

The material volume optimization and the fully stressed design of the truss (buckling exluded) with one case of loading results the statically determinate truss.

General references

/1/. F.P.Beer and E.R.Johnston, Vector Mechanics for Engineers, Statics. McGraw-Hill Inc.

/2/. Arvo Ylinen, Kimmo- ja lujuusoppi I, WSOY.

/3/. J.S.Przemieniecki, Theory of Matrix Structural Analysis. McGraw-Hill Inc. (1968)

/4/. U.Kirsch, Structural Optimization. Springer-Verlag (1993).

/5/. F.Ayers, Jr, Matrices. Schaum Publishing Co. (1962).

/6/. Brochures of British Steel Corporation. Tubes Division.

POSTBUCKLING ANALYSIS OF AN ELASTIC STRUT BY TWO FORMULATIONS

E.-M. SALONEN

Laboratory of Theoretical and Applied Mechanics

Helsinki University of Technology

P.O.Box 1000, FIN-02015 HUT, Finland

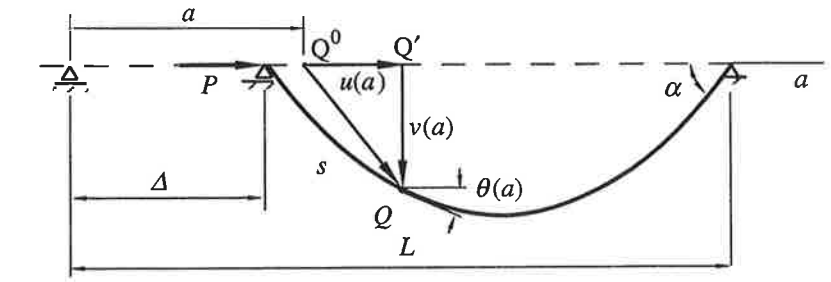
ABSTRACT

Large displacement analysis of a uniform originally straight inextensible elastic strut (the classical elastica problem) is performed employing both the conventional Lagrangian description (normally used in solid mechanics) and the Eulerian description (normally used in fluid mechanics). The article is mainly of pedagogic nature: as the displacements are really large, the differencies between the two formulations can be clearly seen in a rather simple setting. This example can thus be used to illuminate the basic features of the two descriptions. The governing equations are presented. The properties of the two formulations are discussed. The equations are transformed in non-dimensional forms and solved by presenting the equilibrium equations first weakly and by then applying the Galerkin method with trial functions consisting of sines. Iterative solution procedures are necessary due to the nonlinearities. The final discrete system equations are in spite of some linearisations still nonlinear. All the necessary numerical calculations are executed by the Mathematica program.

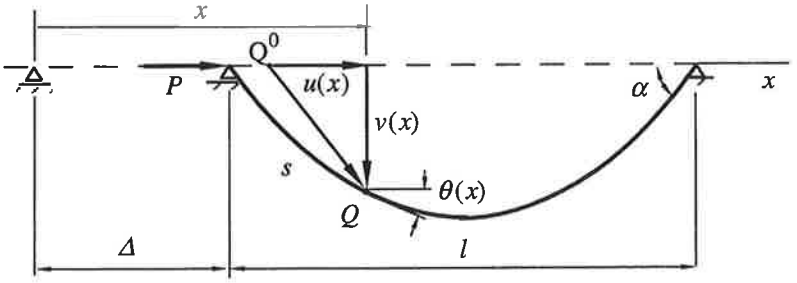
INTRODUCTION

In mechanics, the two main ways to describe the motion of a continuum are the Lagrangian description and the Eulerian description, [1]. The former is usually employed in solid mechanics and the latter in fluid mechanics. To clearly understand the properties of these two descriptions is of utmost importance for the student to proceed in the assimilation of mechanics. We consider in this article the large displacement analysis of a uniform originally straight inextensible elastic strut (the classical elastica problem). Both the Lagrangian description and the Eulerian description are employed. This can be considered

as a rather detailed extension of the subject treated in Reference [2] now for arbitrary large displacements. As the displacements are really large, the differencies between the two formulations become clearly discernible and the presentation may thus be of some pedagogical value as a suitable demonstration example on the theme.



(a)



(b)

FIGURE 1 (a) Lagrangian description. (b) Eulerian description.

Some of the notations are shown in Figure 1. The original — and as the strut is assumed inextensible — the final length of the strut axis is L. The independent variables in the Lagrangian and the Eulerian descriptions are denoted by a and x, respectively. (Normally the first space coordinate is denoted by the symbol x in both the formulations but as we have here to consider the two formulations simultaneously, for clarity the need for different symbols arises.) The horizontal and lateral displacement components are u and v, respectively. The horizontal displacement Δ of the left hand end is considered given and the corresponding axial load P is originally unknown. Due to the inextensibility constraint u can be expressed in terms of v and the final problem is to determine $v(a)$ and P or $v(x)$ and P in the Lagrangian or in the Eulerian formulation, respectively. The constitutive relation is

$$M = EI\kappa, \tag{1}$$

where M is the bending moment, EI the bending stiffness and $\kappa = 1/\rho$ the curvature.

The exact solution of the elastica problem is described for instance in [3]. The classical formulation proceeds considering the slope angle θ as the dependent variable and the curved arclength s as the independent variable (Figure 1). This leads to a rather specialised formulation and to the use of of elliptic integrals. Here we want to consider the deflection v as the dependent variable and the "straight" coordinate a or x as the independent variable so that the student can make direct comparison with the conventional formulation of small displacement beam bending.

LAGRANGIAN DESCRIPTION

Governing equations

Let us consider Figure 1 (a) for the Lagrangian description. It is important to realize that in the Lagrangian description so to speak "physics takes place along the curved buckled strut but mathematics takes place along the straight a-axis"; the domain of the mathematical definition of the problem is the initial domain $0 \le a \le L$. A student accustomed to figures used in deriving the small displacement beam bending theory could easily write down the formulas

$$\tan\theta = \frac{\mathrm{d}v}{\mathrm{d}a} \quad \text{(wrong)} \tag{2}$$

and

$$\kappa = \frac{-\dfrac{\mathrm{d}^2 v}{\mathrm{d}a^2}}{[1+(\dfrac{\mathrm{d}v}{\mathrm{d}a})^2]^{3/2}} \quad \text{(wrong)} \tag{3}$$

The former seems to be true by looking at Figure 1 (a) and the latter apparently corrects the small displacement curvature approximation $-\mathrm{d}^2 v/\mathrm{d}a^2$ into the large displacement range with the help of any mathematics handbook. The pitfall lies in the fact that in the notation $v(a)$ in Figure 1 (a), the measure a does not refer to the perpendicular point Q′ to Q but to point Q^0 at the initial straight position which has displaced to point Q in the current deformed position.

We now derive the correct versions of formulas (2) and (3) and some additional results following roughly the presentation of Reference [2]. Figure 2 shows an infinitesimal strut

element in the initial and in the current position. The arclength element ds is equal to the original length element da due to the inextensibility assumtion. From the figure,

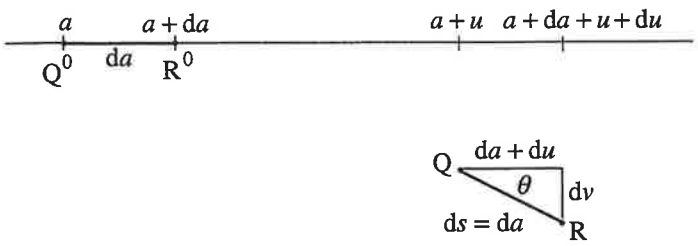


FIGURE 2 Initial and current position of an infinitesimal strut element.

$$\tan\theta = \frac{dv}{da+du} = \frac{\frac{dv}{da}}{1+\frac{du}{da}} \tag{4}$$

and

$$\sin\theta = \frac{dv}{da}. \tag{5}$$

The curvature is the rate of change of the slope angle with respect to the arc length (here with the minus sign):

$$\kappa = -\frac{d\theta}{ds} = -\frac{d\theta}{da} = -\frac{d[\sin^{-1}\frac{dv}{da}]}{da} = \frac{-\frac{d^2v}{da^2}}{[1-(\frac{dv}{da})^2]^{1/2}}. \tag{6}$$

Further from Figure 2,

$$(da)^2 = (da+du)^2 + (dv)^2 \tag{7}$$

and division by $(da)^2$ gives

$$1 = (1+\frac{du}{da})^2 + (\frac{dv}{da})^2. \tag{8}$$

Solving this inextensibility constraint equation for du/da leads first to

$$1+\frac{du}{da} = \pm[1-(\frac{dv}{da})^2]^{1/2} \tag{9}$$

and finally to

$$\frac{du}{da} = -1 + [1 - (\frac{dv}{da})^2]^{1/2}. \tag{10}$$

The plus sign has been selected because the minus sign would lead in the case $v(a) \equiv 0$ to the unphysical result $du/da = -2$. Using expression (10), we obtain

$$\begin{aligned} u(a) &= u(0) + \int_0^a \frac{du}{da} da = \Delta + \int_0^a \{-1 + [1 - (\frac{dv}{da})^2]^{1/2}\} da \\ &= \Delta - a + \int_0^a [1 - (\frac{dv}{da})^2]^{1/2} da. \end{aligned} \tag{11}$$

At $a = L$, $u = 0$ since the right hand side support is fixed, and we obtain from (11) a global inextensibility constraint equation

$$\Delta = L - \int_0^L [1 - (\frac{dv}{da})^2]^{1/2}] da. \tag{12}$$

We see from Figure 1 (a) that the bending moment at the generic point Q is Pv. Equating this with expression (1) using (6) gives

$$Pv = -EI \frac{\frac{d^2v}{da^2}}{[1 - (\frac{dv}{da})^2]^{1/2}}, \qquad 0 < a < L. \tag{13}$$

The kinematical boundary conditions are

$$v(0) = 0, \qquad v(L) = 0 \tag{14}$$

and as the bending moments also vanish at the ends, (1) and (6) give

$$\frac{d^2v}{da^2}(0) = 0, \qquad \frac{d^2v}{da^2}(L) = 0. \tag{15}$$

When Δ is considered given, equations (12) to (15) form the governing system from which the unknown function $v(a)$ and the unknown constant P are to be determined. If needed, function $u(a)$ can be calculated by post-processing from (11).

Numerical solution

To proceed numerically, it is convenient to use nondimensional quantities. We define

$$\xi = a/L, \quad \bar{v} = v/L, \quad \bar{\Delta} = \Delta/L, \quad \bar{P} = PL^2/EI, \quad ()' = d()/d\xi. \tag{16}$$

After some manipulations, equations (12) to (15) are transformed into

$$\bar{\Delta} - 1 + \int_0^1 [1 - (\bar{v}')^2]^{1/2} \mathrm{d}\xi = 0, \tag{17}$$

$$\bar{P}\bar{v} + \frac{\bar{v}''}{[1 - (\bar{v}')^2]^{1/2}} = 0, \quad 0 < \xi < 1, \tag{18}$$

$$\bar{v}(0) = 0, \quad \bar{v}(1) = 0, \tag{19}$$

$$\bar{v}''(0) = 0, \quad \bar{v}''(1) = 0. \tag{20}$$

To effect the numerical solution, the differential equation (18) is cast into a weak form

$$\bar{P}\int_0^1 \bar{v}\bar{w}\,\mathrm{d}\xi + \int_0^1 \frac{\bar{v}''}{[1 - (\bar{v}')^2]^{1/2}}\bar{w}\,\mathrm{d}\xi = 0, \tag{21}$$

where $\bar{w}(\xi)$ is the weighting function. We employ the Galerkin method for the discretization. As a demonstration case, simple three term approximation

$$\bar{v}(\xi) \approx \sum_j \bar{b}_j \bar{\varphi}_j = \bar{b}_1\bar{\varphi}_1(\xi) + \bar{b}_2\bar{\varphi}_2(\xi) + \bar{b}_3\bar{\varphi}_3(\xi) \tag{22}$$

is used with the basis functions

$$\bar{\varphi}_1(\xi) = \sin \pi\xi, \quad \bar{\varphi}_2(\xi) = \sin 3\pi\xi, \quad \bar{\varphi}_2(\xi) = \sin 5\pi\xi. \tag{23}$$

The anticipated symmetry of the solution with respect to point $\xi = 1/2$ has been taken into account in the selection of the basis functions. As the basis functions satisfy separately the boundary conditions (19) and (20) so does also the linear combination (22).

Due to the highly nonlinear nature of the problem, the equations must be solved iteratively. To this end, we put

$$\bar{v} = \bar{v}_0 + \Delta\bar{v}, \tag{24}$$

where $\bar{v}_0$ is assumed to be known from a previous iteration and try to solve for the unknown change $\Delta\bar{v}$. Expanding the nonlinear terms in (21) and (17) in series and keeping only the terms up to linear in $\Delta\bar{v}$ gives the forms

$$\bar{P}\int_0^1 (\bar{v}_0 + \Delta\bar{v})\bar{w}\,\mathrm{d}\xi + \int_0^1 \left(\frac{\bar{v}_0''}{\bar{r}_0^{1/2}} + \frac{1}{\bar{r}_0^{1/2}}\Delta\bar{v}'' + \frac{\bar{v}_0'\bar{v}_0''}{\bar{r}_0^{3/2}}\Delta\bar{v}'\right)\bar{w}\,\mathrm{d}\xi = 0, \tag{25}$$

$$\bar{\Delta} - 1 + \int_0^1 \left(\bar{r}_0^{1/2} - \frac{\bar{v}_0'}{\bar{r}_0^{1/2}}\Delta\bar{v}'\right)\mathrm{d}\xi = 0, \tag{26}$$

where the shorthand notation

$$\bar{r}_0 = 1 - (\bar{v}_0')^2 \tag{27}$$

has been used. Employing approximation (22), we have

$$\bar{v}_0 \approx \sum_j \bar{b}_j{}^0 \bar{\varphi}_j \tag{28}$$

and

$$\Delta\bar{v} \approx \sum_j \Delta\bar{b}_j \bar{\varphi}_j \tag{29}$$

with obvious meaning. After substituting (28) and (29) into (25) into (25) and (26), discrete equations are produced from the weak form by the Galerkin method by selecting consecutively $\bar{w} = \bar{\varphi}_1$, $\bar{w} = \bar{\varphi}_2$, $\bar{w} = \bar{\varphi}_3$. The final system equations are found to be

$$\begin{aligned} &\sum_j \bar{K}_{ij}\Delta\bar{b}_j + (\sum_j \bar{L}_{ij}\Delta\bar{b}_j)\bar{P} + \bar{M}_i\bar{P} + \bar{c}_i = 0, \quad i = 1,2,3, \\ &\sum_j \bar{N}_j \Delta\bar{b}_j + \bar{d} = 0, \end{aligned} \tag{30}$$

where

$$\begin{aligned} \bar{K}_{ij} &= \int_0^1 (\frac{1}{\bar{r}_0^{1/2}}\bar{\varphi}_i\bar{\varphi}_j'' + \frac{\bar{v}_0'\bar{v}_0''}{\bar{r}_0^{3/2}}\bar{\varphi}_i\bar{\varphi}_j')\,\mathrm{d}\xi = \int_0^1 (\frac{1}{\bar{r}_0^{1/2}}\bar{\varphi}_i\bar{\varphi}_j'' - \frac{\bar{P}_0\bar{v}_0\bar{v}_0''}{\bar{r}_0}\bar{\varphi}_i\bar{\varphi}_j')\,\mathrm{d}\xi, \\ \bar{L}_{ij} &= \int_0^1 \bar{\varphi}_i\bar{\varphi}_j\,\mathrm{d}\xi, \\ \bar{M}_i &= \int_0^1 \bar{v}_0\bar{\varphi}_i\mathrm{d}\xi, \\ \bar{c}_i &= \int_0^1 \frac{\bar{v}_0''}{\bar{r}_0^{1/2}}\bar{\varphi}_i\,\mathrm{d}\xi = -\bar{P}_0\int_0^1 \bar{v}_0\bar{\varphi}_i\,\mathrm{d}\xi, \\ \bar{N}_j &= -\int_0^1 \frac{\bar{v}_0'}{\bar{r}_0^{1/2}}\bar{\varphi}_j'\mathrm{d}\xi, \\ \bar{d} &= \bar{\Delta} - 1 + \int_0^1 \bar{r}_0^{1/2}\,\mathrm{d}\xi. \end{aligned} \tag{31}$$

The unknowns are $\Delta\bar{b}_1$, $\Delta\bar{b}_2$, $\Delta\bar{b}_3$, $\bar{P}$. It should be noted that that the equations are in spite of the linearisations still nonlinear. Terms (31) are evaluated numerically by the NIntegrate command and set (30) by the FindRoot command of the Mathematica program. The latter formulas for K_{ij} and c_i in (31) are used. They are obtained by replacing $\bar{v}_0''/\bar{r}_0^{1/2}$ with $-\bar{P}_0\bar{v}_0$ due to equation (18). It is hoped that this increases the accuracy of the calculations similarly as is explained in an analogous case in a buckling analysis on p. 90 in Reference [3]. (This remains to be checked via further calculations.) The value of the current $\bar{P}_0$ is estimated from (21) with $\bar{v} = \bar{v}_0$ and by taking $\bar{w} = \bar{v}_0$ which gives

$$\bar{P}_0 = -\frac{\int_0^1 \bar{v}_0 \bar{v}_0'' \bar{r}_0^{-1/2}\,\mathrm{d}\xi}{\int_0^1 \bar{v}_0^{\,2}\,\mathrm{d}\xi}. \tag{32}$$

The specific case $\alpha = 60^\circ$ (Figure 1) is is considered here. Reference [3] gives as the corresponding exact value $\Delta = 0.258980\cdots L$, which is used in the following. The calculations are started with the initial guess

$$\bar{b}_1^0 = 0.25, \quad \bar{b}_2^0 = 0, \quad \bar{b}_3^0 = 0, \tag{33}$$

for which (32) gives

$$\bar{P}_0 = 10.90 \quad (P_0 = 10.90\,EI/L^2). \tag{34}$$

The calculations converge in practice in three iterations giving the results

$$\bar{b}_1 = 0.2894, \quad \bar{b}_2 = -0.00145, \quad \bar{b}_3 = -0.000007, \quad \bar{P} = 11.45. \tag{35}$$

The values obtained for the maximum deflection and for the axial load are given in Table 1.

TABLE 1 Results for $v_{\max}$ and P

	$v_{\max}/L$	$P/(EI/L^2)$
Galerkin	0.291	11.45
Exact, [3]	0.2967…	11.367…

Considering the crude three term approximation, the accuracy achieved is rather satisfactory. The graph of the approximate function $v(a)$ is shown in Figure 3.

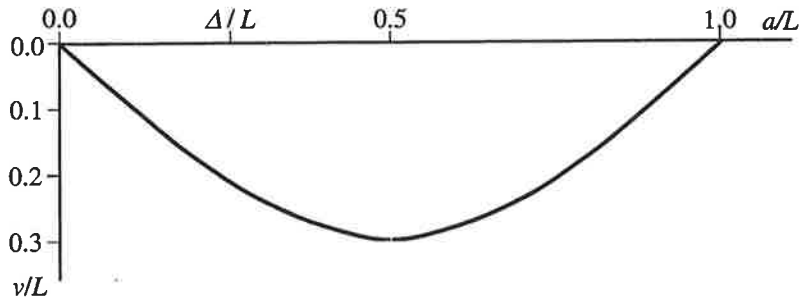


FIGURE 3 Function $v(a)$.

EULERIAN DESCRIPTION

Governing equations

Let us consider Figure 1(b) for the Eulerian description. If we for a while take the case where the force P is given and thus the displacement Δ is unknown, immediately one essential drawback of the Eulerian description in solid mechanics can be seen: the solution domain of the problem is not known in advance and must be determined as part of the problem. (In fluid mechanics this feature is not generally a difficulty as usually the solution domain is taken to be a given fixed domain of space through which the fluid flows. However, for instance in free surface flows this difficulty remains present.) To proceed simply, we thus consider Δ as given and the solution domain is now $\Delta \le x \le L$ at least for small enough Δ.

The counterparts of the incorrect formulas (2) and (3) are here now valid:

$$\tan\theta = \frac{\mathrm{d}v}{\mathrm{d}x}, \tag{36}$$

$$\kappa = \frac{-\frac{\mathrm{d}^2 v}{\mathrm{d}x^2}}{[1+(\frac{\mathrm{d}v}{\mathrm{d}x})^2]^{3/2}}. \tag{37}$$

An infinitesimal arc length element

$$\mathrm{d}s = [1+(\frac{\mathrm{d}v}{\mathrm{d}x})^2]^{1/2}. \tag{38}$$

From Figure 1(b), the original strut axis length from the left hand end to point Q^0 is $x-u(x)$. This must be equal to the curved axis length s to point Q in the deformed state due to the inextensibility assumption:

$$x-u(x) = s = \int_{\Delta}^{x}[1+(\frac{\mathrm{d}v}{\mathrm{d}x})^2]^{1/2}\mathrm{d}x. \tag{39}$$

From this, the horizontal displacement

$$u(x) = x - \int_{\Delta}^{x}[1+(\frac{\mathrm{d}v}{\mathrm{d}x})^2]^{1/2}\mathrm{d}x. \tag{40}$$

At $x = L$, $u = 0$ and we have from (40) a global inextensibility constraint equation

$$L = \int_{\Delta}^{x}[1+(\frac{\mathrm{d}v}{\mathrm{d}x})^2]^{1/2}\mathrm{d}x. \tag{41}$$

The bending moment at a generic point Q is again Pv, and making use of expressions (1) and (37) produces the equation

$$Pv = -EI\frac{\frac{d^2v}{dx^2}}{[1+(\frac{dv}{dx})^2]^{3/2}}, \qquad \Delta < x < L. \tag{42}$$

The boundary conditions are obtained similarly as in the Lagrangian formulation:

$$v(\Delta) = 0, \qquad v(L) = 0, \tag{43}$$

$$\frac{d^2v}{dx^2}(\Delta) = 0, \qquad \frac{d^2v}{dx^2}(L) = 0. \tag{44}$$

Equations (41) to (44) form the governing system from which the unknown function $v(x)$ and the unknown constant P are to be determined. If needed, function $u(x)$ can be calculated by post-processing from (40).

Numerical solution

To proceed numerically, it is again convenient to use nondimensional quantities. We define

$$\eta = (x-\Delta)/l, \quad \hat{v} = v/l, \quad \hat{\Delta} = \Delta/l, \quad \hat{P} = Pl^2/EI, \quad ()' = d()/d\eta \tag{45}$$

and repeat the steps used in the Lagrangian formulation. After some manipulations, equations (41) to (44) are transformed into

$$\hat{\Delta} - \int_0^1 [1-(\hat{v}')^2]^{1/2} d\eta = 0, \tag{46}$$

$$\hat{P}\hat{v} + \frac{\hat{v}''}{[1+(\hat{v}')^2]^{3/2}} = 0, \qquad 0 < \eta < 1, \tag{47}$$

$$\hat{v}(0) = 0, \qquad \hat{v}(1) = 0, \tag{48}$$

$$\hat{v}''(0) = 0, \qquad \hat{v}''(1) = 0. \tag{49}$$

To effect the numerical solution, differential equation (47) is cast into a weak form

$$\hat{P}\int_0^1 \hat{v}\hat{w}\, d\eta + \int_0^1 \frac{\hat{v}''}{[1+(\hat{v}')^2]^{3/2}}\hat{w}\, d\eta = 0, \tag{50}$$

where $\hat{w}(\eta)$ is the weighting function. The Galerkin method is employed for the discretization. A three term approximation

$$\hat{v}(\xi) \approx \sum_j \hat{b}_j \hat{\varphi}_j = \hat{b}_1 \hat{\varphi}_1(\eta) + \hat{b}_2 \hat{\varphi}_2(\eta) + \hat{b}_3 \hat{\varphi}_3(\eta) \tag{51}$$

is used with the basis functions

$$\hat{\varphi}_1(\eta) = \sin \pi\eta, \quad \hat{\varphi}_2(\eta) = \sin 3\pi\eta, \quad \hat{\varphi}_2(\eta) = \sin 5\pi\eta. \tag{52}$$

The anticipated symmetry of the solution with respect to point $\eta = 1/2$ has been taken into account in the selection of the basis functions. As the basis functions satisfy separately the boundary conditions (48) and (49) so does also the linear combination (51).

The equations must be solved iteratively. To this end, we put

$$\hat{v} = \hat{v}_0 + \Delta\hat{v}, \tag{53}$$

where $\hat{v}_0$ is assumed to be known from a previous iteration and try to solve for the unknown change $\Delta\hat{v}$. Expanding the nonlinear terms in (50) and (46) in series and keeping only the terms up to linear in $\Delta\hat{v}$ gives the forms

$$\hat{P}\int_0^1 (\hat{v}_0 + \Delta\hat{v})\hat{w}\,\mathrm{d}\eta + \int_0^1 \left(\frac{\hat{v}_0''}{\hat{s}_0^{3/2}} + \frac{1}{\hat{s}_0^{3/2}}\Delta\hat{v}'' - \frac{3\hat{v}_0'\hat{v}_0''}{\hat{s}_0^{5/2}}\Delta\hat{v}'\right)\hat{w}\,\mathrm{d}\eta = 0, \tag{54}$$

$$\hat{\Delta} + 1 - \int_0^1 \left(\hat{s}_0^{1/2} + \frac{\hat{v}_0'}{\hat{s}_0^{1/2}}\Delta\hat{v}'\right)\mathrm{d}\eta = 0, \tag{55}$$

where the notation

$$\hat{s}_0 = 1 + (\hat{v}_0')^2 \tag{56}$$

has been used. Employing approximation (51), we have

$$\hat{v}_0 \approx \sum_j \hat{b}_j{}^0 \hat{\varphi}_j \tag{57}$$

and

$$\Delta\hat{v} \approx \sum_j \Delta\hat{b}_j \hat{\varphi}_j \tag{58}$$

with obvious meaning. After substituting (57) and (58) into (54) and (55), discrete equations are produced from the weak form by the Galerkin method by selecting consecutively $\hat{w} = \hat{\varphi}_1$, $\hat{w} = \hat{\varphi}_2$, $\hat{w} = \hat{\varphi}_3$. The final system equations are found to be

$$\sum_j \hat{K}_{ij}\Delta\hat{b}_j + (\sum_j \hat{L}_{ij}\Delta\hat{b}_j)\hat{P} + \hat{M}_i\hat{P} + \hat{c}_i = 0, \quad i = 1,2,3,$$
$$\sum_j \hat{N}_j\Delta\hat{b}_j + \hat{d} = 0, \tag{59}$$

where

$$\hat{K}_{ij} = \int_0^1 (\frac{1}{\hat{s}_0^{3/2}}\hat{\varphi}_i\hat{\varphi}_j'' - \frac{3\hat{v}_0'\hat{v}_0''}{\hat{s}_0^{5/2}}\hat{\varphi}_i\hat{\varphi}_j')\,d\eta = \int_0^1 (\frac{1}{\hat{s}_0^{3/2}}\varphi_i\varphi_j'' + \frac{3\hat{P}_0\hat{v}_0\hat{v}_0'}{\hat{s}_0}\hat{\varphi}_i\hat{\varphi}_j')\,d\eta,$$
$$\hat{L}_{ij} = \int_0^1 \hat{\varphi}_i\hat{\varphi}_j\,d\eta,$$
$$\hat{M}_i = \int_0^1 \hat{v}_0\hat{\varphi}_i d\eta,$$
$$\hat{c}_i = \int_0^1 \frac{\hat{v}_0''}{\hat{s}_0^{3/2}}\hat{\varphi}_i\,d\eta = -\hat{P}_0\int_0^1 \hat{v}_0\hat{\varphi}_i\,d\eta, \tag{60}$$
$$\hat{N}_j = -\int_0^1 \frac{\hat{v}_0'}{\hat{s}_0^{1/2}}\hat{\varphi}_j'\,d\eta,$$
$$\hat{d} = \hat{\Delta} + 1 - \int_0^1 \hat{s}_0^{1/2}\,d\eta.$$

The unknowns are $\Delta\hat{b}_1$, $\Delta\hat{b}_2$, $\Delta\hat{b}_3$, $\hat{P}$. The latter formulas for $\hat{K}_{ij}$ and $\hat{c}_i$ in (60) are used. They are obtained by replacing $\hat{v}_0''/\hat{s}_0^{1/2}$ with $-\hat{P}_0\hat{v}_0$ due to equation (42). The idea behind this is similar to that explained in connection with the Lagrangian formulation. The value of the current $\hat{P}_0$ is estimated from (50) with $\hat{v} = \hat{v}_0$ and by taking $\hat{w} = \hat{v}_0$ which gives

$$\hat{P}_0 = -\frac{\int_0^1 \hat{v}_0\hat{v}_0''\hat{s}_0^{-3/2}d\eta}{\int_0^1 \hat{v}_0^2 d\eta}, \tag{61}$$

which is used in the following. The calculations are started with the initial guess (which corresponds to (33) with $\hat{b}_1 \approx L/l \bullet \bar{b}_1$)

$$\hat{b}_1^0 = 0.34, \quad \hat{b}_2^0 = 0, \quad \hat{b}_3^0 = 0, \tag{62}$$

for which (61) gives

$$\hat{P}_0 = 7.32 \quad (P_0 = 13.34\,EI/L^2). \tag{63}$$

The calculations converge in practice in three iterations giving the results

$$\hat{b}_1 = 0.4237, \quad \hat{b}_2 = 0.0101, \quad \hat{b}_3 = 0.0013, \quad \hat{P} = 6.45. \tag{64}$$

The values obtained for the maximum deflection and for the axial load are given in Table 2.

TABLE 2 Results for v_{max} and P

	v_{max}/L	$P/(EI/L^2)$
Galerkin	0.307	11.75
Exact, [3]	0.2967…	11.367…

The graph of the approximate function $v(x)$ is shown in Figure 4.

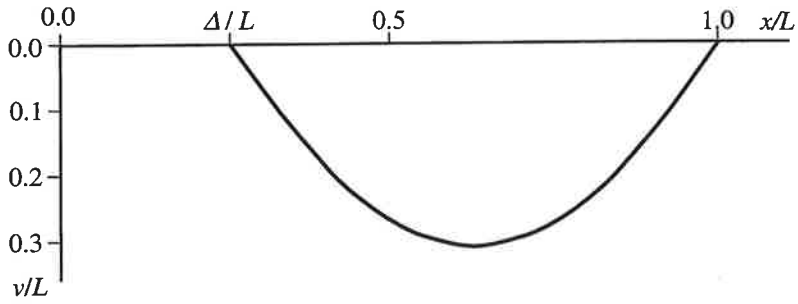


FIGURE 4 Function $v(x)$.

DISCUSSION

The Figures 3 and 4 presenting the lateral displacement as a function of the independent variable clearly show the differencies between the two formulations. In this respect the Eulerian presentation may be considered somewhat more illuminating. However, several problems of the Eulerian description in a more general case become apparent. One difficulty was mentioned in the beginning of the chapter on Eulerian description. Further, let us consider the case where the strut is not uniform in its properties, say its bending stiffness EI is not constant. We then know function $EI(a)$ but function $EI(x)$, which would be needed in the Eulerian description is in advance unknown. In the numerical example considered here the computational effort in the two formulations was about the same and no dramatic differencies in the accuracy were found. However, considering Figure 5 sketching the deformed shape of the strut in the case of a large Δ, again problems in the Eulerian description can be detected. First, the solution domain is no more known in advance; $v(x)$ is clearly double valued at certain subdomains of x. Second, the derivative dv/dx becomes unbounded at points R and S. (Some difficulties are to be expected at these points also in the Lagrangian description as now $dv/da = 1$ at R and $dv/da = -1$ at

S; see for instance formula (13).) Altogether, even if the simple strut problem under consideration could be solved by both the formulations with a comparable effort, in the light of this example it is quite obvious why the Lagrangian description is to be preferred in solid mechanics problems in general.

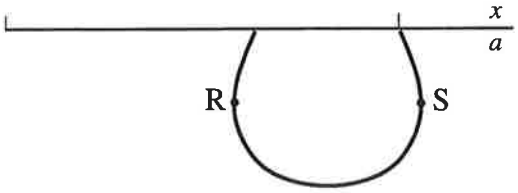


FIGURE 5 Deformation with a large Δ.

REFERENCES

1. L. E. Malvern, *Introduction to the Mechanics of a Continuous Medium*, Prentice-Hall, Englewood Cliffs, 1969.

1. M. S. El. Nashie, *Stress, Stability and Chaos in Structural Engineering: An Energy Approach*, McGraw-Hill, London, 1990.

3. S. P. Timoshenko and J. M. Gere, *Theory of Elastic Stability*, 2nd ed., Mc-Graw-Hill, New York, 1961.

4. S. Wolfram, *Mathematica, A System for Doing Mathematics by Computer*, 2nd ed., Addison-Wesley, Redwood City, 1991.

EQUATIONS FOR THE LATERAL BUCKLING OF THIN-WALLED BEAMS

MARTTI MIKKOLA and JUHA PAAVOLA
Laboratory of Structural Mechanics
Helsinki University of Technology
P.O. Box 2100, FIN-02015, HUT, FINLAND

ABSTRACT

In the present paper, the problem of non-distortional lateral buckling of a straight beam with a thin-walled cross-section is investigated. The principle of stationary potential energy is applied in deriving the equilibrium equations. The discrepancies which were observed before, between the basic equations derived by applying the energy principle and by the traditional equilibrium consideration are analyzed and explained.

1. INTRODUCTION

In the proceedings of the 5th Finnish Mechanics Days, held in Jyväskylä 1994 [1], the authors presented the derivation of the fundamental equilibrium equations for the combined lateral and torsional buckling of a straight beam with an arbitrarily shaped thin-walled cross-section. The application of the energy principle produced equations which slightly deviated from those traditionally obtained by the equilibrium consideration of an incremental element. In this paper it is shown that in traditional applications of various energy principles, the mutually inconsistent, linear or non-linear kinematics for the strain energy and the potential of external loads produces the discrepancies observed. The results obtained by using consistently same non-linear kinematics are in good agreement with the equilibrium consideration. However, the linear kinematics is sufficient to result in similar equations when certain more complete procedures are utilized. Taking in the initial state into account the transverse normal stress components σ_y^o and σ_z^o, in addition to the traditional ones σ_x^o and τ_{xs}^o, leads to correct results. This means, however, a step toward the two-dimensional analysis since the transverse normal stresses are not determined in the classical beam theory.

The deviations found concerned the lateral buckling of the beam only, while no problem was due to the torsional buckling. Hence in the continuation, in order to simplify the expressions, the torsional buckling will be left out of consideration.

2. BASIC KINEMATICS

As a frame, a global Cartesian coordinate system x, y, z with unit vectors $\vec{e}_x, \vec{e}_y, \vec{e}_z$ is defined. The axial coordinate x coincides with the beam axis, i.e. goes through the centroid of each cross-section plane. Coordinates y and z are the principal axes of the cross-section. In addition, coordinate s with unit vector $\vec{e}_s$ follows the centreline of the cross-section's wall and coordinate n with $\vec{e}_n$ is its normal. For simplicity, the beam is assumed to be composed of planar plates so that the cross-section is formed of piecewise straight sections as shown in Fig. 1.

While no distortion is present, the displacement field of a point on the middle surface of the wall follows from the usual assumptions made in the theory of thin-walled beams,

$$\vec{u}_o = (u - yv' - zw' - \omega\phi')\vec{e}_x + (v - (z - z_v)\phi)\vec{e}_y + (w + (y - y_v)\phi)\vec{e}_z, \qquad (1)$$

with u the axial displacement of the centroid, v and w the displacements of the shear center and ϕ the rotation of the cross section. The sectorial coordinate ω is defined according to Vlasov so that the linear strain e_{xs} disappears on the centreline of the cross section and, in addition, $e_{xn} \equiv 0$. y_v, z_v are the coordinates of the shear center. The loading is assumed to include only distributed loads $p_y^o(x)$ in the principal xy-plane and $p_z^o(x)$ in the xz-plane with no torsion. Coordinates a_y and a_z define the location of the line loads on the cross-section plane. When no axial load exists, i.e. $p_x^o(x) \equiv 0$, and thus $N^o(x) \equiv 0$ the axial displacement u in (1) can be dropped.

The expressions for strain components are composed of linear and non-linear parts:

$$\begin{aligned}
\epsilon_x &= e_x + \tfrac{1}{2}(e_x^2 + \theta_y^2 + \theta_z^2) \approx e_x + \tfrac{1}{2}(\theta_y^2 + \theta_z^2),\\
\epsilon_y &= e_y + \tfrac{1}{2}(\theta_z^2 + \theta_x^2),\\
\epsilon_z &= e_z + \tfrac{1}{2}(\theta_x^2 + \theta_y^2),\\
\gamma_{yz} &= 2e_{yz} - \theta_y\theta_z, \qquad (2)\\
\gamma_{zx} &= 2e_{zx} + e_x\theta_y - \theta_z\theta_x \approx 2e_{zx} - \theta_z\theta_x,\\
\gamma_{xy} &= 2e_{xy} - e_x\theta_z - \theta_x\theta_y \approx 2e_{xy} - \theta_x\theta_y.
\end{aligned}$$

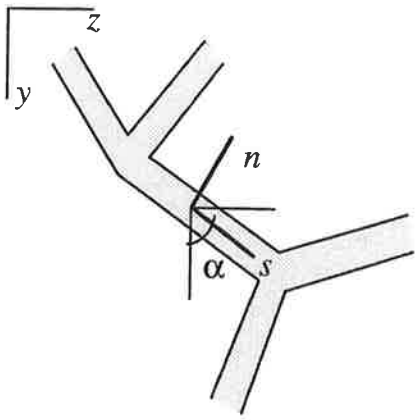


FIGURE 1. Thin-walled cross-section.

Here, the second order terms, quadratic in rotations are included only. The expressions for linear strain components derived in [1] are

$$\begin{aligned}
e_x &= -yv'' - zw'' - \omega\phi'' + n\left[(v'' - (z - z_v)\phi'')\sin\alpha - (w'' + (y - y_v)\phi'')\cos\alpha\right]\\
&\approx -yv'' - zw'' - \omega\phi'',\\
e_y &= e_z = e_{yz} = e_{zx} = e_{xy} \equiv 0. \qquad (3)
\end{aligned}$$

The term dropped in the normal strain is related to the bending of the walls of the cross-section. In addition, the shear strain $2e_{xs} = -2n\phi'$ associated with Saint Venant torsion is taken into account. The rotation components in (2) are

$$\theta_x = \tfrac{1}{2}\left(\frac{\partial \vec{u}}{\partial y}\cdot\vec{e}_z - \frac{\partial \vec{u}}{\partial z}\cdot\vec{e}_y\right) = \phi,$$
$$\theta_y = \tfrac{1}{2}\left(\frac{\partial \vec{u}}{\partial z}\cdot\vec{e}_x - \frac{\partial \vec{u}}{\partial x}\cdot\vec{e}_z\right) = -w' - (y - y_v)\phi', \tag{4}$$
$$\theta_z = \tfrac{1}{2}\left(\frac{\partial \vec{u}}{\partial x}\cdot\vec{e}_y - \frac{\partial \vec{u}}{\partial y}\cdot\vec{e}_x\right) = v' - (z - z_v)\phi'.$$

3. PRINCIPLE OF MINIMUM POTENTIAL ENERGY

The procedure follows exactly the linearized theory, called also EULER method, presented for example by NOVOZHILOV [2] and WASHIZU [3], according to which the incremental strain energy of the beam is

$$\Delta U = U_L + U_{NL} = \tfrac{1}{2}\int_V (\sigma_x e_x + \tau_{xs} 2e_{xs})\mathrm{d}V + \int_V (\sigma_x^o \epsilon_x + \tau_{xs}^o \gamma_{xs})\mathrm{d}V. \tag{5}$$

In the linear part, HOOKE's law between linear incremental strains and stresses is adopted. σ_x^o and τ_{xs}^o are the initial stresses in the initial position the stability of which will be studied,

$$\sigma_x^o = \frac{M_z^o}{I_z} y + \frac{M_y^o}{I_y} z, \qquad \tau_{xs}^o = -\frac{Q_y^o S_z(y)}{I_z t} - \frac{Q_z^o S_y(z)}{I_y t}. \tag{6}$$

The linear part U_L takes according to [1] the form

$$U_L = \tfrac{1}{2}\int_V \left[E(y^2(v'')^2 + z^2(w'')^2 + \omega^2(\phi'')^2) + 4Gn^2(\phi')^2\right]\mathrm{d}V, \tag{7}$$

while the non-linear one is splitted into two separate parts $U_{NL} = U_{NL,1} + U_{NL,2}$ in which the absence of torsional load at the initial state is taken into account

$$U_{NL,1} = \int_V \sigma_x^o(-yv'' - zw'' - \omega\phi'')\mathrm{d}V, \tag{8a}$$

$$U_{NL,2} = \int_V \tfrac{1}{2}\sigma_x^o[(v' - (z - z_v)\phi')^2 + (w' + (y - y_v)\phi')^2]\mathrm{d}V$$
$$+ \int_V \underline{\tau_{xs}^o}\phi[-(v' - \underline{(z - z_v)\phi')\sin\alpha} + (w' + \underline{(y - y_v)\phi')\cos\alpha}]\Big\}\mathrm{d}V. \tag{8b}$$

The expression for the potential of external loads is [1]

$$V = -\int_0^L (p_y^o v + p_z^o w)\mathrm{d}x, \tag{9}$$

---

[1] Traditionally in energy methods, as also in [1], the second order terms due to the kinematics are included in the expression of the potential energy, i.e.

$$V = -\int_0^L [p_y^o(v - a_y(1 - \cos\phi)) + p_z^o(w - a_z(1 - \cos\phi))]\mathrm{d}x$$
$$\approx -\int_0^L (p_y^o(v - \tfrac{1}{2}a_y\phi^2) + p_z^o(w - \tfrac{1}{2}a_z\phi^2))\mathrm{d}x, \tag{9'}$$

which is inconsistent with the initial kinematics applied.

The energy principle defines that the first variation of the total potential energy disappears, i.e. $\delta\Pi = \delta U_L + \delta U_{NL,1} + \delta U_{NL,2} + \delta V = 0$ with

$$\delta U_L = \int_0^L (EI_z v''\delta v'' + EI_y w''\delta w'' + EI_\omega \phi''\delta\phi'' + GI_t\phi'\delta\phi')\mathrm{d}x, \tag{10a}$$

$$\delta U_{NL,1} = \int_0^L (-M_z^o\delta v'' - M_y^o\delta w'')\mathrm{d}x, \tag{10b}$$

$$\delta U_{NL,2} = \int_0^L \Big\{M_z^o[(w' + 2\beta_y\phi')\delta\phi' + \phi'\delta w'] - M_y^o[(v' - 2\beta_z\phi')\delta\phi' + \phi'\delta v'] \tag{10c}$$

$$+\underline{Q_y^o[\beta_y(\phi'\delta\phi + \phi\delta\phi')}) + (w'\delta\phi + \phi\delta w')] + \underline{Q_z^o[\beta_z(\phi'\delta\phi + \phi\delta\phi')} - (v'\delta\phi + \phi\delta v')]\Big\}\mathrm{d}x,$$

$$\delta V = -\int_0^L (p_y^o\delta v + p_z^o\delta w)\mathrm{d}x. \tag{10d}$$

The familiar notation for the axial, two bending, warping and torsional rigidities, and the so-called Wagner coefficients

$$\int_A \mathrm{d}A = A, \quad \int_A y^2\mathrm{d}A = I_z, \quad \int_A z^2\mathrm{d}A = I_y, \quad \int_A \omega^2\mathrm{d}A = I_\omega, \quad \int_A 4n^2\mathrm{d}A = I_t,$$

$$\beta_y = \frac{1}{2I_z}\int_A y(y^2 + z^2)\mathrm{d}A - y_v, \qquad \beta_z = \frac{1}{2I_y}\int_A z(z^2 + y^2)\mathrm{d}A - z_v, \tag{11}$$

are introduced, and due to the assumptions concerning the coordinate system the following integrals over the cross-sectional area are assumed to vanish

$$\int_A y\mathrm{d}A = \int_A z\mathrm{d}A = \int_A \omega\mathrm{d}A = \int_A yz\mathrm{d}A = \int_A y\omega\mathrm{d}A = \int_A z\omega\mathrm{d}A = 0. \tag{12}$$

Under these assumptions, the terms $\delta U_{NL,1}$ in (10b) and δV in (10d) together result in the equilibrium conditions of the initial state, while δU_L in (10a) and $\delta U_{NL,2}$ in (10c) produce the differential equation system

$$\begin{aligned}
&EI_z v'''' + (M_y^o\phi)'' = 0,\\
&EI_y w'''' - (M_z^o\phi)'' = 0,\\
&EI_\omega\phi'''' - GI_t\phi'' - M_z^o w'' - 2\beta_y(M_z^o\phi')' + M_y^o v'' - 2\beta_z(M_y^o\phi')' + \underline{p_y^o\beta_y\phi} + \underline{p_z^o\beta_z\phi} = 0.
\end{aligned} \tag{13}$$

for lateral buckling of a thin-walled beam, and the boundary conditions

$$\begin{aligned}
\text{if}\quad \delta v' \neq 0, \qquad & EI_z v'' = 0,\\
\delta v \neq 0, \qquad & -EI_z v''' - (M_y^o\phi)' = 0,\\
\delta w' \neq 0, \qquad & EI_y w'' = 0,\\
\delta w \neq 0, \qquad & -EI_y w''' + (M_z^o\phi)' = 0,\\
\delta\phi' \neq 0, \qquad & EI_\omega\phi'' = 0,\\
\delta\phi \neq 0, \qquad & -EI_\omega\phi''' + GI_t\phi' + M_z^o(w' + 2\beta_y\phi') - M_y^o(v' - 2\beta_z\phi')\\
& \quad + \underline{Q_y^o\beta_y\phi} + \underline{Q_z^o\beta_z\phi} = 0,
\end{aligned} \tag{14}$$

at the ends of the beam $x = 0$ and $x = L$. The discrepancy arises from the fact that the traditional equilibrium consideration leads to equations

$$\begin{aligned} &EI_z v'''' + (M_y^o \phi)'' = 0, \\ &EI_y w'''' - (M_z^o \phi)'' = 0, \\ &EI_\omega \phi'''' - GI_t \phi'' - M_z^o w'' - 2\beta_y (M_z^o \phi')' + M_y^o v'' - 2\beta_z (M_y^o \phi')' + \underline{p_y^o a_y \phi} + \underline{p_z^o a_z \phi} = 0. \end{aligned} \tag{15}$$

where the underlined terms $p_y^o \beta_y \phi$ and $p_z^o \beta_z \phi$ in eq.$(13)_3$ are replaced by $p_y^o a_y \phi$ and $p_z^o a_z \phi$, respectively, and to the traditional boundary conditions,

$$\begin{aligned} \text{if} \quad \delta v' \neq 0, \quad & EI_z v'' = 0, \\ \delta v \neq 0, \quad & -EI_z v''' - (M_y^o \phi)' = 0, \\ \delta w' \neq 0, \quad & EI_y w'' = 0, \\ \delta w \neq 0, \quad & -EI_y w''' + (M_z^o \phi)' = 0, \\ \delta \phi' \neq 0, \quad & EI_\omega \phi'' = 0, \\ \delta \phi \neq 0, \quad & -EI_\omega \phi''' + GI_t \phi' + M_z^o (w' + 2\beta_y \phi') - M_y^o (v' - 2\beta_z \phi') \\ & \quad + \underline{Q_y^o a_y \phi} + \underline{Q_z^o a_z \phi} = 0, \end{aligned} \tag{16}$$

where the underlined terms $Q_y^o \beta_y \phi$ and $Q_z^o \beta_z \phi$ in eq.$(14)_6$ are replaced by $Q_y^o a_y \phi$ and $Q_z^o a_z \phi$, respectively.

4. PROBLEM DESCRIPTION

Consider a bit more accurately the underlined terms in (8b) which produce the deviating terms in the final differential equation system (13). Evaluating the variation of them and integrating by parts with respect to x yield the expression

$$\begin{aligned} &\delta \int_V \tau_{xs}^o [(z - z_v) \sin\alpha + (y - y_v) \cos\alpha] \phi \phi' \mathrm{d}V \\ &\quad = \int_V [(z - z_v) \sin\alpha + (y - y_v) \cos\alpha] (\tau_{xs}^o \phi' \delta\phi + \tau_{xs}^o \phi \delta\phi') \mathrm{d}V \\ &\quad = -\int_V [(z - z_v) \sin\alpha + (y - y_v) \cos\alpha] (\tau_{xs}^o)' \phi \delta\phi \mathrm{d}V \\ &\quad + \left[\int_A [(z - z_v) \sin\alpha + (y - y_v) \cos\alpha] \tau_{xs}^o \mathrm{d}A \phi \delta\phi \right]_0^L. \end{aligned} \tag{17}$$

When now the shear stress distribution (6) for τ_{xs}^o is substituted into this, following the classical beam theory, the expression

$$\begin{aligned} &\int_0^L \int_A [(z - z_v) \sin\alpha + (y - y_v) \cos\alpha] \Big(\frac{(Q_y^o)' S_z(y)}{I_z t} + \frac{(Q_z^o)' S_y(z)}{I_y t} \Big) \phi \delta\phi \mathrm{d}A \mathrm{d}x \\ &\quad + \left[\int_A [(z - z_v) \tau_{xz}^o + (y - y_v) \tau_{xy}^o] \mathrm{d}A \phi \delta\phi \right]_0^L. \end{aligned} \tag{18}$$

is obtained. Here, the term at the ends of the beam, with $\tau^o_{xs} = \tau^o_{xz}\sin\alpha + \tau^o_{xy}\cos\alpha$, is related to the rotation of the shear stress components, and is not to be avoided. It necessarily leads to the underlined terms in boundary conditions (14). In the integral term, the basic equations of the beam theory $(Q^o_y)' = -p_y(x)$ and $(Q^o_z)' = -p_z(x)$ are applied, in which the information concerning the location of the line loads on the cross-section disappears. As the result the underlined terms in equilibrium equations (13) are obtained. The surface integrals over the cross-sectional area required are

$$\begin{aligned}\int_A [(z-z_v)\sin\alpha + (y-y_v)\cos\alpha]S_z(y)\mathrm{d}A &= -\beta_y I_z t,\\ \int_A [(z-z_v)\sin\alpha + (y-y_v)\cos\alpha]S_y(z)\mathrm{d}A &= -\beta_z I_y t.\end{aligned} \tag{19}$$

This consideration shows inevitably that the shear stress distribution defined according to the beam theory can not take into account the location of loading at the cross-section plane and thus describe accurately enough the two-dimensional behaviour of a beam in lateral buckling.

5. TRANSVERSE NORMAL STRESSES

To avoid the problems due to the inaccurate shear stress distribution and the discrepancies following it, the general equilibrium equations are utilized to eliminate the shear stresses. The equilibrium equations at the initial state are

$$\begin{aligned}\frac{\partial\tau^o_{xy}}{\partial x} + \frac{\partial\sigma^o_y}{\partial y} + \frac{\partial\tau^o_{zy}}{\partial z} + f^o_y = 0 \quad &\Longrightarrow \quad \frac{\partial\tau^o_{xy}}{\partial x} = -\frac{\partial\sigma^o_y}{\partial y} - f^o_y,\\ \frac{\partial\tau^o_{xz}}{\partial x} + \frac{\partial\tau^o_{yz}}{\partial y} + \frac{\partial\sigma^o_z}{\partial z} + f^o_z = 0 \quad &\Longrightarrow \quad \frac{\partial\tau^o_{xz}}{\partial x} = -\frac{\partial\sigma^o_z}{\partial z} - f^o_z,\end{aligned} \tag{20}$$

with f^o_y and f^o_z the volume forces and $\tau^o_{yz} = \tau^o_{zy} \equiv 0$. Consider expression (17), and particularly the integral term in its final stage. Taking into account that $\tau^o_{xs} = \tau^o_{xz}\sin\alpha + \tau^o_{xy}\cos\alpha$ and substituting (20) yields

$$\begin{aligned}&-\int_V [(z-z_v)\sin\alpha + (y-y_v)\cos\alpha](\tau^o_{xs})'\phi\delta\phi\mathrm{d}V\\ &\quad = -\int_V [(z-z_v)(\tau^o_{xz})' + (y-y_v)(\tau^o_{xy})']\phi\delta\phi\mathrm{d}V\\ &\quad = \int_0^L\int_A [(z-z_v)(\frac{\partial\sigma^o_z}{\partial z} + f^o_z) + (y-y_v)(\frac{\partial\sigma^o_y}{\partial y} + f^o_y)]\mathrm{d}A\phi\delta\phi\mathrm{d}x.\end{aligned} \tag{21}$$

Integrating once by parts gives further

$$\begin{aligned}&\int_0^L (p^o_z a_z + p^o_y a_y)\phi\delta\phi\mathrm{d}x - \int_V (\sigma^o_z + \sigma^o_y)\phi\delta\phi\mathrm{d}V\\ &\quad = \tfrac{1}{2}\int_0^L (p^o_z a_z + p^o_y a_y)\delta(\phi)^2\mathrm{d}x - \tfrac{1}{2}\int_V (\sigma^o_z + \sigma^o_y)\delta(\phi)^2\mathrm{d}V,\end{aligned} \tag{22}$$

in which the volume forces and boundary forces are combined to the terms

$$\begin{aligned} p_z^o a_z &= \int_A (z - z_v) f_z^o \mathrm{d}A + \Big[\int_y (z - z_v)\sigma_z^o \Big]_{z=z_l}^{z=z_u} \mathrm{d}y, \\ p_y^o a_y &= \int_A (y - y_v) f_y^o \mathrm{d}A + \Big[\int_y (y - y_v)\sigma_y^o \Big]_{y=y_l}^{y=y_u} \mathrm{d}z. \end{aligned} \tag{23}$$

The first term in (22) is the same which is picked up into the traditional solution in the expression for the potential of external loads (9'), and the second one will be cancelled if in the the non-linear part of the strain energy (5) the transverse normal stresses σ_y^o and σ_z^o are also included, i.e.

$$U_{NL} = \int_V (\sigma_x^o \epsilon_x + \tau_{xs}^o \gamma_{xs} + \sigma_y^o \epsilon_y + \sigma_z^o \epsilon_z) \mathrm{d}V, \tag{24}$$

It is assumed, however, that in (24) for ϵ_y and ϵ_z only the rotation component θ_x is taken into account, while the components θ_y and θ_z in (4) are dropped. Including all the rotation components produces some terms which are dependent on the transverse normal stresses. However, in the final differential equation system they are summed with terms which are considerably greater. Hence, the meaning of the terms abandoned is small and the procedure is well argued. The term at the ends of the beam in (18) is still left in the boundary conditions. It reflects the fact that the shear stress distribution $(6)_2$ does not correspond to the distribution of the external load.

6. NON-LINEAR KINEMATICS

One way to derive the equilibrium equations is necessarily to apply in the expression for the strain energy the same kinematics which is referred and applied in the expression for the potential of external loads in footnote (9'). Hence, the displacement field

$$\begin{aligned} \vec{u}_o =& (-yv' - zw' - \omega\phi')\vec{e}_x + (v - (z - z_v)\phi - \underline{\tfrac{1}{2}(y - y_v)\phi^2})\vec{e}_y \\ &+ (w + (y - y_v)\phi - \underline{\tfrac{1}{2}(z - z_v)\phi^2})\vec{e}_z, \end{aligned} \tag{25}$$

instead (1) is applied. This assumption is in the slight disagreement with Vlasov's original idea about vanishing shear strains on the centreline of the cross-section. The additional components in the expressions for strains in (3) due to the underlined terms of (25) are

$$\begin{aligned} e_y = e_z &= -\tfrac{1}{2}\phi^2, \\ 2e_{zx} &= -(z - z_v)\phi\phi', \\ 2e_{xy} &= -(y - y_v)\phi\phi', \end{aligned} \tag{26}$$

of which the two latter ones produce, actually to the expression (8a) for $U_{NL,1}$, when $e_{xs} = e_{xy}\cos\alpha + e_{zx}\sin\alpha$, an additional term

$$U_{NL,1}^{+} = \int_V 2\tau_{xs}^o e_{xs} \mathrm{d}V = -\int_V \tau_{xs}^o [(z - z_v)\sin\alpha + (y - y_v)\cos\alpha]\phi\phi' \mathrm{d}V. \tag{27}$$

The comparison with (17) shows that the additional expression (27) directly cancels the problematic term in (8b), while the non-linear terms in the expression for the potential of external loads (9') produce the replacing terms required. This procedure can also, however, be interpreted as inconsistent, since the additional non-linear terms are included selectively by leaving most of them out of consideration. Its correspondence to the analogy of the traditional equilibrium consideration is obvious.

7. CONCLUSION

The traditional energy methods are usually applied by using mutually inconsistent kinematics for the strain energy and potential of external loads. This ad-hoc type of method has been working well with the experimental results, and possibly been interpreted as ingenious though its background seems not to be very solid. In this paper, it is proved that the linear kinematics is sufficient to yield the complete equilibrium equations when the transverse normal stresses are included together with the traditional beam stresses. It has also been shown that applying consistently the identical non-linear kinematics, although selectively by including only certain non-linear terms, the good agreement with the traditional equilibrium consideration can be reached.

REFERENCES

1. M. Mikkola and J.Paavola, "Re-Examination of the Equations of Lateral Buckling of Thin-Walled Beams", Proceedings of the 5th Finnish Mechanics Days, ed. R.A.E. Mäkinen and P. Neittaanmäki, University of Jyväskylä, Jyväskylä 1994, 87-94
2. V.V. Novozhilov, "Foundations of the Nonlinear Theory of Elasticity", 3rd ed., Graylock Press Rochester, New York, 1963
3. V.Z. Vlasov, "Thin-Walled Elastic Beams", Israel Program for Scientific Translations, Israel, 1963
4. K. Washizu, "Variational Methods in Elasticity and Plasticity", Pergamon Press Ltd., 2nd Edition, 1975
5. J.T. Oden, "Finite Elements of Nonlinear Continua", McGraw-Hill Book Company, Advanced engineering series, New York, 1972
6. W.F. Chen and T. Atsuta, "Theory of Beam-Columns, Volume 2 Space Behavior and Design", McGraw-Hill Book Company, New York, N.Y. 1977

MODELLING MONOLITHIC JOINTS OF PLANE FRAMES

MIKA REIVINEN*, EERO-MATTI SALONEN* and JUHA PAAVOLA**
*Laboratory of Theoretical and Applied Mechanics
**Laboratory of Structural Mechanics
Helsinki University of Technology, P.O.Box 1100
FIN-02015 HUT, Finland

ABSTRACT

A method for the plane frame analysis developed in [2] is shortly reviewed. The method is based on the use of a refined model for the monolithic joints of plane frames. A direct method which is used in optimizing the model for selected joint geometry is presented. In addition three iterative methods derived from it for the plane frame analyses are given. As a numerical example the convergence rate of these methods in the case of a simple statically indeterminate plane frame is studied.

1. INTRODUCTION

Traditionally in the plane frame analysis the beams are considered as one-dimensional structural members with dimensionless nodes. If the axes of the members do not intersect at a single point, a small rigid domain is usually employed. Figure 1 depicts a part of a plane frame and some notations used in the refined corner model presented by Reivinen, Salonen and Paavola [1].

In the model the corner is described using n+1 adjustable design parameters a_j. The first n parameters are the linear measures fixing the theoretical end points of the beam axes as shown in Figure 1. The last parameter $a_{n+1} = 1/A_C$, where A_C is the area of the so-called

core region. Six degrees of freedom (displacements u, v, rotation θ, constant strains ε_x, ε_y and γ_{xy}) are associated with the corner centroid node. Knowing the force resultants acting on the beam ends the values of the strains can be calculated for example using the principle of virtual work.

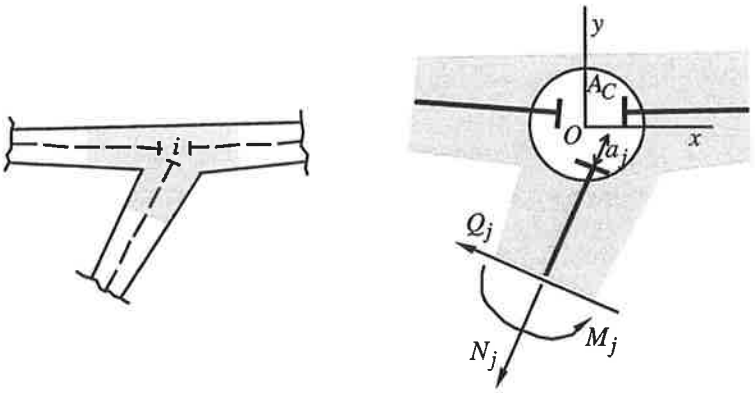


Figure 1 Frame corner i (shaded) and some details of the corner model.

For simplicity, a constant shear deformation γ_o is also assumed to each beam attached to the corner. For each joint geometry the design parameters a_j can be optimized in such a way that the flexibility matrices of the plane frame model and corresponding continuum model coincide as well as possible [1].

2. DIRECT DETERMINATION OF THE CORNER STRAINS

We consider next the joint i of the plane frame to be analyzed (Figure 1). It is assumed to be connected with n other joints by beams. Each joint has six degrees of freedom u_i, v_i, θ_i, ε_{xi}, ε_{yi}, γ_{xyi}. In the direct method all six degrees of freedom are unknown variables, which are solved from the system equations. The direct method is used in optimizing the design parameters for the selected joint geometry. We use the following notations

$$\{u\}_i = \left[u_i, v_i, \theta_i\right]^T, \tag{1}$$

$$\{\varepsilon\}_i = \left[\varepsilon_{x_i}, \varepsilon_{y_i}, \gamma_{xy_i}\right]^T \tag{2}$$

and (6×1)-vectors

$$\{d\}_i = \begin{Bmatrix} \{u\}_i \\ \{\varepsilon\}_i \end{Bmatrix}, \tag{3}$$

$$\{f\}_i = \begin{Bmatrix} \{f_1\}_i \\ \{f_2\}_i \end{Bmatrix}, \tag{4}$$

where $\{f_1\}_i$ and $\{f_2\}_i$ are (3×1)-load vectors. Let $\{n^i\}$ be a $(n\times1)$-vector, which contains the neighbour node numbers of node i, for example n^i_j refers to j'th node. The equations of equilibrium of the structure derived using the displacement method are of the form

$$[K]_{ii}\{d\}_i + \sum_{j=1}^{n}[K]_{ik}\{d\}_k = \{f\}_i, \quad k = n^i_j. \tag{5}$$

Each (6×6)-stiffness matrix in (5) can be consisted of submatrices using notations

$$[K]_{ik} = \begin{bmatrix} [K_{11}]_{ik} & [K_{12}]_{ik} \\ [K_{21}]_{ik} & [K_{22}]_{ik} \end{bmatrix}, \tag{6}$$

to obtain following more detailed system equations

$$[K_{11}]_{ii}\{u\}_i + [K_{12}]_{ii}\{\varepsilon\}_i + \sum_{j=1}^{n}[K_{11}]_{ik}\{u\}_k + \sum_{j=1}^{n}[K_{12}]_{ik}\{\varepsilon\}_k = \{f_1\}_i, \quad k = n^i_j, \tag{7}$$

$$[K_{21}]_{ii}\{u\}_i + [K_{22}]_{ii}\{\varepsilon\}_i + \sum_{j=1}^{n}[K_{21}]_{ik}\{u\}_k + \sum_{j=1}^{n}[K_{22}]_{ik}\{\varepsilon\}_k = \{f_2\}_i, \quad k = n^i_j. \tag{8}$$

The former of the equations is obtained by differentiating the potential energy Π with respect to the degrees of freedom $\{u\}_i$ and the latter by differentiating with respect to the degrees of freedom $\{\varepsilon\}_i$, respectively.

3. ITERATIVE FORMULATIONS

To enter the iterative methods we derive two subsystems from the equations (7) - (8). The iterative versions are derived mainly keeping in the mind the possible generalization of the joint model for the arbitrary shell structures where the joint "glues" two or more shell elements together. The quantities marked with the upper bar are assumed to be known from the previous iteration cycle. The equations concerning displacement degrees of freedom $\{u\}$ for each node i are

$$[K_{11}]_{ii}\{u\}_i + \sum_{j=1}^{n}[K_{11}]_{ik}\{u\}_k = \{f_1\}_i - [K_{12}]_{ii}\{\bar{\varepsilon}\}_i - \sum_{j=1}^{n}[K_{12}]_{ik}\{\bar{\varepsilon}\}_k, \quad k = n_j^i. \tag{9}$$

The equations concerning the strain quantities $\{\varepsilon\}$ for each node i are respectively

$$[K_{22}]_{ii}\{\varepsilon\}_i + \sum_{j=1}^{n}[K_{22}]_{ik}\{\varepsilon\}_k = \{f_2\}_i - [K_{21}]_{ii}\{\bar{u}\}_i - \sum_{j=1}^{n}[K_{21}]_{ik}\{\bar{u}\}_k, \quad k = n_j^i. \tag{10}$$

The displacements $\{u\}_i$ are solved from equations (9) and the strain quantities $\{\varepsilon\}_i$ from equations (10). The equations (9) - (10) form *the iteration version 1*. From the equations (10) we obtain further

$$[K_{22}]_{ii}\{\varepsilon\}_i = \{f_2\}_i - \sum_{j=1}^{n}[K_{22}]_{ik}\{\bar{\varepsilon}\}_k - [K_{21}]_{ii}\{\bar{u}\}_i - \sum_{j=1}^{n}[K_{21}]_{ik}\{\bar{u}\}_k, \quad k = n_j^i. \tag{11}$$

This leads to a more practical iteration form, where only 3×3 systems are to be solved. This is *the iteration version 2*. The final form can be obtained by splitting the matrix $[K_{22}]_{ii}$ in two parts

$$\left[K_{22}\right]_{ii} = \left[K_{22}^{c}\right]_{ii} + \left[K_{22}^{b}\right]_{ii}, \tag{12}$$

where c refers to the contribution of the core i and b to the contribution of the beams which are connected to the core. This leads to the following equations for the strain quantities

$$\left[K_{22}^{c}\right]_{ii}\{\varepsilon\}_i = \{f_2\}_i - \left[K_{22}^{b}\right]_{ii}\{\bar{\varepsilon}\}_i +$$

$$-\sum_{j=1}^{n}[K_{22}]_{ik}\{\bar{\varepsilon}\}_k - [K_{21}]_{ii}\{\bar{u}\}_i - \sum_{j=1}^{n}[K_{21}]_{ik}\{\bar{u}\}_k, \quad k = n_j^i, \tag{13}$$

which is *the iteration version 3.*

4. A NUMERICAL EXAMPLE

As a numerical example we consider a simple statically indeterminate plane frame depicted in Figure 2. The design parameters for the T-joints (nodes 2 and 3) are derived in the reference [2] and the values of parameters are found to be $a_1 = a_2 = a_3 = 0{,}270h$, $a_4 = a_5 = 0{,}999h$ and the inverse of the joint area $1/A_C = 1{,}596/h^2$, where h is the thickness of the vertical beams 4 and 5. The magnitude of the point load applied at the node 1 is $F = 0{,}025\,Eh^2$, where E is the Young´s modulus. The angle between load vector F and the horizontal plane is $\pi/4$.

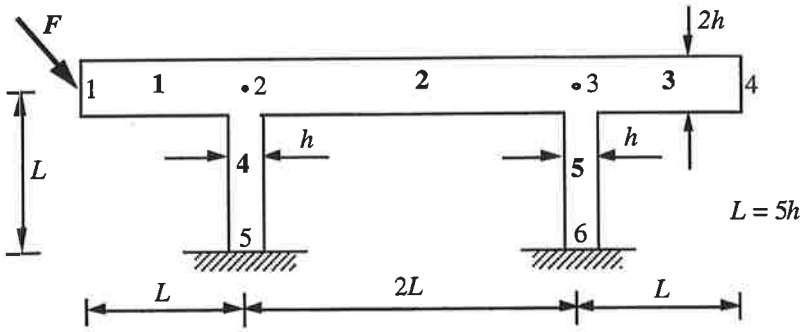


Figure 2 A statically indeterminate plane frame.

The ratio of the horizontal displacement of iterative version to the final displacement of node 2 as a function of iteration cycle is presented in the Figure 3. Using the design parameters the direct method gives for the horizontal displacement the value 0,176 *F/Eh*. The traditional

plane frame model (without design parameters) gives for the displacement the value 0,270 *F/Eh* whereas the continuum model gives the value 0,154 *F/Eh*.

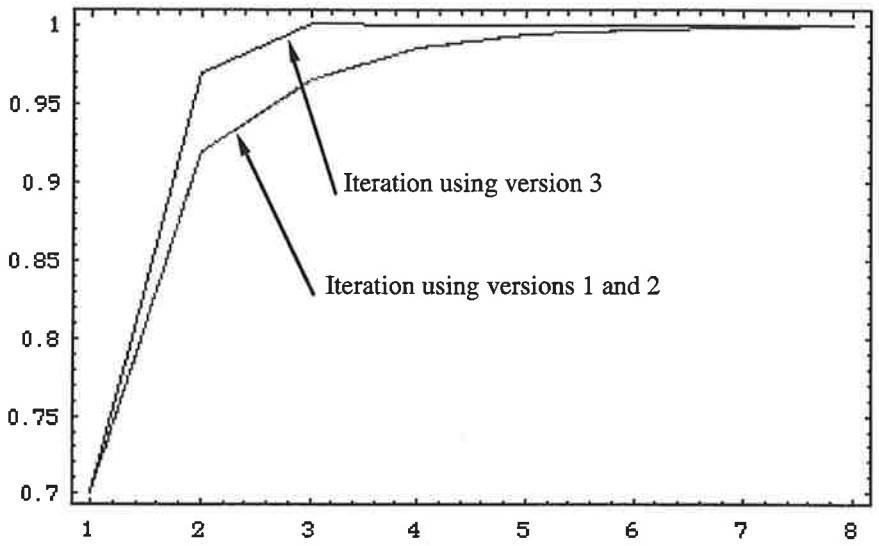


Figure 3 Horizontal displacement of node 2 using iteration versions 1 ... 3.

The refined model seems to describe the behaviour of the plane frame consisting of moderately thick beams with essentially better accuracy compared to the traditional plane frame analysis. In addition to this the third iteration version has a considerably faster converge rate compared to the two other versions.

5. CONCLUDING REMARKS

In this study three iterative versions with different convergence rate for the plane frame analysis with refined joint model derived from the direct method are presented. The constant shear deformation assumed for each plane frame beam can be handled in a similar way iteratively and an improvement to the convergence rate can also be obtained by a proper arrangement of the displacement terms as shown in the reference [2].

REFERENCES

1. Reivinen, M., Salonen, E-M. and Paavola J., *A Method for Refined Analysis of Box Beam Cross Sections,* Proceedings of the 5th Finnish Mechanics Days (Editors: Mäkinen Raino A. E., Neittaanmäki Pekka), Report 3/1994, University of Jyväskylä, Department of Mathematics, Laboratory of Scientific Computing.

2. Reivinen Mika, *Modelling the monolithic joints of plane frames* (in Finnish), Licentiate's thesis on Structural Mechanics, Helsinki University of Technology (1996).

A METHOD FOR SOLVING MARGINS OF SAFETY IN COMPOSITE FAILURE ANALYSIS

PETRI KERE and MARKKU PALANTERÄ
Helsinki University of Technology
Laboratory of Lightweight Structures
P.O. Box 4100
FIN – 02015 HUT, Finland

ABSTRACT

In failure analyses of structures, the distance between the applied load and the predicted failure load is often indicated in terms of so-called margin of safety. The probabilistic distributions of applied loads and material strengths can be taken into account by using appropriate factors of safety and material allowables. A generalized method for solving the margin of safety of a composite laminate with an iterative solver has been developed. Failure criterion functions are handled as "black boxes" whose internal formulation has no effect on the solution procedure. The problem is formulated as an unconstrained minimization problem where the objective function in the layer stress or strain space is minimized over a closed bounded interval by iteratively reducing the interval of uncertainty. The necessary definitions of margins of safety in the so-called constant and variable load approach are given. An example run is used for illustrating the method.

INTRODUCTION

Closed form solutions for reserve factors can be found for various failure criteria used for composite materials. For the constant and variable load approach, the determination of reserve factors for quadratic failure criteria is outlined in [3]. However, the formulations of failure criteria of composite materials may be quite complicated, for instance [4]. The failure criterion functions may be piecewise continuous by their nature, consisting of different expression for the assessment of different types of failure (e.g., fiber and matrix failure). Finding closed form solutions for this type of criteria is difficult or even impossible. Therefore, an iterative method was developed. The method has been implemented in ESAComp [5], an analysis and design software of composites, where numerous failure criterion functions are available. Moreover, the system provides support for user defined failure criterion functions.

The constant and variable load approach adopted for ESAComp is based on the partitioning of the applied loads into constant loads and variable loads [3]. Let the applied load vector be defined as

$$\mathbf{F} = \mathbf{F}^c + \mathbf{F}^v, \tag{1}$$

where the superscript c denotes the constant load part and the superscript v denotes the variable load part of the load vector. Separate FoS are associated with the constant and variable load parts. Hence, the effective load is defined by

$$\mathbf{F}_e = FoS^c\,\mathbf{F}^c + FoS^v\,\mathbf{F}^v. \tag{2}$$

The reserve factor (RF) measures the criticality of the effective load with respect to the failure load or, in other words, the applied load with respect to the allowed load. The RF values greater than one indicate positive margin to the allowed load. In the constant and variable load approach, the criticality of the load case is studied with with respect to the variation of the variable load. Thus, the critical load leading to failure of the layer is expressed as

$$\mathbf{F}_f = FoS^c\,\mathbf{F}^c + RF\,FoS^v\,\mathbf{F}^v. \tag{3}$$

The margin of safety (MoS) corresponding to the RF value is defined by the relation

$$MoS = RF - 1. \tag{4}$$

Margins of safety are often expressed as percentages. A negative MoS indicates how much the load has to be reduced to obtain an acceptable load level.

The failure loads corresponding to the constant and variable load vectors applied individually are defined by

$$\mathbf{F}_f = RF^c\,FoS^c\mathbf{F}^c \tag{5}$$

and

$$\mathbf{F}_f = RF^v\,FoS^v\mathbf{F}^v, \tag{6}$$

where RF^c and RF^v are the reserve factors for the constant and variable load vectors, respectively. The corresponding margins of safety are denoted by MoS^c and MoS^v, respectively.

For the computation of reserve factors and margins of safety, definitions for the RF (and MoS) values *infinity* and *indefinite* are given.

Definition. *Infinity* as a value of the reserve factors RF^c or RF^v means that the load vector can be increased without a limit (e.g., the midplane of a symmetric laminate in pure bending). In other words, no actual constant or variable load is applied. If a constant load is applied, the reserve factor RF is *infinity* only when RF^v equals to *infinity* and $RF^c \geq 1$ or RF^c equals to *infinity*.

Definition. A reserve factor is *indefinite* when the effective load causes failure and decreasing the magnitude of the variable load does not make the effective load non-critical. The limiting case where the effective load reaches the failure envelope from the outside is also defined as *indefinite*. For RF^c and RF^v *indefinite* is not a possible value since the origin of the load vector is always inside the failure envelope.

SEARCH BY GOLDEN SECTION

Among the derivative-free line search methods (the Fibonacci method, the golden section method, the dichotomous search method, and the uniform search method [1]) the Fibonacci method and the golden section method are the most effective algorithms. They make two functional evaluations at the first iteration and then only one evaluation at each of the subsequent iterations. However, to reach the desired accuracy the Fibonacci search requires the total number of iterations n to be chosen beforehand. Thus, the golden section method for minimizing a srictly quasiconvex function over the interval $[a_1, b_1]$ was selected for the solver.

The golden section ratio is defined by making the number k of allowed measurements points to approach infinity

$$\lim_{k\to\infty}\frac{F_{k-1}}{F_k}, \tag{7}$$

where the integers F_k are members of the Fibonacci sequence generated by the recurrence relation

$$F_k = F_{k-1} + F_{k-2},\ F_0 = F_1 = 1, k = 2,\ldots,n. \tag{8}$$

Hence, the resulting sequence is 1, 1, 2, 3, 5, 8, 13, ... The golden section ratio is achieved as a solution to the equation [1]

$$\gamma^2 + \gamma - 1 = 0. \tag{9}$$

Since γ must be in the interval (0, 1) then

$$\gamma = \frac{-1+\sqrt{5}}{2} \cong 0.618. \tag{10}$$

In the golden section method, the interval of uncertainty is reduced each time by a factor of 0.618. The number of observations required to achieve the desired degree of accuracy is given by

$$(0.618)^{k-1} \geq \frac{2\delta}{b_1 - a_1}, \tag{11}$$

where 2δ is the final length of uncertainty and $(b_1 - a_1)$ is the initial interval of uncertainty. For k large enough, $1/F_k = 2\delta/(b_1 - a_1)$ is asymptotic to $(0.618)^{k-1}$.

AN ITERATIVE METHOD FOR SOLVING MARGINS OF SAFETY

Let the layer actual stresses in the plane stress state (in a layer coordinate plane 12) caused by the load vectors $FoS^c\ \mathbf{F}^c$ and $FoS^v\ \mathbf{F}^v$ be σ^c and σ^v, respectively. In a linear analysis, the layer stress vector corresponding to the failure load in eq. (3) is written as

$$\boldsymbol{\sigma}_f = \boldsymbol{\sigma}^c + RF\boldsymbol{\sigma}^v \tag{12}$$

The resultant load and the loads corresponding to the constant and variable load vectors applied individually are written as

$$\boldsymbol{\sigma}(\lambda) = \boldsymbol{\sigma}^c + \lambda\boldsymbol{\sigma}^v \tag{13}$$

$$\boldsymbol{\sigma}(\lambda) = \lambda\boldsymbol{\sigma}^c \tag{14}$$

$$\boldsymbol{\sigma}(\lambda) = \lambda\boldsymbol{\sigma}^v, \tag{15}$$

where λ denotes the load criticality factor. The value of the generalized failure criterion function is obtained from

$$f(\lambda) = f(\boldsymbol{\sigma}(\lambda)). \tag{16}$$

The failure criterion function indicates failure when

$$f(\lambda) \geq 1. \tag{17}$$

Accordingly, λ can be determined in strain space by applying the equivalent strains corresponding to the actual stresses caused by the constant and variable load vectors.

The problem is formulated as an unconstrained minimization problem

$$\begin{aligned} &\min \theta(\lambda) = |f(\lambda) - 1| \\ &\lambda = \{\lambda_k \,|\, \lambda_k \in [a_k, b_k]\}' \end{aligned} \tag{18}$$

where the objective function $\theta(\lambda)$ is minimized over the closed bounded interval by iteratively reducing the interval of uncertainty $[a_k, b_k]$. First, the initial interval of uncertainty is determined. In the main step, the golden section line search method is used to search for the point where the failure occurs.

Searching the initial interval of uncertainty is summarized as follows.

Initialization step. Let $\sigma^c \neq 0$ and $\sigma^v \neq 0$. Let the allowable final length of uncertainty be $2\delta>0$ and the terminal condition for the infinite λ be INF>0, and let $k=1$ and $\lambda_1=1$. Thus, $f_1=f(\lambda_1)$ and go to the main step.

Main step.

1. If $f_1 < 1$, go to step 2.
   If $f_1 > 1$, go to step 3.
   If $f_1 = 1$, go to step 6.

2. $\lambda_{k+1} = 2\lambda_k$;
   $k = k+1$;
   $f_k = f(\lambda_k)$, go to step 4.

3. $\lambda_{k+1} = 0.5\lambda_k$;
   $k = k+1$;
   $f_k = f(\lambda_k)$, go to step 5.

4. If $f_k < 1$ and $\lambda_k \leq$ INF, go to step 2.
   If $f_k > 1$ and $\lambda_k \leq$ INF, $[a_k, b_k] = [\lambda_{k-1}, \lambda_k]$, go to the sequential line search.

5. If $f_k > 1$ and $\lambda_k \geq 2\delta$, go to step 3.
   If $f_k > 1$ and $\lambda_k < 2\delta$, λ is *indefinite*.
   If $f_k < 1$ and $\lambda_k > 2\delta$, $[a_k, b_k] = [\lambda_k, \lambda_{k-1}]$, go to the sequential line search.

6. If $f_k(1-\delta) < 1, \lambda_k = 1$.
   If $f_k(1-\delta) \geq 1, \lambda_k$ is *indefinite*.

The golden section line search method is next used to focus on the failure point. The golden section method makes two functional evaluations at the first iteration and then only one evaluation at each of the subsequent iterations.

Initialization step. Let $2\delta>0$ be the allowed final length of the uncertainty and let $[a_1, b_1]$ be the initial interval of uncertainty at $k=1$. Now $x_1=a_1+\gamma(b_1-a_1)$ and $y_1=b_1+\gamma(b_1-a_1)$. Evaluate $\theta(x_1)$ and $\theta(y_1)$, and go to the main step.

Main step.

1. if $b_k - a_k < 2\delta$, stop; the solution lies in the interval $[a_k, b_k]$ and is thus $\lambda = (a_k + b_k)/2$.
   if $\theta(x_k) > \theta(y_k)$, go to step 2.
   if $\theta(x_k) \leq \theta(y_k)$, go to step 3.

2. $a_{k+1} = a_k$ and $b_{k+1} = x_k$;
   $x_{k+1} = y_k$ and $y_{k+1} = b_{k+1} - \gamma(b_{k+1} - a_{k+1})$;
   $\theta(x_{k+1}) = \theta(y_k)$;
   Evaluate $\theta(y_{k+1})$; and go to step 4.

3. $a_{k+1} = y_k$ and $b_{k+1} = b_k$;
   $y_{k+1} = x_k$ and $x_{k+1} = a_{k+1} + \gamma(b_{k+1} - a_{k+1})$;
   $\theta(y_{k+1}) = \theta(x_k)$;
   Evaluate $\theta(x_{k+1})$; and go to step 4.

4. $k = k + 1$, and go to step 1.

The value λ corresponds to RF, RF^c, or RF^v depending on whether the stress state of equation (13), (14), or (15) is used. The use of the golden section method is demonstrated in the following example.

AN EXAMPLE RUN

The iterative method is studied through an example run in which the first ply failure analysis based on the classical lamination theory is performed for a symmetric graphite/epoxy laminate. The laminate structure and the ply properties corresponding to a typical unidirectional T800/epoxy ply are given in Table 1.

The Hashsin criterion for unidirectional plies introduced in [2] is applied. Under tensile longitudinal loads the expression for predicting longitudinal failure is

$$f_a = \left(\frac{\sigma_1}{X_t}\right)^2 + \left(\frac{\tau_{12}}{S}\right)^2 \qquad \sigma_1 \geq 0. \tag{19}$$

Under compressive loads longitudinal failure is predicted with an independent stress condition

$$f_a = -\frac{\sigma_1}{X_c} \qquad \sigma_1 < 0\,. \tag{20}$$

In the case of tensile transverse stress the expression for matrix failure is

$$f_b = \left(\frac{\sigma_2}{Y_t}\right)^2 + \left(\frac{\tau_{12}}{S}\right)^2 \qquad \sigma_2 \geq 0\,. \tag{21}$$

A more complicated expression is used when the transverse stress is compressive

$$f_b = \left(\frac{\sigma_2}{2S}\right)^2 + \left[\left(\frac{Y_c}{2S}\right)^2 - 1\right]\frac{\sigma_2}{Y_c} + \left(\frac{\tau_{12}}{S}\right)^2 \qquad \sigma_2 < 0\,. \tag{22}$$

The more critical of the two failure modes is selected

$$f = \max(f_a, f_b). \tag{23}$$

Table 1. Laminate structure and ply properties.

Laminate lay-up	(2(+45/–45)/2(0/0/90)/0)s	
Laminate thickness	4.40 mm (t = 0.20 mm)	
Ply engineering constants	**Ply failure stresses**	
E_1 = 155.0 GPa	X_t = 2000 MPa	X_c = 1500 MPa
E_2 = 8.5 GPa	Y_t = 40 MPa	Y_c = 220 MPa
G_{12} = 5.5 Gpa	S = 80 MPa	
ν_{12} = 0.30		
Ply thermal expansion coefficients	α_1 = –0.50 e–6/°C	α_2 = 30 e–6/°C

A load case used in the study includes a temperature difference ΔT=–50°C in the constant load part with FoS^c=1.3 and a mechanical load N_x=–1000 kN/m and N_{xy}=700 kN/m in the variable load part with FoS^v=1.5. The temperature difference from the stress-free environment of the laminate is constant through the laminate.

Let the allowable final length of uncertainty be at most 0.01. Table 2 shows the computations for determining the initial interval of uncertainty for RF. Summary of computations by golden section method is given in Table 3.

Table 4 summarizes layer failure modes, reserve factors, and margins of safety. The values of the most critical layers are taken as the laminate values. Failure mode is predicted based on which of the equations (19)...(22) is active when failure load is reached. The critical layers with respect to the combined effective load are the –45° oriented layers whose margins of safety are negative. The effective constant and variable loads do not cause failure when applied alone.

Table 2. Summary of computations for determining the initial interval of uncertainty.

k	f_k	a_k	b_k
Layer orientation +45°			
1	0.271495	0	1
2	1.11687	1	2
Layer orientation –45°			
1	1.09473	0	1
2	0.496539	0.5	1
Layer orientation 0°			
1	0.971548	0	1
2	3.4737	1	2
Layer orientation 90°			
1	0.838801	0	1
2	3.36011	1	2

Table 3. Summary of computations by golden section method.

k	a_k	b_k	x_k	y_k	$\theta(x_k)$	$\theta(y_k)$
Layer orientation +45°						
1	1	2	1.61803	1.38197	0.273797	0.473364
2	1.38197	2	1.76393	1.61803	0.134461	0.273797
3	1.61803	2	1.8541	1.76393	0.0422359	0.134461
4	1.76393	2	1.90983	1.8541	0.017096	0.0422359
5	1.8541	2	1.94427	1.90983	0.0546567	0.017096
6	1.8541	1.94427	1.90983	1.88854	0.017096	0.0057772
7	1.8541	1.90983	1.88854	1.87539	0.0057772	0.0197836
8	1.87539	1.90983	1.89667	1.88854	0.00292889	0.0057772
9	1.88854	1.90983	1.9017	1.89667	0.00832853	0.00292889
10	1.88854	1.9017	1.89667	1.89357	0.00292889	0.00040102
11	1.88854	1.89667	1.89357	1.89165	0.00040102	0.00245625

Table 3. Summary of computations by golden section method (continued).

k	a_k	b_k	x_k	y_k	$\theta(x_k)$	$\theta(y_k)$
Layer orientation –45°						
1	0.5	1	0.809017	0.690983	0.16297	0.304185
2	0.690983	1	0.881966	0.809017	0.068798	0.16297
3	0.809017	1	0.927051	0.881966	0.00796211	0.068798
4	0.881966	1	0.954915	0.927051	0.0306427	0.00796211
5	0.881966	0.954915	0.927051	0.90983	0.00796211	0.0314369
6	0.90983	0.954915	0.937694	0.927051	0.00669289	0.00796211
7	0.927051	0.954915	0.944272	0.937694	0.0158063	0.00669289
8	0.927051	0.944272	0.937694	0.933629	0.00669289	0.00108194
9	0.927051	0.937694	0.933629	0.931116	0.00108194	0.00237763
10	0.931116	0.937694	0.935182	0.933629	0.0032232	0.00108194
Layer orientation 0°						
1	1	2	1.61803	1.38197	1.32122	0.730544
2	1	1.61803	1.38197	1.23607	0.730544	0.411929
3	1	1.38197	1.23607	1.1459	0.411929	0.232755
4	1	1.23607	1.1459	1.09017	0.232755	0.128795
5	1	1.1459	1.09017	1.05573	0.128795	0.0671323
6	1	1.09017	1.05573	1.03444	0.0671323	0.0300114
7	1	1.05573	1.03444	1.02129	0.0300114	0.00744701
8	1	1.03444	1.02129	1.01316	0.00744701	0.00635429
9	1	1.02129	1.01316	1.00813	0.00635429	0.0148289
10	1.00813	1.02129	1.01626	1.01316	0.00109567	0.00635429
11	1.01316	1.02129	1.01818	1.01626	0.00216238	0.00109567
Layer orientation 90°						
1	1	2	1.61803	1.38197	1.19843	0.603232
2	1	1.61803	1.38197	1.23607	0.603232	0.282267
3	1	1.38197	1.23607	1.1459	0.282267	0.101809
4	1	1.23607	1.1459	1.09017	0.101809	0.00287866
5	1	1.1459	1.09017	1.05573	0.00287866	0.0649664
6	1.05573	1.1459	1.11146	1.09017	0.0364918	0.00287866
7	1.05573	1.11146	1.09017	1.07701	0.00287866	0.0268297
8	1.07701	1.11146	1.0983	1.09017	0.0120695	0.00287866
9	1.07701	1.0983	1.09017	1.08514	0.00287866	0.0120615
10	1.08514	1.0983	1.09328	1.09017	0.0028179	0.00287866
11	1.09017	1.0983	1.09519	1.09328	0.00634668	0.0028179

Table 4. Summary of layer failure modes (tf = tensile fiber mode, tm = tensile matrix mode, cf = compressive fiber mode, and cm = compressive matrix mode), reserve factors, and margins of safety.

Layer Orientation	*RF*	*MoS* (%)	Failure mode	RF^c	MoS^c (%)	Failure mode	RF^v	MoS^v (%)	Failure mode
+45°	1.89	89	tf	2.62	162	tm	1.87	87	tf
−45°	0.93	−7	cm	2.62	162	tm	1.21	21	cf
0°	1.02	2	cm	2.71	171	tm	1.09	9	tm
90°	1.09	9	tf	2.53	153	tm	1.09	9	tf
Laminate									
	0.93	−7	cm	2.53	153	tm	1.09	9	tf

CONCLUSIONS

Closed form solutions for reserve factors can be found for various failure criteria used for composite materials. However, finding closed form solutions for some criteria may be difficult or even impossible. Therefore, a generalized method for solving the laminate reserve factors with an iterative solver has been developed. The internal formulation of failure criterion functions has no effect on the solution procedure. On the basis of the laminate reserve factors, margins of safety of the composite laminate are computed.

The problem is formulated as an unconstrained minimization problem where the objective function in the layer stress or strain space is minimized over a closed bounded interval by iteratively reducing the interval of uncertainty. Results show that the golden section line search method that is used in the solver is an efficient algorithm for finding the layer failure point. In a reasonably number of iterations, an adequate degree of accuracy for composite laminate design is obtained.

REFERENCES

1. Bazaraa, M. S., *Nonlinear Programming: Theory and Algorithms*. 2nd edition. Wiley, New York (NY), USA, 1993.

2. Hashin, Z., *Failure Criteria for Unidirectional Fiber Composites*. Journal of Applied Mechanics Materials Vol. 47 (1980), pp. 329–334.

3. Palanterä, M., Klein, M., *Constant and Variable Loads in Failure Analyses of Composite Laminates*. In Computer Aided Design in Composite Material Technology IV, pp. 221–228. Southampton, Computational Mechanics Publications, 1994.

4. Puck, A., *Progress in Composite Component Design through Advanced Failure Models*. In Proceedings of the 17th International SAMPE Europe Concference of the Society for the Advancement of Material and Process Engineering, pp. 83–96. Basel, SAMPE European Chapter, 1996.

5. Saarela O., Häberle J., Klein M., *Composite Analysis and Design System ESAComp*. In Computer Aided Design in Composite Material Technology IV, pp. 23–30. Southampton, Computational Mechanics Publications, 1994.

ANALYSIS OF REINFORCED CONCRETE STRUCTURES FOR FAST EXPLOSION LOAD

Pentti Varpasuo
IVO POWER ENGENEERING LTD
01019 IVO, FINLAND

ABSTRACT: The study forms a part of ongoing development project of the VVER-91 NPP concept. The reactor cavity of the plant is equipped with a core catcher. The pressure histories on the cavity wall at various locations were developed by T.G. Theofanous. The results of the analyses show that the cavity can carry the prescribed loads within allowable deformation limits and that the zones where the crushing of concrete takes place are local and restricted.

1 INTRODUCTION

The strength of the VVER-91 reactor cavity has been studied before by using ABAQUS/STANDARD, ABAQUS/EXPLICIT and ANSYS programs. These studies have been reported in reference[1]. None of these studies gave satisfactory results. It appeared that ABAQUS/STANDARD was able to estimate the static load capacity of the structure but could not be used for the dynamic analysis. ABAQUS/EXPLICIT was capable to do the dynamic analysis of the structure but the results were not fully consistent. The time histories for hoop reinforcement and the time histories for strain in hoop direction were not compatible with each other. The next program to try was ANSYS. ANSYS was capable to perform the static analysis and gave reliable results whereas the ANSYS results in dynamic analysis were completely unsatisfactory and clearly erroneous.
So it was decided in the beginning of 1995 to acquire a new special purpose reinforced concrete analysis program ANACAP from Anatech Research Corporation San Diego, California. ANACAP is advanced finite element modelling program for the thermal and stress evaluation of reinforced and prestressed concrete structures.

The aim of this study is to analyse the VVER-91 reactor cavity for non-axisymmetric loads using ANACAP program. The task is the evaluation of the dynamic response of the reactor cavity against the pressure impulse caused by the steam explosion. The steam explosion occurs when the molten core from the reactor pressure vessel reacts with the water in the reactor cavity. The resulting loading is a very fast impulse type load. In cavity strength analyses two different reinforcement amounts in hoop and vertical directions were stipulated, namely, 200 kg/m3 or 2.4 % of the concrete area or 300 kg/m3 or 3.6 % of the concrete area.

2 DESCRIPTION OF THE STRUCTURE

The reactor cavity is a reinforced concrete cylinder which supports the reactor pressure vessel and surrounds the lower part of the pressure vessel. The thickness of the cylinder varies along the height. The lower five meters are 2.6 meters thick and the upper three meters are 1.83 meters thick. The total height of the cylinder is eight meters. The inner radius of the cylinder is 2.9 meters. At the bottom of the cavity there is an opening which opens to the control room of the reactor inspection machine.
The cylinder of the reactor cavity is supported by the base slab of the reactor building on the level +3.00. The upper edge of the cavity is supported on the level +11.00. For the analysis the following basic assumptions concerning the materials were made. The concrete is of class K35-1 and the reinforcement is of the class A500 HW. The amount of the reinforcement is 50 kg/m3 in radial direction. The amount of reinforcement in hoop and vertical directions was varied from 200 kg/m3 (2.4 % of the concrete area) to 300 kg/m3 (3.6 % of the concrete area).

3 DESCRIPTION OF THE LOAD

When the molten material discharges from the reactor vessel, it comes into contact with water in the reactor cavity. A violent ex-vessel steam explosion is a possible outcome of such a contact. The ex-vessel steam explosion may threaten the structural integrity of the reactor cavity walls, which support the reactor vessel and the primary piping.
In order to make quantitative evaluations of the pressure impulse on the cavity walls, one has to consider first the core melt discharge from the vessel. Two possible mechanisms for the vessel failure has been studied in reference [2]. Scenario I is the local creep rupture, and Scenario II is the global lower head failure.
The calculations of ex-vessel steam explosion energetics were made for typical ex-vessel conditions with subcooled water, entry velocities of melt to the water are about 10 m/s and the water pool depth is 1 - 3 m.
The applied analysis methods PM-ALPHA and ESPROSE.m are mechanistic except for two assumptions. These assumptions are the melting pour rate and the degree of

melt break-up during the premixing. The melting pour rate is 1000 kg/s of oxidic pour with a diameter of 60 cm, 100K superheat and a particle size of 1 cm.
Off-center pours with distances of 0.5, 1, 1.5 and 2 meters from the sidewall in reactor cavity equipped with a core catcher are considered in reference [3]. The core catcher is a cone-shaped steel cone, resting on the cavity floor and extending all the way to the cavity sidewalls. The water depth is taken as 2 meters, the melt pour rate at 1000 kg/s, and the melt particle size at 1 cm, so as to match the conditions in previous calculations run for the cavity and reported in reference [2].
The results indicate that local peak loads can be quite high, due to the proximity of the explosion zone; however, the duration becomes increasingly shorter as the pour approaches the side and effective water depth decreases due to the conical shape. The results also show a complicated wave reflection pattern that affects both timing and load distribution on the boundaries.
In the following Figures 1 and 2 the wall pressures and impulses at selected elevations are presented for the off-center pour with 0.5 meter distance from the side wall. This run appeared to yield the highest wall pressure and impuse values from all considered pour distances. The run was carried out with the aid of ESPROSE.m using planar cavity model.

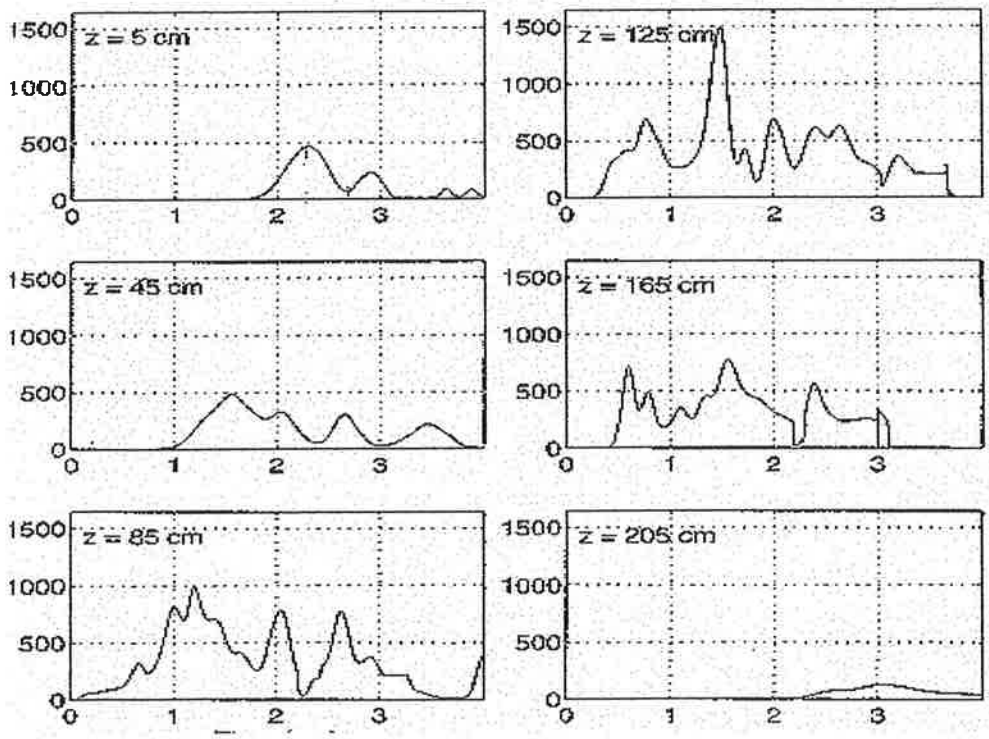


Figure 1. Wall pressure at selected elevations after Theofanous [2]. Pressure in bars, time in milliseconds.

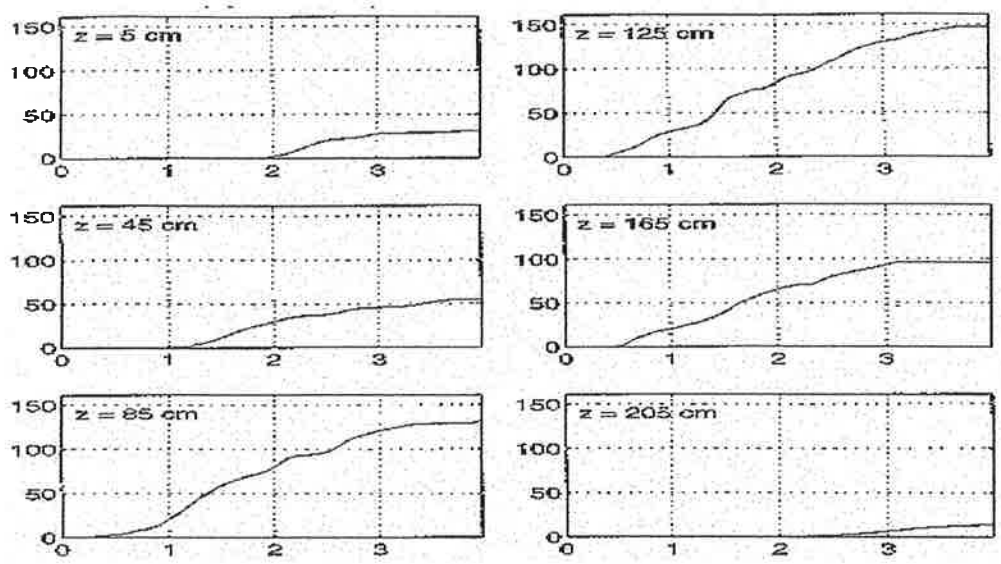


Figure 2. Wall impulse (kPa-s) at selected elevations after Theofanous [2].

4 DESCRIPTION OF ANACAP REINFORCED CONCRETE MODEL

The ANACAP model is a development of the smeared-crack model [4]. Since its inception the model has been developed to include compressive plasticity, tensile and compressive strain softening, cyclic behavior, shear transfer across cracks, high temperature creep and temperature induced stiffness and strength degradation.
The smeared-crack models the local crack by redefining the incremental constitutive matrix to reflect the decreased stiffness in the direction of the crack. The model requires that the cracks at specific material point are mutually orthogonal but cracks can have independent histories of opening and closing under cyclic load. In contrast with the fracture mechanics based models which consider the propagation of an existing crack along a single trajectory, the smeared crack model is a predictive crack initiation and distribution model. This property makes it suitable large-scale 3D computations.

5 3D FINITE ELEMENT MODEL OF THE CAVITY

The cavity structure was modelled with 354 solid quadratic elements. The number of nodes in the model was 2019. Reinforcing bars in the model were generated on individual bar basis in hoop, vertical and radial directions. The rod distribution was uniform throughout the model.
One half of the cavity structure was modeled and the symmetry conditions were imposed on the end faces of the model. The bottom of the model was fixed and the upper edge was restrained in horizontal plane. The cavity model is depicted in Figure 3 and 4. Figure 3 is hidden line and Figure 4 shaded plot of the Fem-model.

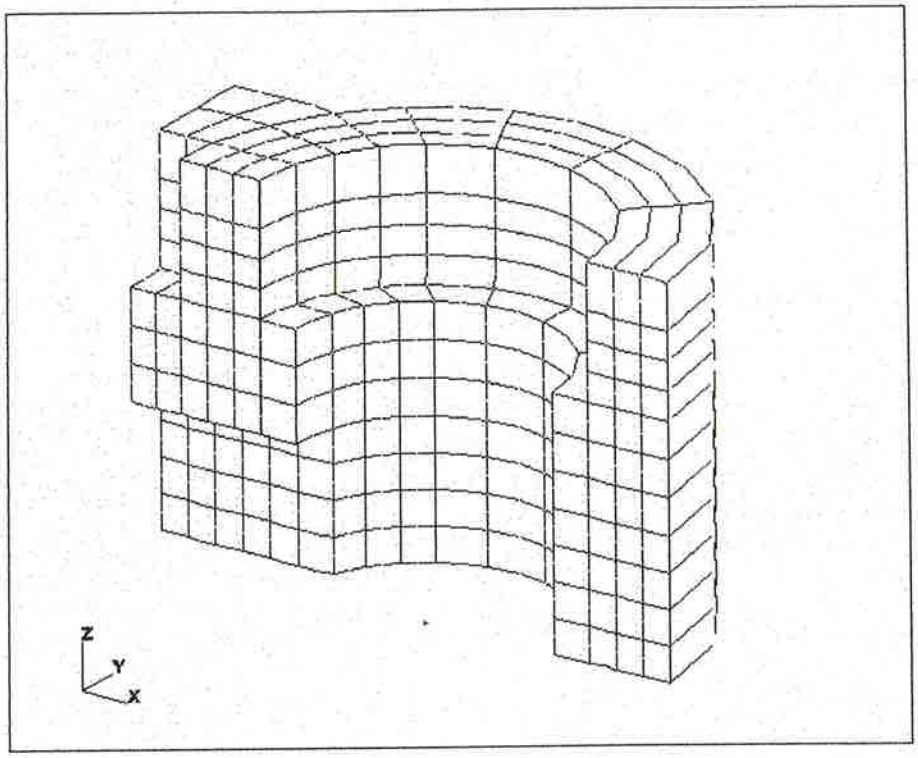


Figure 3 Hidden line plot of the reactor cavity model.

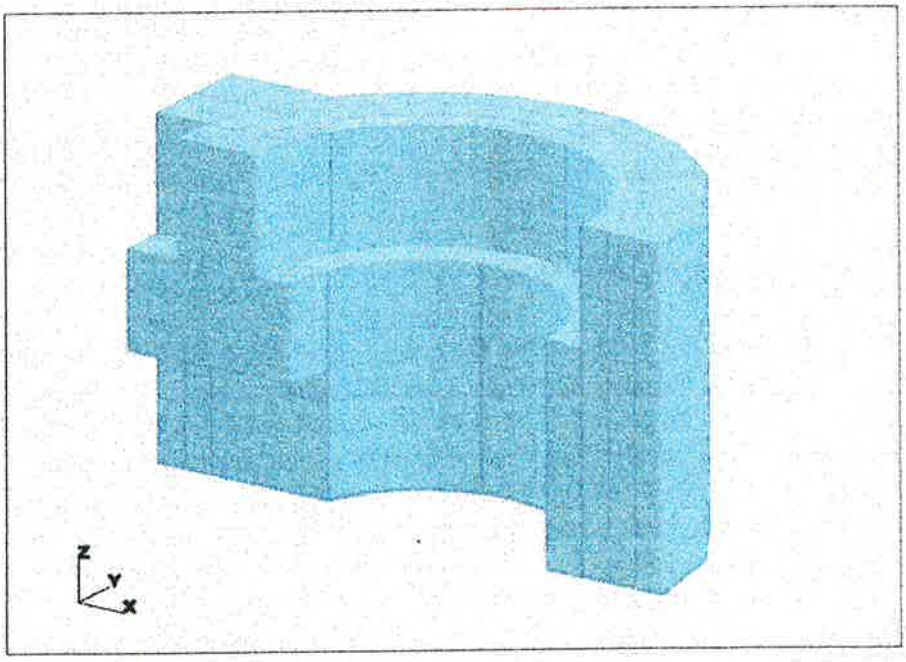


Figure 4 Smooth shaded plot of the reactor cavity model.

The spatial and temporal variation of the pressure load applied to the FEM model was devised in accordance with Figures 1 and 2. The vertical distribution of the load was assumed to be trapezoidal and the circumferential distribution was assumed to be sine half wave with the extent of half of the circumference. The shape of the load impulse in time was taken to be triangle with the impulse value of 150 kPa-s for the maximum load amplitude. The load characterization is given in Figure 5.

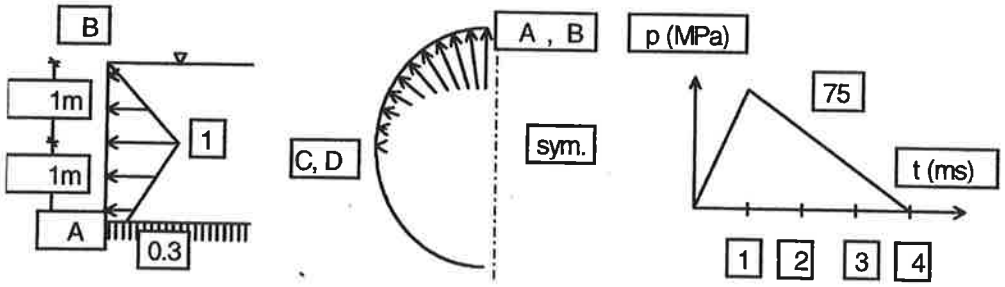


Figure 5 Spatial and temporal variation of the pressure load.

6 MATERIALS

The grade of K35-1 was selected according to the Finnish reinforced concrete code for the concrete material. The Young's modulus for concrete used in the analysis was E=29580 MPa, Poisson's ratio was 0.2 and the concrete density was 2400 kg/m3. The uniaxial, crushing strength was 24.5 MPa and the uniaxial, tensile, cracking strength was 2.14 MPa and the corresponding fracture tensile strain was 7.235E-5. The reinforcement bars of class A500 HW according to Finnish reinforced concrete code were selected. The Young's modulus for the reinforcement was 200 Gpa, the density was 7700 kg/m3. The Poisson's ratio for the reinforcement was 0.3 and the yield limit was 500 MPa. The ideal elasto-plastic behaviour was chosen for the reinforcement.

7 SOME FEATURES OF ANACAP PROGRAM [5],[6],[7],[8]

ANACAP is a structural analysis program developed for the 2D or 3D, static or dynamic response and failure analysis of reinforced or prestressed concrete structures. ANACAP is cast in implicit finite element methodology and is especially devoted for the highly non-linear material response involved with concrete due to cracking, creep, ageing or crushing. ANACAP can be applied over the whole range of analysis requirements from design acceptance criteria verification to construction optimisation to failure predictions. The capabilities in ANACAP have evolved from over twenty years of research and experimental verification by Anatech Research Corporation. Anatech Research Corporation is an internationally known leader in failure prediction, concrete modelling and analysis.

The following is the summary list of capabilities of ANACAP for modelling concrete:

- smeared-crack model for general 3D stress states
- history-dependent cracking
- crack closure and re-opening for cyclic loading
- pre-set crack direction for modelling construction joints
- rough-crack modelling for aggregate interlock effects
- compressive plasticity utilising the full stress-strain curve, including post ultimate strain softening and crushing
- hereditary creep relations with age and temperature dependence
- shrinkage
- material damping at locally damaged regions of cracking
- rebar plasticity with strain hardening
- rebar bond slip models for member ductility demand evaluation

Cracking

ANACAP employs the history-dependent smeared-crack model which predicts crack formation according to principal stress-principal strain interaction cracking criterion. Crack orientation, and hence stiffness anisotropy, is dictated by the stress and strain principal directions at each material integration point.

Cracks are allowed to form in three directions, and once a crack forms, it may close and re-open but can never heal. This crack memory feature is essential for analyses involving load reversals. The model includes algorithms for residual tension stiffness for the gradual transfer of load to the reinforcement during crack formation and for shear retention to simulate the effect of crack roughness through aggregate interlock. The model also allows definition of precracked directions for analysing structures with existing cracks.

Crushing

Plastic flow of material under compressive stresses is implemented through modified Drucker-Prager Yield condition. The compressive stress-strain curve is followed up to ultimate strength and into the strain softening regime where the material begins to unload due to internal damage and crushing. The model also includes hysteretic behaviour due to loading and unloading in the strain softening regime.

Temperature Dependence & Degradation

At elevated temperatures, usually encountered in nuclear power applications, concrete exhibits a significant departure from its elastic and creep properties at lower temperatures. This occurs because of thermally activated damage that is evidenced by the degradation of the material properties, especially the elastic modulus. This

material property degradation with time and temperature has been implemented for temperatures up to 450°F based on experimental data for the modulus, compressive strength and ultimate tensile strength of concrete. This feature is required for evaluation of structural integrity involving long-term thermal creep.

Damping

For dynamic applications, a concrete response model must include the effects of internal damping. ANACAP employs a cracking consistent damping model that introduces energy absorption directly at the element integration point level. This damping is treated as a function of time and the cracking orientation and status of the concrete material. This is done in a similar manner to the modelling of plasticity in metal structures, where hysteretic energy losses are modelled directly at the location of the affected material, rather than as a smeared effect uniformly distributed over the entire structure. This is a unique, specialised feature not found in any other commercially available Finite Element program.

Rebar-concrete Interaction

In areas of large stiffness discontinuities, major cracks develop and the interaction between the concrete and the reinforcement plays a major in determining the structural response and failure state. Because of dislocation displacements and rebar debonding, slippage can develop between the steel and concrete. The ANACAP concrete model has capabilities for modelling rebar bond slip based on confinement and anchorage utilising bond strength data from rebar pull tests.

8 NON-LINEAR DYNAMIC ANALYSIS FOR 200 KG/M3 HOOP AND VERTICAL REINFORCEMENT

In non-linear dynamic analysis the equations of motion of the system are solved by implicit direct integration method and displacements, velocities and accelerations as well as stresses and strains are calculated by the equilibrium iteration for every time step.
The results are shown at three faces of the Fem- model denoted by PL1, PL2 and PL3. PL1 is situated at the symmetry plane of the structure at the azimuth of positive x-axis, where the pressure load has its maximum value. This face is shown in Figures 3 and 4 at the right hand side of the plots. The faces PL2 and PL3 are situated at 90 degrees intervals from the face PL1 in counter clockwise direction. The results are shown in form of displacement time histories, cracking pattern charts and strain fringe plots.
The displacement time histories are presented in inner and outer face of the cavity wall in top, middle and bottom position in the vertical direction. In two faces situated at symmetry plane , namely PL1 and PL3, the displacements are given in x-direction and in face PL2 the displacements are given in global x- and y-directions.

In Figure 6, 7 and 8 the displacement histories of the cavity wall are given in the middle height of the wall and in the faces PL1 and PL2 and in the inner and outer surface of the wall.

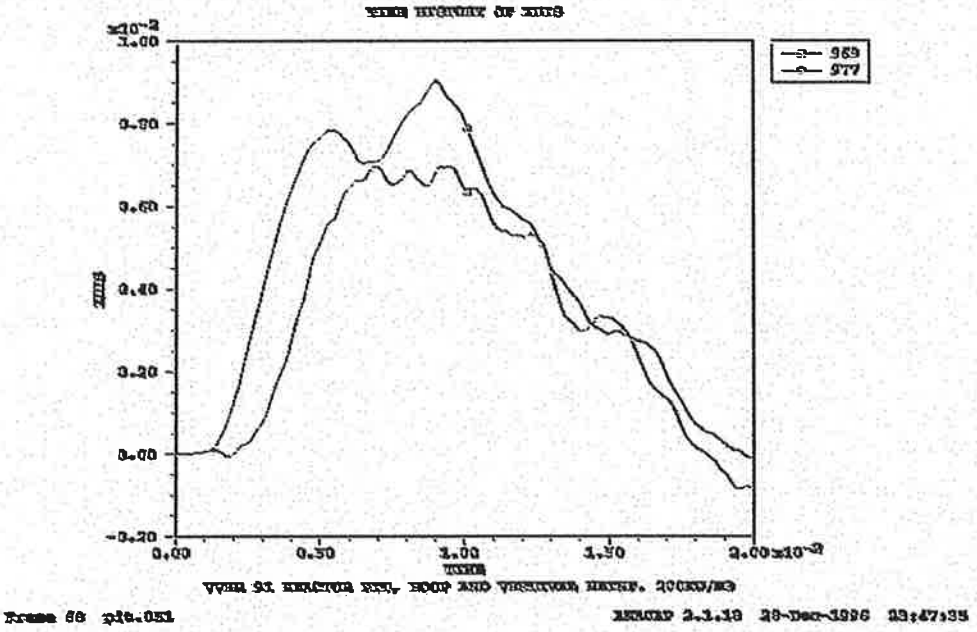


Figure 6 X-displacement time histories at middle height of face PL1 of the cavity wall

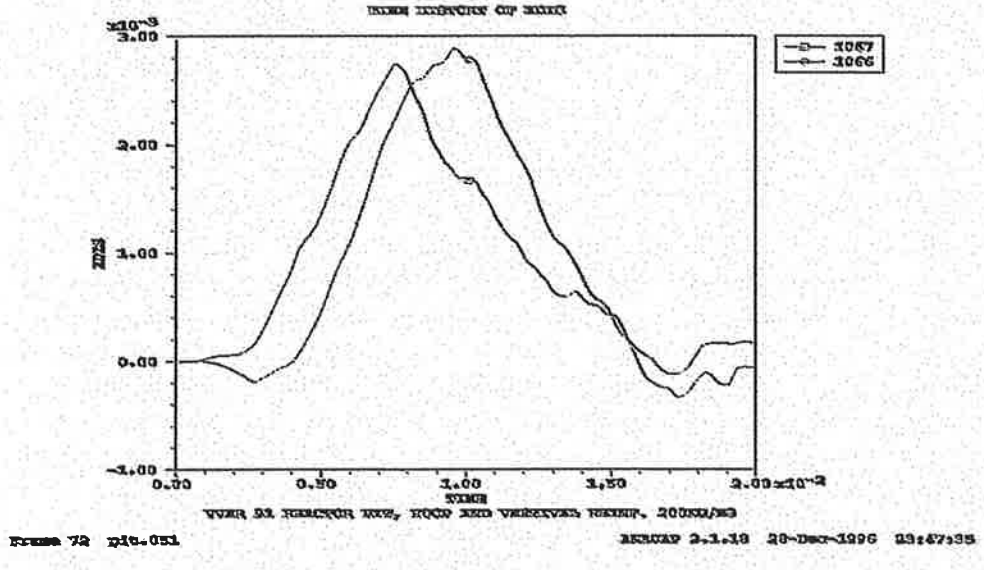


Figure 7 X-displacement time histories at middle height of face PL2 of the cavity wall

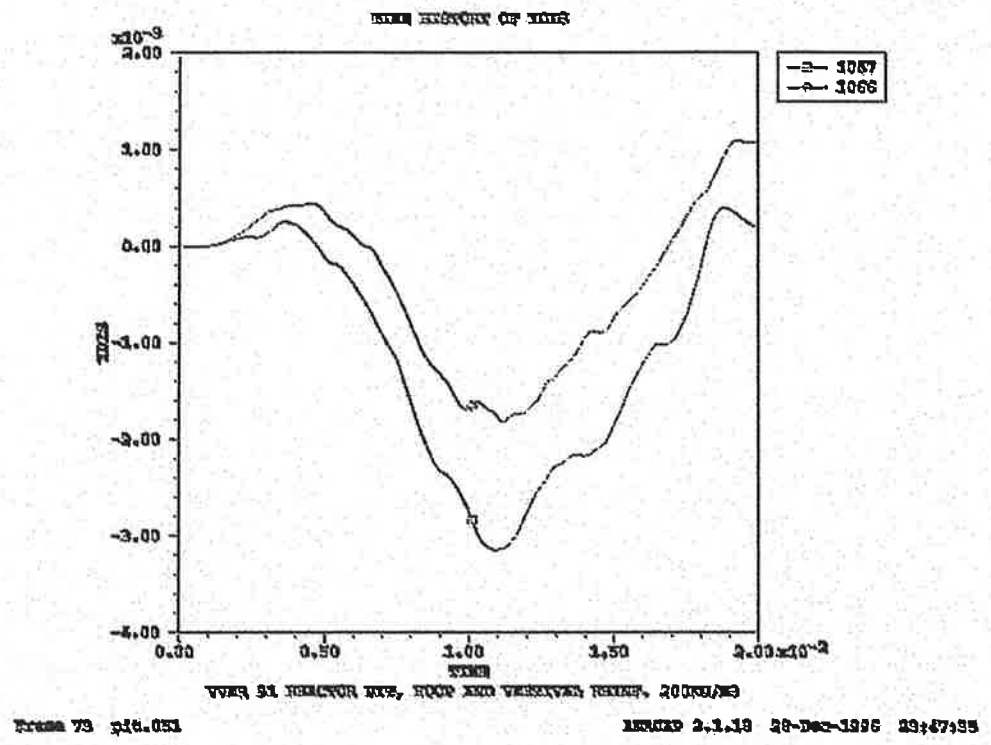


Figure 8 Y-displacement time histories at face PL2 at middle height of the cavity wall

In Figures 6,7 and 8 the displacements are given in meters and time in seconds. The duration of the time histories is 20 milliseconds which is five times the duration of the load pulse given in Figure 4. The curve depicted by square denotes the inner face displacement and the curve depicted by circle denotes the outer face displacement. The maximum value of the x-displacement in Figure 6 is about 1 cm and Figure 7 about 3 mm. The maximum value of the y-displacement in Figure 8 is about 3 mm. Its sign is minus and so the wall moves invards in face PL2.

A crack in a calculation point is marked by two concentric circles in the crack plane an when two concentric circles are seen in the picture it means that the cracking plane is approximately same as the picture plain or xz plane, that is to say that the concrete has cracked in the hoop direction. When the cracks appear as ellipses or even just lines it means that the planes of cracks deviate from the picture plain. The lines mean cracking in a plane perpendicular to the picture plain. When concrete has cracked in two or three directions in a calculation point the graphical image for each crack is that described above for one crack .

A cross in a calculation point mean that concrete crushes in compression. In cracking pattern charts this is seen near that part of the surface where the pressure load has its maximum value. The cracking and crushing pattern of the wall face PL1 is shown in Figures 9, 10 and 11 at time points 10, 40 and 120 milliseconds.

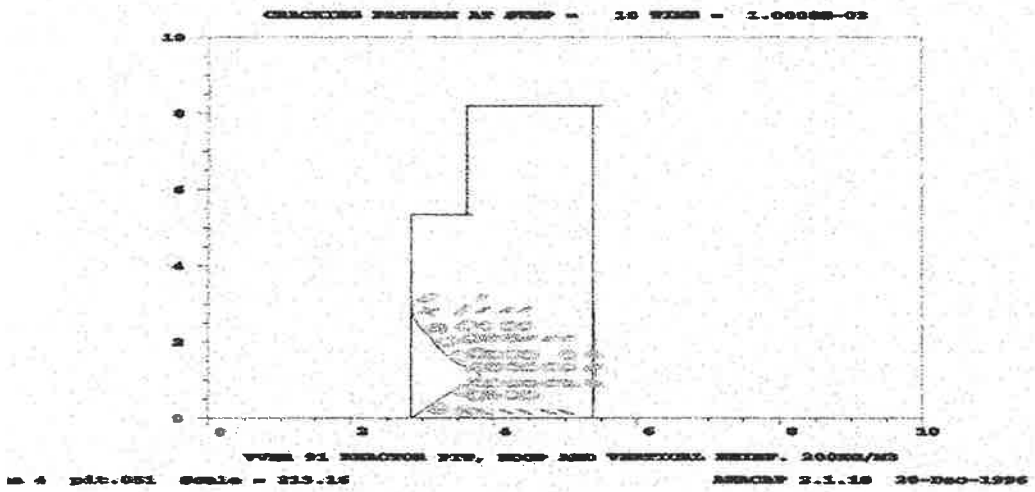


Figure 9 The cracking pattern at PL1 at time 1 millisecond from load application

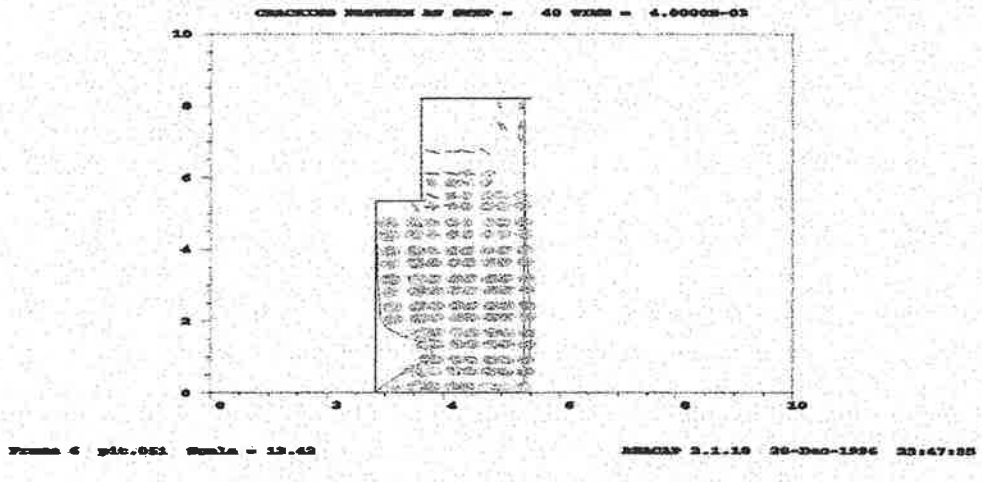


Figure 10 Cracking pattern at face PL1 at time 4 milliseconds from load application

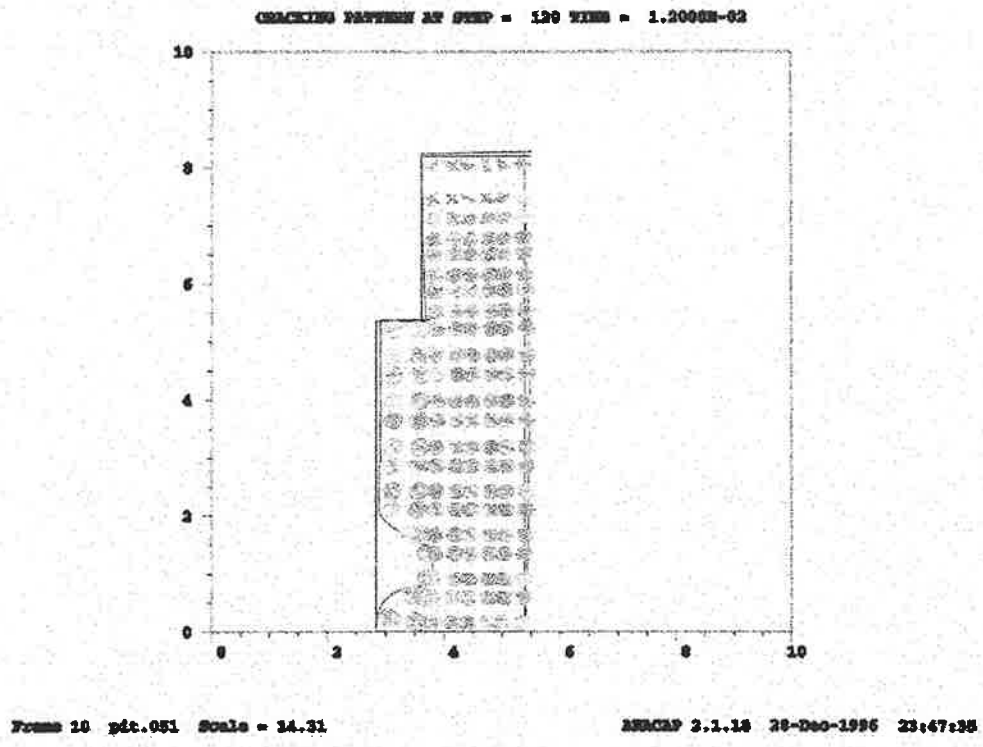


Figure 11 Cracking pattern at face PL1 12 milliseconds from load application

The fringe plots for deformations are shown in Figures 12-14. Strain values vary between 11000 microstrains of compression and 17000 microstrains of tension. Some of the peak values of the fringe plots are caused by fixed boundary condition at the base of cavity wall. Actual boundary condition is more flexible because cavity is supported by concrete slab.

9 DISCUSSION

All the results shown in Figures 6-14 were calculated for the 200 kg/m3 hoop and vertical reinforcement. The same calculations were also carried out for the 300 kg/m3 hoop and vertical reinforcement. Radial reinforcement amount was kept constant in both runs and the amount was 50 kg/m3.
ANACAP does not plot the reinforcement stress histories but the check from print file gave the result that reinforcement remains in elastic range except local concentrated areas near the base of the cavity wall where fixed boundary condition was applied.
The cracking procedure in the case of 300 kg/m3 hoop and vertical reinforcement is almost the same as in the 200 kg/m3 case. Stresses and strains in reinforcing bars are again in elastic region and are somewhat smaller than those in the 200 kg/m3 case. So the crack widths are also smaller.

There is again some compressive crushing near the maximum pressure amplitude and near the base slab. The latter is again explained by a conservative boundary conditions at base slab which do not take into account its deformation.The strain fringe plot shows that strains are somewhat smaller than in the 200 kg/m3 case which was to be expected because of increased reinforcement amount.

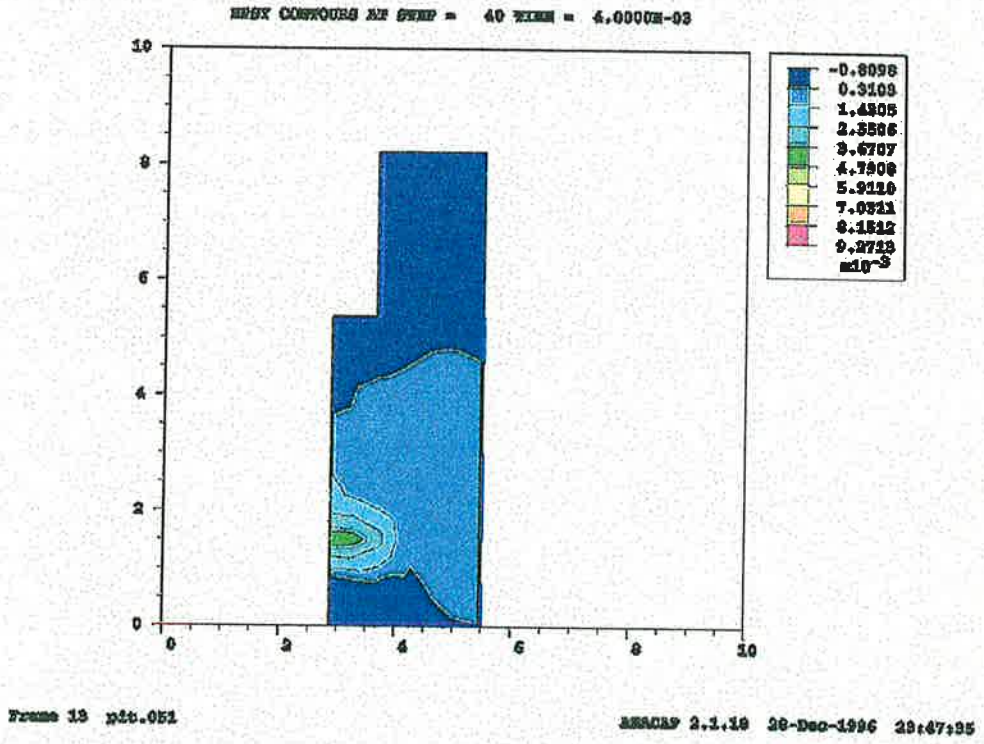


Figure 12 Fringe plot of hoop deformation at the end of load pulse at face PL1.

10 CONCLUSION

The performed analysis demonstrated that the off-center pours of molten core to the cavity can be contained by the cavity walls without impairing their capacity to carry vertical load from vessel weight.
There exists local concrete crushing at the cavity wall adjacent to pour but its extent is limited and its penetration insignificant compared to 2.6 m wall thickness.
Reinforcement yield is very limited and is caused by conservative restraints at wall bottom.
Concrete cracking is spreaded throughout the whole wall section at maximum load amplitude location but because of moderate strains the crack widths remain small.
The obtained results also confirmed the suitability of ANACAP to advanced 3D analyses of reinforced concrete. Similar problems as studied in this report have been studied in references [9],[10],[11] and [12].

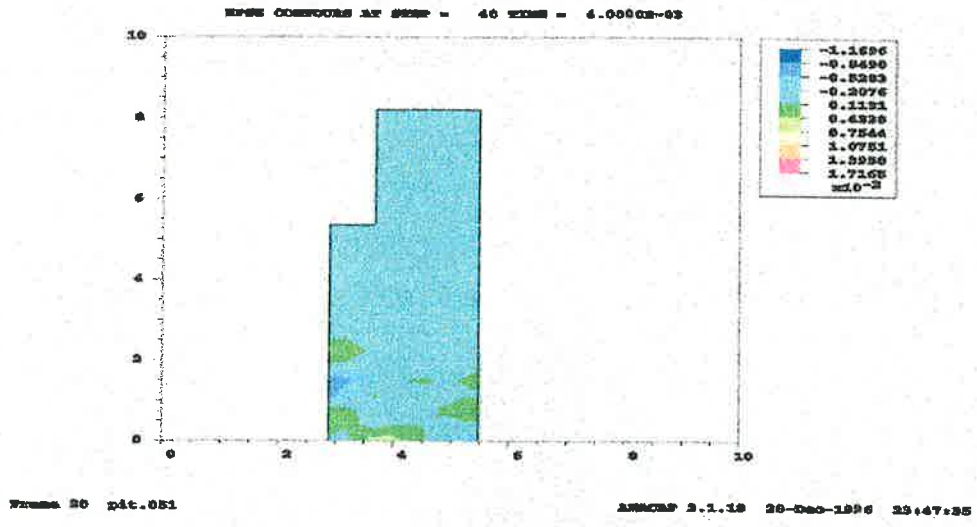


Figure 13 Vertical strains at face PL1 at the end of the load pulse

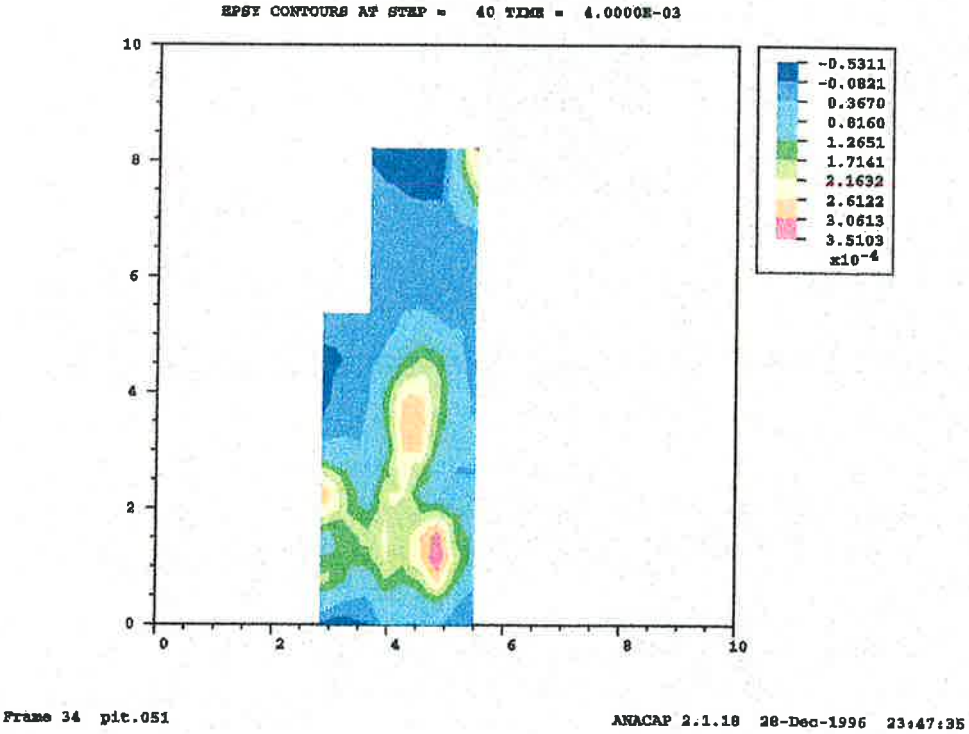


Figure 14 Hoop strains at face PL2 at the end of the load pulse

REFERENCES

1. Varpasuo, P. 1995. The strength of the reactor cavity of VVER-1000 NPP against steam explosion.
Trans. 13th SMIRT:vol IV, pp. 167-173, Editora da Universidade, Universidade do Rio Grande do Sul.
2.Theofanous, T.G. 1994. Consideration of severe accidents in the containment design of the VVER-1000.
TcI-94/02, Theofanous & Co., Inc.
3. Theofanous, T.G., Yuen W.W.1995. Steam Explosions in the VVER-91 Reactor Cavity Equipped with a Core Catcher.
TCI-95/02, Theofanous & Co., Inc.
4. Rashid, Y. R. 1968. Ultimate strength analysis of prestressed concrete pressure vessels.
Nuclear engineering and design ,vol. 7.
5. Minutes of the twenty-fifth Explosive Safety Seminar, Vol IV, Anaheim Hilton, Anaheim, California, August 1992, Department of Defense Explosives Safety Board.
6. ANACAP, ANATECH Concrete Analysis Program, Advanced Finite Element Modelling for the Thermal and Stress Evaluation of Concrete and Steel Structures, ANATECH Research Corp.
7. ANACAP-U, ANATECH Concrete Analysis Package, Version 92-22, Theory Manual, ANATECH Research Corp., November 1992.
8. ANACAP, ANATECH Concrete Analysis Program, User's Guide, Version 2.1.15, ANATECH Research Corp, San Diego, September, 1995.
9. Almström H. et al., 1997. Significance of fluid-structure interaction phenomena for containment response to ex-vessel steam explosions, OECD/CSNI Specialist Meeting on Fuel-Coolant Interactions, JAERI, Tokyo.
10. Engelbrektson A. et al., 1996. An assessment of containment structural response to ex-vessel steam explosion loads. International Conference on Nuclear Containment University of Cambridge, Cambridge, England.
11. Toader A. B. et al., 1997. Constitutive Model for 3D Cyclic Analysis of Concrete Structures, ASCE Journal of Engineering Mechanics, February 1997.
12. Zuchuat O. et al., 1997. Steam explosion-induced containment failure studies for Swiss Nuclear Power Plants. OECD/CSNI Specialist Meeting on Fuel-Coolant Interactions, JAERI, Tokyo.

ANALYSIS OF CFST MEMBERS BY LBE METHOD

Matti V. LESKELÄ
Structural Engineering Laboratory
University of Oulu
P.O.BOX 191, FIN 90101 OULU, FINLAND
E-mail Matti.Leskela@oulu.fi or Matti.V..Leskela@mvles.pp.fi

ABSTRACT

Concrete filled steel tubes (= CFST) are employed as columns in steel and composite frames and also as compressed members in pile foundations, and they are loaded mainly axially, but some bending may also be introduced by transverse loads and due to eccentricity in the axial load. Bond stresses are introduced to the steel-concrete interface by the flexural load effects, which may also cause cracking in concrete, making the behaviour non-linear. Therefore, the possibilities to apply the LBE method (= layered beam elements) for the CFST problems are discussed and a system of two parallel element layers is introduced. Of special interest is the problem of the load introduction from the steel cover to the concrete core, as there are requirements for this case in ENV-Eurocode 4. Principal modes of behaviour of composite columns and bent CFST are enlighted by the results of three numerical examples.

1. INTRODUCTION

Concrete filled steel tubes are one form of composite columns covered in ENV Eurocode 4 [1]. The cross-sections of the columns in the scope of the code are such that both of the material components shall have a symmetric section and the centroids should coincide. In order to behave compositely, both materials of the column shall be strained, which is normally true after some introduction length also for the loads introduced from the steel cover. No parametric evaluation of the load introduction is given in Eurocode 4, and it is only stated that the loads introduced to the column should be transferred to act on the whole cross-section within a length not longer than twice the diameter of the cross-section. While the real transfer length will depend on the properties of the shear interface, there should also be reasonable tools for making parametric studies of the real behaviour.

The method of layered beam elements (LBE) can be suited for the problem described, and this paper introduces the setup of the method for discretizing the CFST members into two layers of elements, (1) the steel tube surrounding the concrete core and (2) the concrete core with its reinforcement (Fig. 1). In the initial state of the composite member the concrete core is uncracked and the centroids of the components of the cross-section coincide. For large eccentricities of the load, the concrete is liable to crack and higher interface bond stresses are required to maintain the composite behaviour.

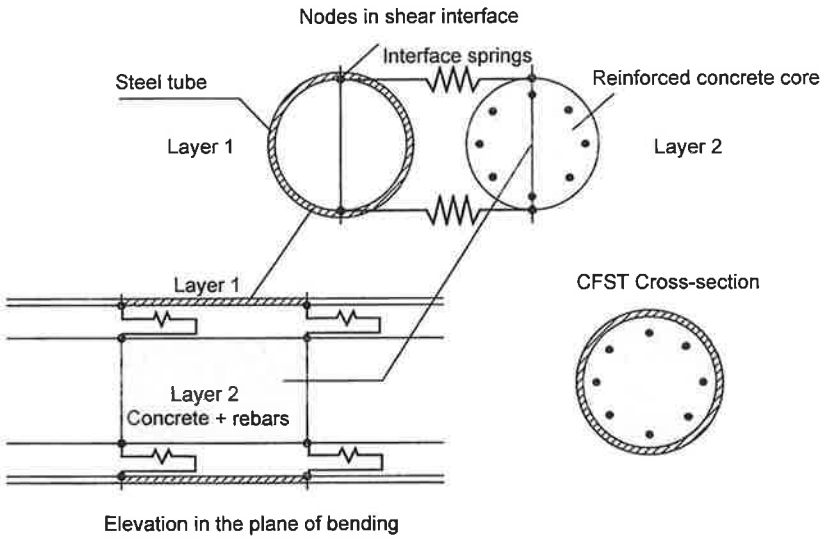


FIGURE 1: A typical CFST and its discretization into LB-elements

2. LB ELEMENTS

The LB elements are a transformation of the normal two-noded beam element in which the degrees of freedom included in a single node are separated and transformed so as to make them suitable for connecting the similar degrees of freedom in various elements together by interface springs. The structure of the element stiffness matrix is discussed e.g. in [2] and it should be noted that the essential information required for the element stiffness matrix is quite limited and similar with that included in the source element, i.e. axial stiffness (EA), flexural stiffness (EI) and the location of the neutral axis for bending (Fig. 2). This makes them practical for well-controlled parametric studies frequently required for composite structures. While in many applications the properties of the interface spring elements can directly be derived from simple tests which represent the properties of the shear connection, this cannot be done in the case considered here, but the average bond stress - slip relationship of the shear interface can be interpreted into spring properties by numerical integration of the average behaviour,

however. The average properties of the shear interface are based on simple push-out tests (Fig. 3), employed normally for composite structures, e.g. in Eurocode 4 [1].

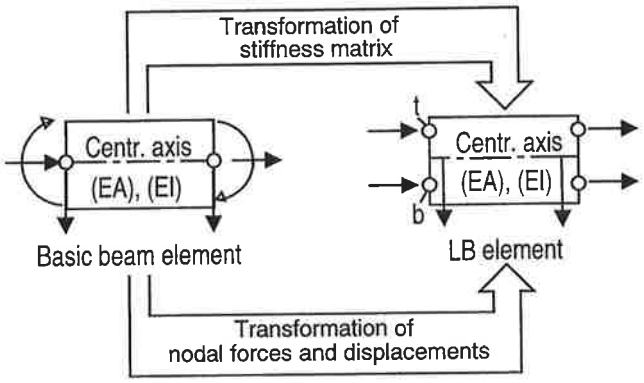


FIGURE 2: Transformations required for the source element

2.1 Derivation of the load-slip relationship for the interface springs

The average load-slip or bond stress-slip relationship for the steel-concrete shear interface of the composite tube is evaluated from simple push-out tests, the principle for which is shown in Fig. 3 below. τ_b is the mean interface shear stress due to applied load and δ_s is the slip of the concrete core with respect to the steel tube. When mechanical connectors are not involved, the behaviour is highly non-linear, resulting in softening curves, in which the slopes are positive for all δ_s, and the bond strengths, τ_u, applicable for the design purposes depend on the maximum slip that can be allowed without impairing the system behaviour.

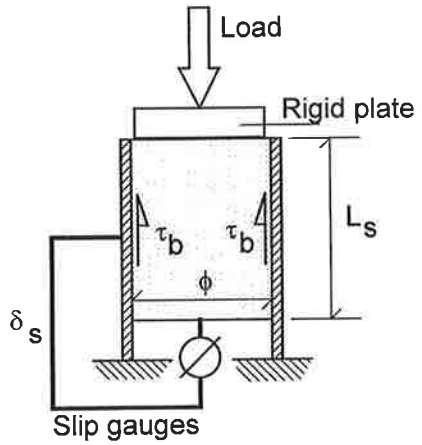


FIGURE 3: Push-out test setup for defining the average τ_b - δ_s relationship

Denoting the total load in the push-out test by F, the average bond stress is simply $\tau_b = F/\pi\phi$. It is now assumed that strength τ_u is obtained at a slip of δ_{s1}, which then fixes a unit function, $f(\delta_s)$, derived from the experimental τ_b - δ_s relationship so that $f(\delta_{s1}) = 1$. Then $\tau_b = \tau_u f(\delta_s)$ and the relationship for the interface shear flow and slip, v_l and $\delta_{s.i}$, respectively is implicitly obtained from:

$$v_l = \phi\tau_u \int_0^{\pi/2} \sin\varphi\, f(\delta_{s.i}\sin\varphi)\, d\varphi \qquad (1)$$

In practice the values of v_l for slips $\delta_{s.i}$ are calculated from summation:

$$v_l = 2\sum_{k=1}^{n} \tau_u f_k \sin\varphi_k\, r\frac{\pi}{2n} = \frac{\pi\phi\tau_u}{2n}\sum_{k=1}^{n} f_k \sin\varphi_k \qquad (2)$$

in which $f_k = f(\delta_k)$, $\delta_k = \delta_{s.i}\sin\varphi_k$ and $\varphi_k = (k - 0{,}5)\pi/(2n)$ with k = 1, 2, ..., n.

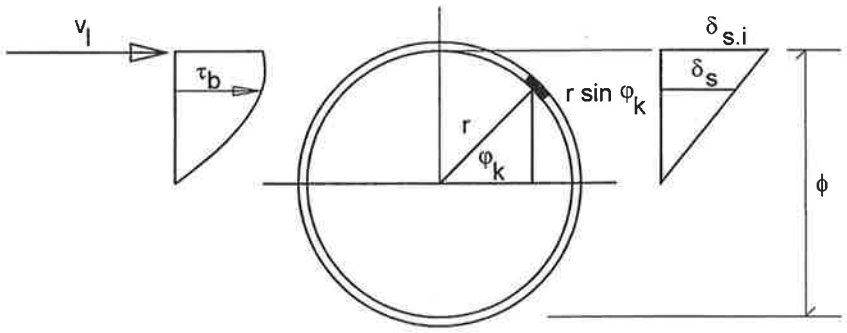


FIGURE 4: Geometry for the derivation of v_l - $\delta_{s.i}$ relationship

Figure 5 shows an example of the shear flow - slip relationship derived for a welded tube with a diameter of 200 mm. It is pointed out that the relationship is not general and the properties of the shear interface should always be considered as based on tests. A comparison of the curve in Fig. 5 with function $f(\delta_s)$ would reveal that there are only secondary differences in the envelopes of the curves, when the curves are scaled into similar coordinates and a good approximation for the values of v_l would be to assume the average bond stress from the push-out test to act on one third of the sphere of the tube.

The relationship in Fig. 5 is applied hereafter in the numerical examples prepared for studying some typical load cases for composite columns loaded primarily axially (concentric and eccentric axial loadings).

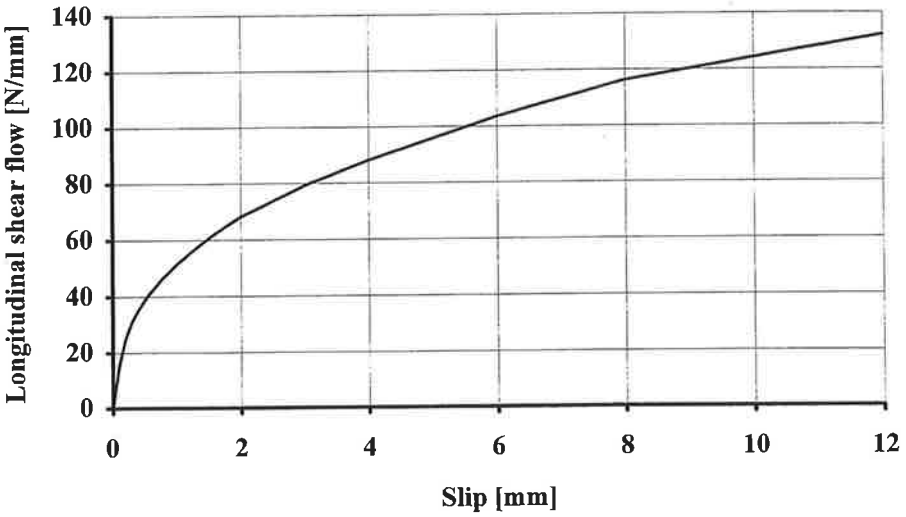


FIGURE 5: An examle of v_l - $\delta_{s.i}$ relationship for a tube of 200 mm

2. NUMERICAL EXAMPLES

A two-storey column shown in Fig. 6 below is analyzed in two configurations of loads F_1 and F_2 so as to demonstrate the behaviour of an intermediate column (case a, F_1, F_2 > 0, balanced loading) and an end column (case b, F_1 > 0, F_2 = 0) in a frame, in which the loads of the flooring are transferred to the columns by the fin plates of the beam-column connection. The location of the reactions, e, is assumed to be 50 mm from the outer face of the tube.

2.1 Intermediate column, case a

The column is loaded axially and the loads, $F_1 + F_2$, ($F_1 = F_2$) are increased in steps of 100 kN up to 1000 kN. Two options are considered, (1) a column without mechanical shear connectors and (2) a column provided with stud shear connectors [5] in the section of the load introduction (middle-height of the column). The strength of the connectors is assumed to be twenty times of that of the bond connection within the length of the elements. The differences in the axial load distribution between the options are seen in Fig. 7, in which the diagrams with markers represent the thrust or pull in the steel member of the column. In the second storey of the column the steel tube is in tension and the concrete core in compression, and the absolute values of the axial forces are identical. The difference between the options is clearly visible and the load transferred by the connectors is approximately 180 kN (= the difference in the axial force of the concrete core when passing the mid-height of the column).

It is pointed out that for the load of 1000 kN the column behaves totally elastically and the stresses are well below the point of yielding. This should also be seen when evaluating the design axial resistance of the column according to Eurocode 4 [1]. The calculation was terminated only as the design loads transferred normally from the flooring to the columns are in the range of 1000 kN.

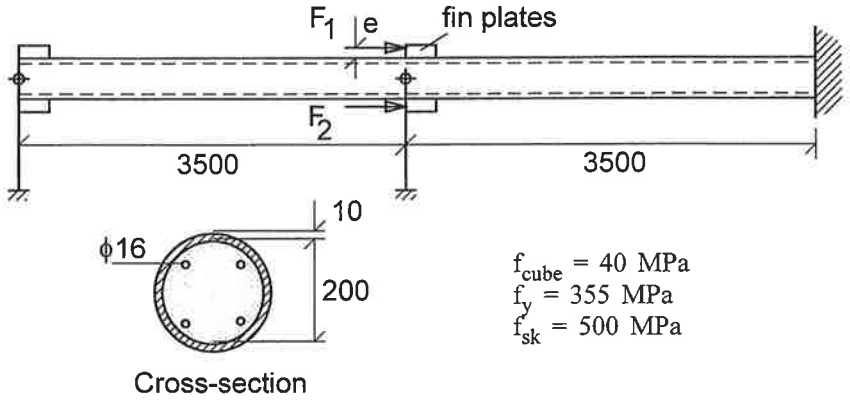


FIGURE 6: Column configuration, a two-storey CFST

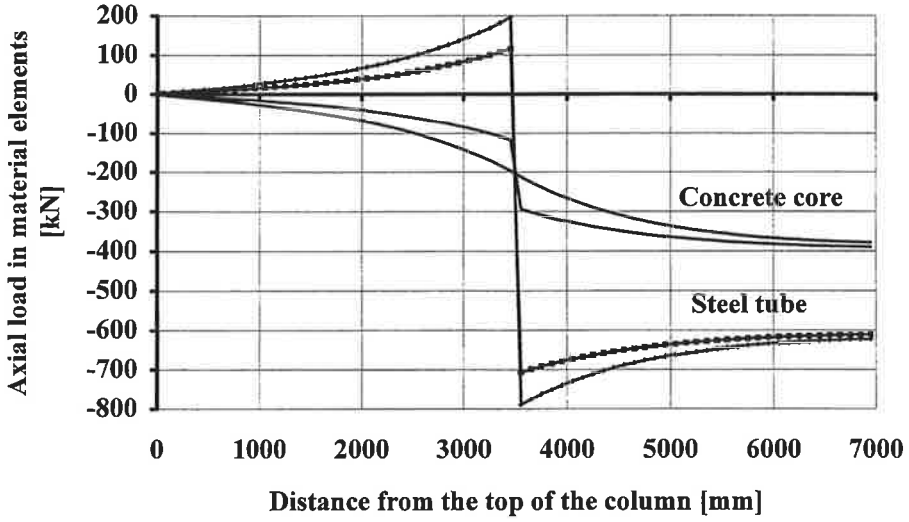


FIGURE 7: Axial load distributions for $F_1 + F_2 = 1000$ kN

2.2 End column, case b

The column is loaded by an eccentric load, F_1, on a fin plate in steps of 100 kN up to 1000 kN. Similar options as in case a are considered. As compared to the previous case, load 1000 kN strains the column also by bending, making thus the maximum strains higher. Maximum deflections of 5,6 mm and 4,6 mm in the opposite directions are observed in the second and first storey, respectively, but the strains are mostly below the limit of yielding and local plasticity starts only developing in the region of the load introduction. There are only minor differences in the deflections between options (1) and (2), but while the purpose of the shear connectors is to prevent slipping in the section of the load introduction, the yielding of the tube is slightly delayed in option (2). The distribution of slip deformations at the connection interfaces is shown in Fig. 8 for load F_1 = 1000 kN. The interface on the side of the load is shown without markers.

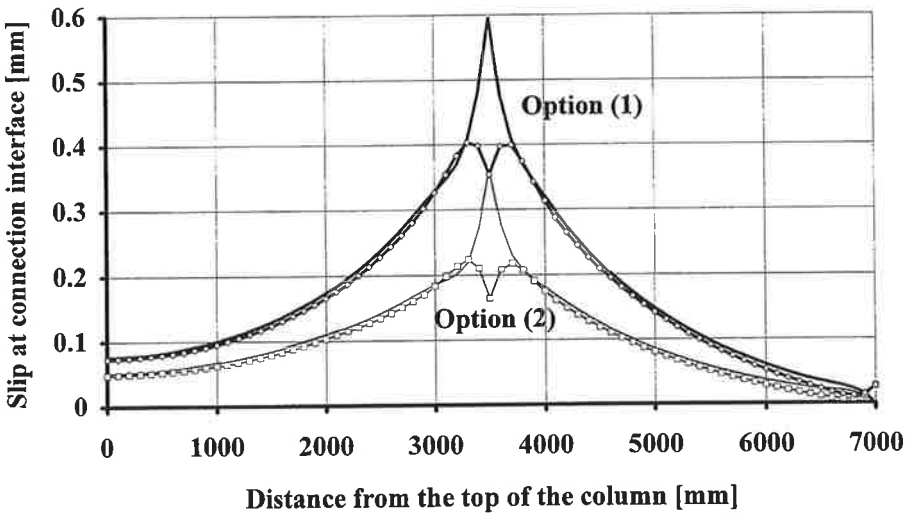


FIGURE 8: Slip deformations for the load of 1000 kN in case b

Although the form of the curves in Fig. 8 is similar for options (1) and (2), the slips are smaller when mechanical connectors are involved.

Bending moments in material components (curves with markers), as well as the sum of the component moments are shown in Fig. 9 for F_1 = 1000 kN. Second order effects were considered in the calculation, i.e. the incremental moments due to deflection are included in the diagrams. The sum of the component moments represents approximately the external moment due to load effects, and even exactly, as far as the centroids of the effective sections coincide. This is best true for the second storey (distance 0 - 3500 mm from the top), but the deviation of the sum from the external moment is less than 2 % even in the first storey (distance 3500 - 7000 mm from the top).

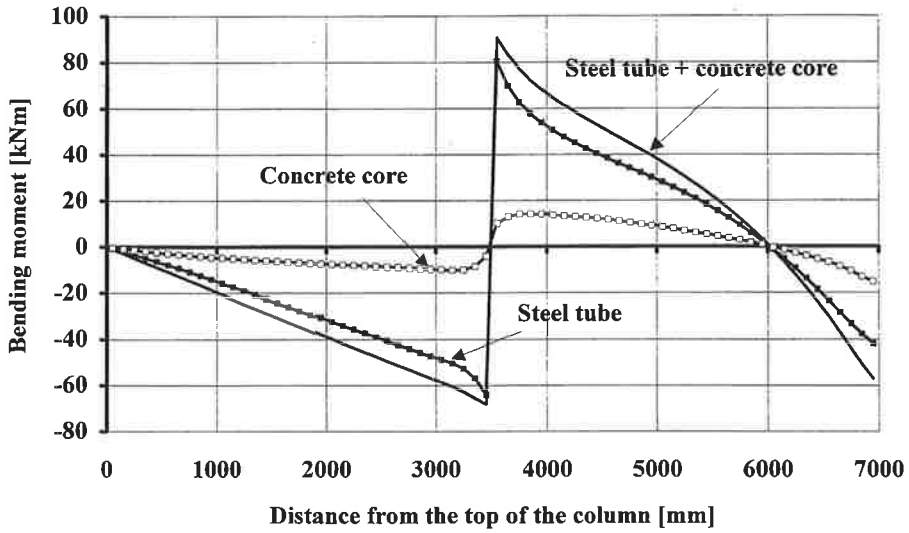


FIGURE 9: Bending moment diagrams for F_1 = 1000 kN in case b

2.3 CFST in pure bending

A CFST loaded by a concentrated central load F in bending as a simply supported member is considered next. The principal dimensions of the structure are:

- outer diameter of the tube = 508 mm
- wall thickness of the tube = 12,5 mm
- reinforcing bars 10 ϕ32 with an axis distance of 80 mm from the steel-concrete interface
- total length of the tube = 4400 mm
- span length as simply supported member = 3800 mm, i.e. there are 300 mm overhangs at both ends.

The material parameters assumed for the calculation are:

- concrete grading, f_{cube} = 42 MPa
- yield strength of the tube, f_y = 355 MPa
- yield strength of the reinforcing bars, f_{sk} = 500 MPa
- Young's modulus for the steel materials, E_a = 190 000 MPa
- the bond stress - slip relationship obtained for 200 mm tubes is applied here, although it may not be exactly valid for larger diameter tubes.

An ideal elastic-plastic behaviour with hardening initiating after a strain of $11f_y/E_a$ is assumed for the tube and reinforcement.

The member was analyzed as discretized into elements having length of 100 mm. The load was increased in steps of 100 kN up to 2000 kN, where the calculation was terminated due to excessive deformations. First yielding at the bottom fibre of the tube is detected at F = 1400 kN, and for increased loading the yielding spreads towards the supports both at the top and bottom fibres of the tube. The maximum deflection at the termination of the loading is in excess of 90 mm.

Figures 10 and 11 introduce some results of the calculation:

- The load-deflection response is shown in Fig. 10.
- The flexural stiffness of the material components for loads F = 1000, 1500 and 2000 kN is shown in Fig. 11.

The values for the elastic bending stiffnesses evaluated by integration by the calculation program are:

- steel tube $(EI)_a$ = 113,4 MNm^2
- concrete core + reinforcement, uncracked, $(EI)_c$ = 116,3 MNm^2

The standard formula for the elastic second moment of area of the circular tube yields $(EI)_a$ = 113,5 MNm^2 for E_a = 190000 MPa, and the respective uncracked stiffness of the concrete core is $(EI)_c$ = 110 ... 125 MNm^2, depending on the assumed location of the reinforcing bars and grading of the concrete. For F > 200 kN, the concrete starts cracking and its stiffness is highly reduced.

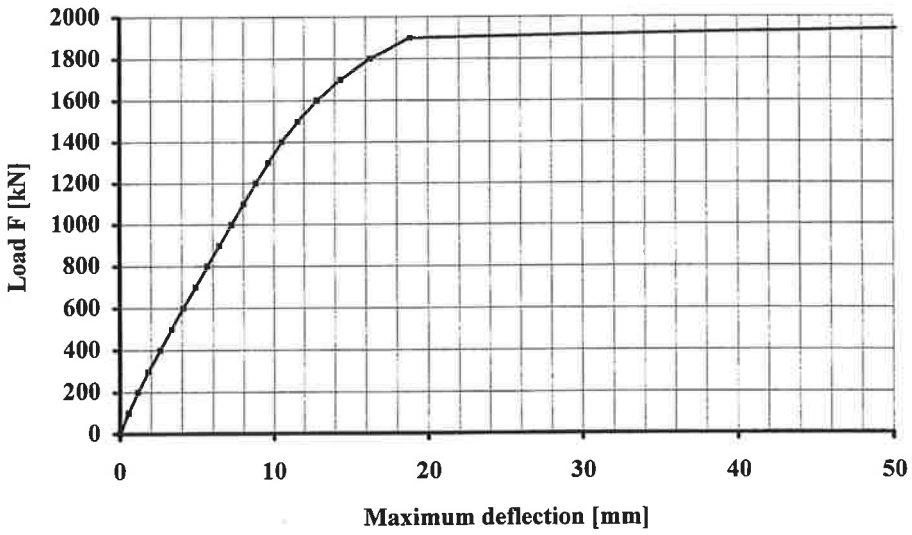


FIGURE 10: Load-deflection response

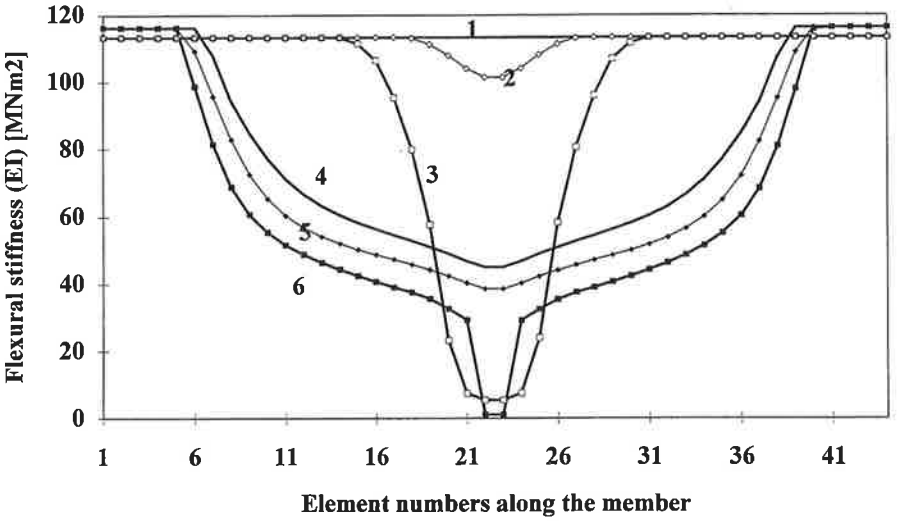


FIGURE 11:
Flexural stiffnesses in elements of the tube and concrete core for three load levels F = 1000 kN, 1500 kN and 2000 kN. Diagrams 1, 2 and 3 represent the flexural stiffness of the tube and diagrams 4, 5 and 6 the stiffness of the concrete core and its reinforcement, respectively.

The bending resistance, $M_{pl.R}$, as evaluated according to section C.6.4 of Eurocode 4 [1] is approximately 1521 kNm, corresponding then to the ultimate value of the concentrated load, $F_{ult} = 4M_{pl.R}/L = 1521 \times 4/3{,}8 = 1600$ kN. A far higher load could be applied in the calculation, as the onset of yielding for the tube is $F_{el} = 1400$ kN, and it is seen in Fig. 11 that for F = 1500 kN, the bending stiffness of the tube is not much reduced. This implies that the method given in Eurocode 4 is highly conservative at least for circular tubes.

3. SUMMARY AND CONCLUSIONS

The method of layered beam elements was here applied for the composite tubes which may efficiently be discretized into two layers of parallel elements. The axial modes as well as the bending modes of the loading can be considered effectively. While the elements were originally developed for various flexural systems of composite structures [e.g. 3, 4], it was shown in the examples that the method is also well applicable for the axial loadings and mixed modes consisting of axial and flexural loads.

The results of the examples for the two-storey column indicate that the load transfer from the tube to the composite cross-section happens both above and below the level

where the load is introduced to the column. Figure 7 would also indicate that in the case of fire design it is justified to 'hang' the loads coming from the flooring above the hot compartment by the steel tube and transfer them to the concrete core in the cool room above the fire. This kind of load transfer can only be considered in the composite columns.

The method can efficiently be programmed to give a good view for the various important parameters of the behaviour in composite structures. As applied for the purpose of the study of composite columns, the standard output for every loadstep include:

- nodal displacements and interface forces,
- axial forces and bending moments for the elements,
- axial and flexural stiffnesses for the elements.

All of these are frequently required for every reasonable parametric study of various composite structures.

4. REFERENCES

[1] Eurocode 4: Design of composite steel and concrete structures, Part 1-1: General rules and rules for buildings. ENV 1994-1-1:1992, CEN 1992.

[2] Leskelä, M.V., A finite beam element for layered structures and its use when analysing steel-concrete composite flexural members. In *Constructional Steel Design. World Developments.* Elsevier Applied Science, London and New York, 1992, 354-358.

[3] Leskelä, M.V., LBE-Modelling Applied to Reinforced Concrete T-Section - Steel U-Section Composite Beam. In *Proc. Estonian Acad. Sci. Engin.*, 1996, **2**, 2, 151-165.

[4] Leskelä, M.V., Shear flow calculations for slim-type composite beams supporting hollow-core slabs. In *Steel-Concrete Composite Structures. Proc. 4th ASCCS Int. Conf.* (Javor, T. ed.). Expertcentrum Bratislava, Slovakia, 1994, 299-302.

[5] Leskelä, M.V., Non-Headed Studs as Shear Connectors for CFST Columns. Papers for *Engineering Foundation Conference Composite Construction III*, 9-14 June 1996, Irsee, Germany. Thurday Evening Sessions, 117-127.

SOME PROBLEMS IN NUMERICAL POST-BIFURCATION ANALYSIS

Reijo KOUHIA and Martti MIKKOLA

Laboratory of Structural Mechanics

Helsinki University of Technology

P.O. Box 2100, 02015 HUT, Finland

ABSTRACT

A critical review of the existing methodology for the numerical treatment of post-bifurcation branches at a multiple bifurcation point is presented. In the critique the main emphasis is given to robustness, but also implementational issues and computational requirements as well unsolved problems are addressed.

1 INTRODUCTION

Path following is the most common procedure to analyze the stability behaviour of complex structures. In these methods a one dimensional equilibrium curve is traced. Difficulties are expected to exist with the basic continuation algorithms near critical points where the tangent stiffness matrix is ill-conditioned. If the multiplicity of the critical point is one, the numerical treatment of post-bifurcation paths is rather straightforward, but the situation changes dramatically when the multiplicity of the critical point is greater than one. Occurrencies of multiple bifurcation points in a numerically traced equilibrium path are quite rare. However, circumstances where the buckling loads are clustered in a small interval just above the smallest critical load are frequent in the stability analysis of various shell structures. Such an occurrence might lead to an exceedingly complicated situation, where the buckling modes interact with each other. Therefore it seems natural to require a branching procedure to be able to handle multiple bifurcation points (or near ones) in a continuation algorithm.

In the scientific literature there are only few papers concerning the numerical treatment of multiple bifurcation. Among them belong the works of Kearfott [1], Allgower and Chien [2] and Huitfeldt [3]. In this article those methods are briefly reviewed and the existing difficulties with these approaches are pointed out.

Also an approach which uses the Koiter type reduction technique is presented and its properties are compared to the above mentioned techniques.

Only the question of the computation of post-bifurcation branches is discussed in the present paper. Relevant related problems like location and computation of the critical point itself have been left out.

2 BASIC DEFINITIONS

Discretized form of static equilibrium equations under single load control can be written in the form

$$\boldsymbol{f}(\boldsymbol{q},\lambda) = \lambda \boldsymbol{p}_r - \boldsymbol{r}(\boldsymbol{q}) = \boldsymbol{0} \tag{1}$$

where $\boldsymbol{p}_r$ and $\boldsymbol{r}$ are the vectors of external and internal forces.

It is assumed that the magnitude of the loading is controlled by a single parameter λ, called the load parameter. The N-dimensional state variable vector is denoted as $\mathbf{q}$ and in most applications N is large. In the solution of the non-linear equilibrium equation (1) the tangent stiffness matrix $\boldsymbol{K}$ is usually needed (at step k)

$$\boldsymbol{K}_k = -\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{q}}\Big|_{\boldsymbol{q}_k}.$$

In order to parametrize the path a length measure, i.e the path parameter s is defined as $\Delta s = \sum_{i=1}^{k} \sqrt{\boldsymbol{t}^T \boldsymbol{C} \boldsymbol{t}}$, where $\boldsymbol{t}^T = \left[\Delta \boldsymbol{q}^T \Delta\lambda\right]$ and $\boldsymbol{C}$ is a weighting matrix, see ref, [4].

It is now assumed that the critical point $(\boldsymbol{q}_{cr}, \lambda_{cr})$ is reached and located for prescribed accuracy between steps $k-1$ and k. Notification of new critical modes during the continuation can be obtained by monitoring the inertia of the tangent stiffness matrix. If the number of unstable modes associated with the appearance of the noticed critical point(s) is M, then

$$|\mathrm{p}(\boldsymbol{K}_k) - \mathrm{p}(\boldsymbol{K}_{k-1})| = M,$$

where $\mathrm{p}(\boldsymbol{K})$ stands for the number of positive eigenvalues. This does not necessarily mean that the lowest critical point itself is a M-fold critical point, i.e.

$$\dim(\ker \boldsymbol{K}_{cr}) = K \leq M.$$

However, it is assumed in the sequel that the rank deficiency of the tangent stiffness matrix at the critical point equals to M and no other critical points lie on the primary path in the last increment.

The number of positive eigenvalues can be determined using the Sylvester law of inertia by counting the number of positive diagonal elements in the $\boldsymbol{LDL}^T$-factorized stiffness matrix. If the linear system is solved by iterative methods, like preconditioned conjugate gradient (PCG) method, the inertia of the stiffness matrix is not easily obtainable. If the preconditioning matrix is denoted by $\boldsymbol{M}$, the outer eigenvalues of the preconditioned operator $\boldsymbol{M}^{-1}\boldsymbol{K}$ can be easily obtained from the tridiagonal matrix related to the underlying Lanczos iteration. However, it is not clear if these eigenvalue estimates can be used in the continuation algorithm.

Multiplicity of the critical point is here defined to be the dimension of the nullspace of the tangent stiffness matrix at the critical point. Other definitions exist. Bauer et al. [5] defined the multiplicity of the bifurcation point to be M if $2M+2$ half rays meet there. Since the number of emanating branches is not known a priori, the previously mentioned definition seems to be more appropriate. It is probably the most common definition adopted in the literature.

3 BRANCH SWITCHING ALGORITHMS

In this section a short review of the existing branch switching techniques for multiple bifurcations is given. The task of these algorithms is to seek solutions for the rate of load parameter $\dot{\lambda}$ and the rates of the projections of the tangent vectors $\dot{a}_i$ of the branches onto the critical eigenmodes $\boldsymbol{\phi}_i, i = 1, ..., M$.

Rheinboldt [6] developed an elegant and computationally favourable branch switching algorithm for simple bifurcation. He also described generalization of his method to multiple bifurcation. However, the question of initial values for the projections $\dot{a}$ remained unanswered.

Keller [7] presented four algorithms, which are denoted methods I-IV. The method I uses a perturbation approach and the solutions for the branch directions are obtained from the algebraic bifurcation equation (ABE). In the evaluation of the coefficients in the ABE, the second derivatives of the residual vector $\boldsymbol{f}$ are needed, or they need to be approximated by finite differences. This method will fail when the ABE is degenerate, e.g. at symmetric bifurcations. In order to avoid the determination of coefficents of the ABE, Keller proposes the method II where the idea is to seek solutions on some subset parallel to the tangent but displaced from the bifurcation point in some direction normal to the tangent. Obviously this method will work well in simple bifurcations, but the problem with multiple bifurcation is how to parametrize in a reasonable way the subset where the solution is to be found. The remaining two methods III and IV seems to be the most robust and also computationally the most demanding. Since they are described in ref. [7] only in the case of simple bifurcation and they have some resemblance with the Koiter's perturbation approach, only the connections to the proposed method are pointed out in the following discussion.

Kearfott [1] developed a technique, where in principle, all solution arcs can be found by locating the minima of $\|\boldsymbol{f}\|$ in the region near the critical point spanned by the critical eigenvectors, i.e finding the solutions branches on a sphere centered to the estimate of the critical point. Drawback of this method is that it needs numerous evaluations of the residual $\boldsymbol{f}$. Determination of the the necessary resolution needed to find all solutions is an open question. If the resolution to scan over the sphere is too low, the probability of missing some brances increases, however tightening the resolution increases the computational cost. Huitfeldt [3] included also the tangent vector of the primary path in the definition of the sphere where the minimization takes place. Pajunen [8] has used the residual minimization technique to solve double bifurcation problem of a truss structure.

Allgower and Chien [2] used the local perturbation method introduced by Georg [9] to multiple bifurcation problems. The idea is to introduce a perturbation near the bifurcation point and solve the perturbed problem

$$\boldsymbol{f}(\boldsymbol{q}, \lambda) + \tau \boldsymbol{b} = \boldsymbol{0} \tag{2}$$

from a point on the primary path and traverse a perturbed path until it is near a point on a branch. The theoretical foundation of this method is based on a version of a generalized Sard's theorem. For successful branching the choice of the perturbation

vectors plays a key role. In their numerical examples the components in the perturbation vectors are chosen in such a way that they oscillated correspondingly to those of the bifurcating solutions. This means that one should have a priori knowlegde of the solution of the problem which has to be solved. No specific theory or rules for the selection of the perturbation vectors was given in ref. [2], and the method seems to be used best as computing the solution curves interactively by trial and error fashion.

A major improvement to the local perturbation algorithm is given by Huitfeldt [3]. He introduced an auxiliary equation which defines with the perturbed equilibrium equations (2) a closed one dimensional curve in a $N+2$-dimensional space. This curve passes exactly one point on each branch (or half branch) of the unperturbed equation (1). When passing such a point the perturbation parameter τ changes sign. The problem is then to locate the zero points of the perturbation parameter τ while traversing the branch connecting curve (BCC). Thus the branch switching problem is reduced to a path following task of the augmented system

$$\boldsymbol{h}(\boldsymbol{q},\lambda,\tau) = \begin{cases} \boldsymbol{f}(\boldsymbol{q},\lambda) + \tau \boldsymbol{b} &= \boldsymbol{0} \\ c_b(\boldsymbol{q},\lambda,\tau) &= 0 \end{cases}, \tag{3}$$

which can be solved with standard continuation algorithms. A constraint that defines a closed surface around the critical point is of spherical (elliptical) form:

$$c_b(\boldsymbol{q},\lambda,\tau) = \tfrac{1}{2}\left(\|\boldsymbol{q}-\boldsymbol{q}_{cr}\|_W^2 + \alpha^2(\lambda-\lambda_{cr})^2 + \beta^2\tau^2 - \rho^2 \right), \tag{4}$$

where α, β are scaling factors and ρ is the radius of the sphere. In principle this method does not need expensive evaluation of the basis of the nullspace of the tangent stiffness matrix. Huitfeldt [3] used a random vector as perturbation $\boldsymbol{b}$.

There are some shortcomings with this conseptually simple and elegant method. It is not known if the branch connecting equation always defines a closed curve. It is believed, as also argued by Huitfeldt, that using a constraint which defines a closed surface, guarantees a closed path defined by the branch connecting equation (3,4). No mathematical proof of this is known to the authors. Secondly, there is no guarantee that all bifurcating branches have been found. This obviously depends on the choice of the perturbation. In addition, the computational expense can be very high for large problems, fortunately it grows only linearily with respect to the emanating branches from the bifurcation point[1]. However, the number of branches in multimode buckling with higher multiplicity can be very large as will be explained in the following.

An essential feature for the construction of a reliable bifurcation procedure is the determination of the number of possible solutions branches emanating from the critical point. This problem has been explored in the late 60's by Sewell [10], [11], Johns and Chilver [12],[13]. Depending on the symmetry properties of the system, the maximum number of different post-buckling branches is

$$2^M - 1 \quad \text{or} \quad \tfrac{1}{2}(3^M - 1)$$

[1]It is assumed that for reliable detection of the zeros of the perturbation parameter on the BCC, a minimum number of steps, say 4-5, has to separate two consecutive roots.

for a system without symmetry or perfectly symmetric, respectively. The minimum number of post-buckling paths is 1 for the former case and M for the latter. The complexity of a multi-mode buckling problem grows enormously with the multiplicity of the critical point.

In order to develope a robust branch switching algorithm it seems natural to reduce the problem into a smaller one and to try to get as much information as possible from the reduced system, see e.g. [14]. Koiter's initial postbuckling theory is based on perturbation formulation resulting in a strongly reduced potential energy function, the variables being the amplitudes of the relevant buckling modes. The number of "post-buckling equilibrium equations" derived from the reduced potential energy expression equals the multiplicity of the buckling load or the number of pertinent interacting modes. A series expansion for the displacement field is used in the form [2]

$$\boldsymbol{q} = \lambda \boldsymbol{q}_r + \sum_{i=1}^{M} a_i(\lambda)\boldsymbol{q}_i + \sum_{i,j=1}^{M} a_i(\lambda)a_j(\lambda)\boldsymbol{q}_{ij},$$

where $\boldsymbol{q}_r$ and $\boldsymbol{q}_i$'s denote the reference displacement vector and buckling modes, $\boldsymbol{q}_{ij}$'s are the second order post-buckling fields and a_i's are the unknown amplitudes. The Koiter's approach consists of the following steps:

1. solution of the eigenvalue problem in order to get the relevant eigenmodes,
2. solving the second-order displacement fields,[3]
3. evaluation of the coefficients of the reduced system,
4. solution of the reduced set of equilibrium equations.

Since the dimension of the reduced problem is very small, any robust solution scheme can be applied. Notice that these equations are polynomial, hence, it is possible to find all the solutions with algorithms described in ref. [15].

Solving the amplitude equation in the vicinity of the critical point gives the local form of the equilibrium surface of the structure. The most severe limitation is that the range of validity of the results obtained is difficult to judge. Therefore the perturbation method has primarily been considered as an "analytical tool" to get qualitative picture of the behaviour of the initial post-buckling regime.

Another problem in the initial post-buckling method is to decide how many eigen-modes are relevant in the expansion. If one interacting mode is left out from the expansion, it will appear in the second order field [16]. However, the range of validity can be extremely small in those cases. An example of that is given in ref. [16] where a T-beam is analysed. The interacting buckling modes comprise two local and one overall mode, the critical load of which is higher than the loads corresponding the local modes. If the overall mode is left out from the series expansion, the resulting two mode analysis deviates rapidly from the three mode analysis after the secondary bifurcation point, which lies in the immediate vicinity of the primary bifurcation point.

[2]Here the behaviour on the primary path is assumed to be almost linear.

[3]In Keller's approach III [7] the second-order fields have to be solved from a non-linear equation system. It is also unclear how the amplitudes in his method are determined in multimode buckling problems.

4 EXAMPLE

A well known example of multiple bifurcation is the double bifurcation of a compressed flat simply supported plate, fig. 1. A plate with aspect ratio $\sqrt{2}$ is chosen as a test example. This problem has also been analysed by Huitfeldt [3] and Lidström [17], however, no post-bifurcation paths have been presented. The plate is discretized by uniform 20×10 quadrilateral mesh using bilinear stabilized MITC type elements with drilling rotations (1304 dof). The stabilization parameter (shear reduction) for the MITC element has been 0.4 [18], [19], [20]. Full 2×2 Gaussian integration is used in evaluation of the element stiffness matrices and internal force vectors. The loaded edges are constrained to remain straight, but the in-plane deflections are allowed for the longitudinal sides (Hemp type boundary conditions). The analytical buckling load has the value $P_{cr} = 4.5\pi^2 D/L$, where L is the length of the loaded side and D is the bending rigidity of the plate $D = Et^3/12(1-\nu^2)$. The buckling modes corresponding to this double bifurcation load have one or two half waves in the x-axis direction. The length to thickness ratio is $L/t = 100$ and the Poisson's ratio has the value $\nu = 0.3$. In the numerical computation the value obtained is $4.39\pi^2 D/L$ interpolated from the zero point of the lowest eigenvalue of the tangent stiffness matrix, which is easily computed at the beginning of each increment by applying few inverse iterations. If an eigenvalue buckling analysis is performed, the double eigenvalue will split in two separate eigenvalues with values 4.32 and $4.40\pi^2 D/L$.

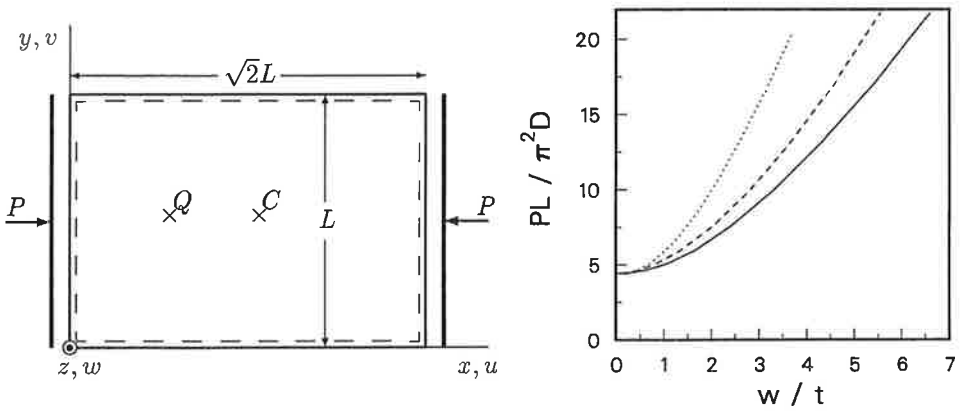


Figure 1: (a) Simply supported plate, (b) load deflection curves; solid line = w_C mode 1 branch, dashed line = w_Q mode 1 branch, dotted line = w_Q mode 2 branch.

This example is not particularly difficult, since there are only two post-buckling branches, deformation patterns of which are just like the buckling modes. The load deflection paths are shown in fig. 1. C and Q refer to the center and quarter points of the plate. Deformed shapes at the end of the continuation are shown in fig. 2.

Only Huitfeldt's approach and the branch-switching scheme based on Koiter's initial post-buckling theory are used in the computations. Results based on Huitfeldt's approach are reported first. Plot of the perturbation parameter τ with respect to the

Figure 2: Deformed states at post-buckling branches 1 and 2.

arc-length along the branch connection curve is shown in fig. 3 as well as the distance from the starting point of the branch connecting curve. The branch connecting curve is traced with 56 increments and it has six roots for the perturbation parameter (two on primary path and four on post-buckling branches) and it constitutes the main computational effort, since the primary path and one branch is traversed within 14 steps.

Magnitude of the perturbation load is chosen in such a way that the maximum deflection caused by the perturbing load equals to the value $0.2t$. In computations, shown in figs. 1, 3 the perturbation load vector is chosen to have only one point load at the quarter point Q. The distance d from the begining of the BCC tracing is defined by

$$d = \left(\|\boldsymbol{q} - \boldsymbol{q}^*\|_w^2 + \alpha^2(\lambda - \lambda^*)^2 + \beta^2\tau^2 \right)^{1/2},$$

where $\boldsymbol{q}^*, \lambda^*$ is the starting point on the primary path.

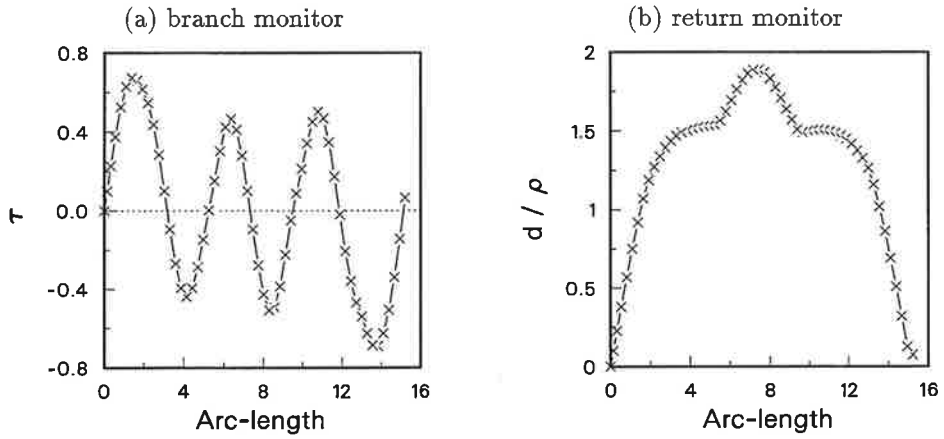


Figure 3: Double bifurcation of simply supported plate. Traversing branch connecting curve. The arc-length is normalized with respect to the radius ρ of the spherical constraint.

If a random vector is used for the pertubation, the BCC-tracing failed for several magnitudes tried. For values $\|\boldsymbol{q}_b\|_{W,max} = 0.2t, 0.3t$ the iterations did not converge even if the step size was halved five times[4]. When using the value $\|\boldsymbol{q}_b\|_{W,max} = 0.1t$ the BCC-tracing was stopped after 200 steps and after finding 28 zeros[5] for the perturbation parameter. Even if the closed surface constraint for the BCC is used, this particular perturbation vector does not seem to define a closed path. The perturbation parameter with respect to the normalized arc-length as well as the distance from departure are shown in fig. 4.

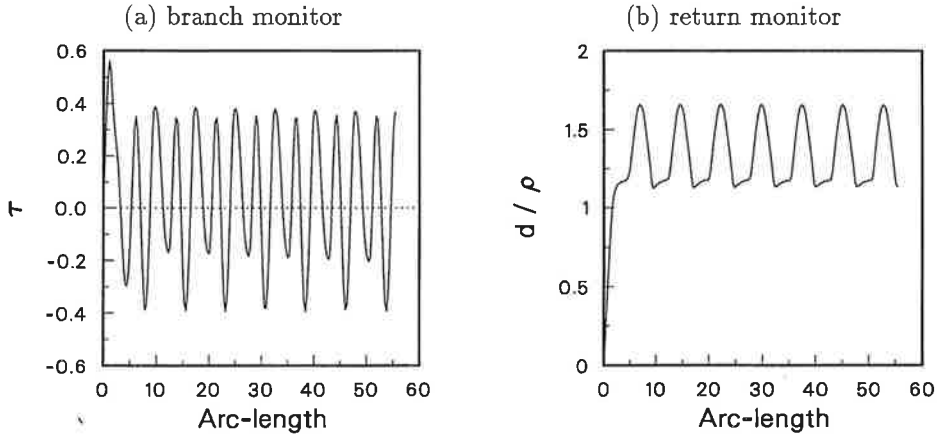


Figure 4: Failure in traversing BCC; the arc-length is normalized with respect to the radius ρ of the spherical constraint.

In the Koiter type perturbation method the reduced potential energy has the form

$$V(a_1, a_2) = \tfrac{1}{2}\left[(1 - (\lambda/\lambda_{cr}))\left(a_1^2 + a_2^2\right)\right] + A_{1111}a_1^4 + A_{1122}a_1^2a_2^2 + A_{2222}a_2^4,$$

where a_1 and a_2 are the dimensionless amplitudes of the buckling modes. In this simple example the two post-buckling branches emanating from the critical point can be solved analytically and they are simply:

- mode 1 branch:
$$\lambda/\lambda_{cr} = 1 + 4A_{1111}a_1^2, \quad \text{and} \quad a_2 \equiv 0,$$
- mode 2 branch:
$$\lambda/\lambda_{cr} = 1 + 4A_{2222}a_2^2, \quad \text{and} \quad a_1 \equiv 0.$$

The values a_1 and a_2 which are used in the prediction step onto the post-buckling branch can be defined by fixing the maximum displacement of the predictor. Solution time

[4]The norm $\|\boldsymbol{q}_b\|_{W,max}$ denotes the maximum norm taken from pure displacement components (rotations excluded). $\boldsymbol{q}_b$ is the displacement vector caused by the perturbing load $\boldsymbol{b}$.

[5]In this example the BCC tracing should be stopped after finding seven zeros without returning to the point of departure.

which is needed for this kind of branch-switching algorithm is a fraction compared to the Huitfeldt's method. Nevertheless, the Huitfeldt's approach can be used to solve the small reduced polynomial equation system instead of using the polynomial continuation methods.

5 DISCUSSION

So far all existing branch switching techniques which can be used in multiple bifurcation problems have some annoying features. In principle Huitfeldt's approach for traversing the branch connecting curve requires only a path following routine, no other specific algorithms are needed. This is in contrast to other branch switching methods which requires the basis of the nullspace of the tangent stiffness matrix, i.e. the eigenmodes. However, in practise also with Huitfeldt's approach, some knowledge on the critical eigenmodes seems to be necessary in order to construct a proper perturbation load.

To solve the eigenvalue problem at the critical point is the most time consuming part of the proposed branch switching algorithm which uses the Koiter-type reduction method. It is believed that this approach is also much more economical with respect to computing time than Kearfott's minimization procedure, in which a lot of residual computations are needed. The price which has to be paid is the formulation of the "second-order" load vectors, where the second order derivatives of the residual appear.

As mentioned before, the range of applicability of the Lyapunov-Schmidt-Koiter type reduction can be very narrow due in the case that some relevant interacting modes are left out from the series expansion. However, this usually manifests itself by the appearance of secondary bifurcations close to the primary one. It is extremely difficult to automate the selection of the relevant buckling modes. Thus, human expertise in performing stability computations involving interactive buckling phenomena is crucial for successful analysis.

REFERENCES

[1] R.B. Kearfott. Some general bifurcation techniques. *SIAM Journal on Scientific and Statistical Computing*, 4:52–68, 1983.

[2] E.L. Allgower and C.-S. Chien. Continuation and local perturbation for multiple bifurcations. *SIAM Journal on Scientific and Statistical Computing*, 7:1265–1281, 1986.

[3] J. Huitfeldt. Nonlinear eigenvalue problems - prediction of bifurcation points and branch switching. Technical Report 17, Department of Computer Sciences, Chalmers University of technology, 1991.

[4] R. Kouhia and M. Mikkola. Strategies for structural stability analyses. In N.E. Wiberg, editor, *Advances in Finite Element Technology*, pages 254–278, 1995.

[5] L. Bauer, H.B. Keller, and E.L. Reiss. Multiple eigenvalue lead to secondary bifurcation. *SIAM Review*, 17(1):101–122, 1975.

[6] W.C. Rheinboldt. Numerical methods for a class of finite dimensional bifurcation problems. *SIAM Journal on Numerical Analysis*, 15:1–11, 1978.

[7] H.B. Keller. Numerical solution of bifurcation and nonlinear eigenvalue problems. In P.H. Rabinowitz, editor, *Applications of Bifurcation Theory*, pages 359–384. Academic Press, 1977.

[8] S. Pajunen. Sauvarakenteiden epälineaarinen analysointi (Nonlinear analysis of bar structures), 1997. Licentiates thesis, (in Finnish) Tampere University of Technology, Department of Civil Engineering.

[9] K. Georg. On tracing an implicitely defined curve by quasi-newton steps and calculating bifurcation by local perturbations. *SIAM Journal on Scientific and Statistical Computing*, 2:35–50, 1981.

[10] M.J. Sewell. A general theory of equilibrium paths through critical points. *Proceedings of the Royal Society - A*, 306:201–238, 1968.

[11] M.J. Sewell. On the branching of equilibrium paths. *Proceedings of the Royal Society - A*, 315:499–518, 1970.

[12] K.C. Johns and A.H. Chilver. Multiple path generation at coincident branching points. *International Journal of Engineering Science*, 13:899–910, 1971.

[13] K.C. Johns. Simultaneous buckling in symmetric structural systems. *Journal of Structural Division, ASCE*, 1972.

[14] M. Potier-Ferry. *Buckling and Post-Buckling*, volume 288 of *Lecture Notes in Physics*, pages 205–223. Springer-Verlag, 1987.

[15] A. Morgan. *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*. Prentice-Hall, 1987.

[16] R. Kouhia, C.M. Menken, M. Mikkola, and G.-J. Schreppers. Computing and understanding interactive buckling. In R.A.E. Mäkinen and P. Neittaanmäki, editors, *Proceedings of the 5th Finnish Mechanics Days*, pages 53–61, 1994.

[17] T. Lidström. *Computational methods for finite element instability analyses*. PhD thesis, Royal Institute of Technology, Department of Structural Engineering, 1996.

[18] M. Lyly, R. Stenberg, and T. Vihinen. A stable bilinear element for the Reissner-Mindlin plate model. *Computer Methods in Applied Mechanics and Engineering*, 110:343–357, 1993.

[19] D.J. Allman. A compatible triangular element including vertex rotations for plane elasticity analysis. *Computers and Structures*, 19:1–8, 1984.

[20] T.J.R. Hughes and F. Brezzi. On drilling degrees of freedom. *Computer Methods in Applied Mechanics and Engineering*, 72:105–121, 1989.

POST-BUCKLING ANALYSIS OF PLATES AND SHELLS

S. PAJUNEN and M. TUOMALA
Laboratory of Structural Mechanics
Tampere University of Technology
PL 600, 33101 Tampere, FINLAND

ABSTRACT

In this study, geometrically non-linear analysis of plate and shell structures is considered. The analyzed structures are discretized by rectangular C_1-continuous conforming plate elements, axisymmetric Reissner shell elements and general thin shell elements. The non-linear equilibrium path is computed by an arc-length method. Critical points on the equilibrium path are located in desired accuracy and identified as limit or bifurcation points. The limit points cause no problem for the arc-length methods, but at bifurcation points special methods for switching to the secondary paths have to be used.

1. INTRODUCTION

In the finite element method the solution of a structural problem can be presented as an equilibrium path in a (n+1)-dimensional space spanned by n nodal point displacement degrees-of-freedom $\mathbf{q}$ and a load parameter λ (assuming proportional loading). At limit points the path tangent is perpendicular to the λ-axis and Newton's iteration method parametrized by λ breaks down. The solution curve or the equilibrium path can be continued past limit points by augmenting the equilibrium equations with a normalizing or constraint equation and parametrizing the solution path by a parameter $s \in \mathbf{R}$. Several choices for the constraint equation have been proposed in the literature, see e.g. [10]. The parameter s can be made to approximate the arc-length of the solution curve.

At bifurcation points the tangent stiffness matrix has a single or multiple zero eigenvalue and special methods have to be adopted in order to switch to a secondary path or otherwise the continuation method keeps following the primary solution branch. For simple bifurcation

points the existing methods are quite reliable but multiple bifurcation points can be much more difficult to handle.

The present study is a continuation to the work reported in [12]. The numerical testing of continuation and branch-switching methods is extended into plate and shell problems. Path following is tested by calculating a rather complicated equilibrium path for a non-linear axisymmetric shell. A mode jumping phenomenon is considered in the context of a plate problem and finally the imperfection sensitivity of a compressed cylindrical panel is analyzed.

2. SOLUTION OF THE EQUILIBRIUM EQUATIONS

The equilibrium equations in the finite element method can be written in the form

$$\mathbf{g}(\mathbf{q},\lambda) \equiv \mathbf{r}(\mathbf{q}) - \lambda\mathbf{p} = \mathbf{0} \tag{1}$$

where $\mathbf{q}$ is nodal point displacement vector, $\mathbf{r}$ is the vector of internal forces, λ is the load factor and $\mathbf{p}$ is a reference load vector. Equations (1) are usually solved incrementally by Newton's method:

$$\mathbf{K}_T(\mathbf{q}_k^i)d\mathbf{q}^{i+1} = \lambda_k\mathbf{p} - \mathbf{r}(\mathbf{q}_k^i) \tag{2}$$

where $\mathbf{K}_T=\partial\mathbf{g}/\partial\mathbf{q}$ is the tangent stiffness matrix, k and i denote the load step and the iteration cycle number, respectively.

For handling possible limit points on the equilibrium path (1) is augmented by a constraint equation yielding a system of n+1 equations

$$\mathbf{G}(\mathbf{x},s) = \begin{Bmatrix} \mathbf{g}(\mathbf{q},\lambda) \\ c(\mathbf{q},\lambda,s) \end{Bmatrix} = \mathbf{0} \tag{3}$$

where $\mathbf{x}^T=(\mathbf{q}^T,\lambda)$. If the Jacobian of $\mathbf{G}$ is nonsingular then the solution can be continued from a known solution point $(\mathbf{q}_0,\lambda_0)$. At regular or limit points

$$\mathbf{G}_x = \begin{bmatrix} \mathbf{g}_q(\mathbf{q},\lambda) & \mathbf{g}_\lambda(\mathbf{q},\lambda) \\ \mathbf{c}_q^T(\mathbf{q},\lambda,s) & c_\lambda(\mathbf{q},\lambda,s) \end{bmatrix} \tag{4}$$

is nonsingular if the vector $\mathbf{c}_q$ is not perpendicular to the tangent of the solution curve [6]. Applying Newton's method to the extended system yields

$$\begin{aligned} \mathbf{g}_q d\mathbf{q} + \mathbf{g}_\lambda d\lambda &= -\mathbf{g}, \\ \mathbf{c}_q^T d\mathbf{q} + \mathbf{c}_\lambda d\lambda &= -\mathbf{c}, \end{aligned} \tag{5}$$

in which $\mathbf{g}_\lambda$=-$\mathbf{p}$. The system (5) is solved in two parts [9]:

$$\begin{aligned} d\mathbf{q}_p &= \mathbf{K}_T^{-1}\mathbf{p} \\ d\mathbf{q}_g &= \mathbf{K}_T^{-1}\mathbf{g} \end{aligned} \tag{6}$$

with

$$d\mathbf{q} = d\mathbf{q}_g + d\lambda d\mathbf{q}_p . \tag{7}$$

In an updated normal plane method, where a constraint

$$\Delta\mathbf{q}^i \cdot d\mathbf{q}^{i+1} + \Delta\lambda^i d\lambda^{i+1} = 0 \tag{8}$$

with $\Delta\mathbf{q}^i = \mathbf{q}_{n+1}^i - \mathbf{q}_n$ is imposed, the change of load factor is

$$d\lambda^{i+1} = -\frac{\Delta\mathbf{q}^i \cdot d\mathbf{q}_g^{i+1}}{\Delta\lambda^i + \Delta\mathbf{q}^i \cdot d\mathbf{q}_p^{i+1}} . \tag{9}$$

Fried [5] has proposed an alternative version, the so called orthogonal trajectory method, in which

$$d\lambda^i = -\frac{d\mathbf{q}_p^i \cdot d\mathbf{q}_g^i}{1 + d\mathbf{q}_p^i \cdot d\mathbf{q}_p^i} . \tag{10}$$

3. BIFURCATION ALGORITHM

In this context the 'bifurcation algorithm' means an algorithm that locates and classifies the singular points on computed equilibrium paths and searches the directions of possible secondary paths. The bifurcation algorithm is switched on when the number of negative pivot elements of $\mathbf{K}_T$ changes during an arc-length step k, say. The algorithm returns the arc-length solver to the previous equilibrium state ($\mathbf{u}_{k-1}$, λ_{k-1}) and locates the observed singular point in desired accuracy by bisecting the step size, see Fig. 1. In the next phase the bifurcation algorithm classifies the

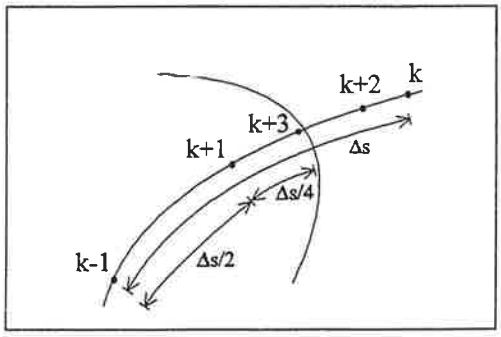


Figure 1. Schematic picture on the bisection method.

singular point as a limit or a bifurcation point according to following test functions [2,9,19]:

$$\left|\frac{\Delta\lambda_k \Delta\mathbf{q}_1 \cdot \mathbf{p}}{\Delta\lambda_1 \Delta\mathbf{q}_k \cdot \mathbf{p}}\right| \le tol \quad \Rightarrow \text{ limit point},$$

$$\frac{|\varphi^T \mathbf{p}|}{|\varphi||\mathbf{p}|} \le tol \quad \Rightarrow \text{ bifurcation point},$$

in which a typical value of *tol* is 10^{-2}. In this study only simple singular points are considered but application to the case of higher order singularity of $\mathbf{K}_T$ is rather straightforward: only the search of secondary branch directions has to be done by adopting more advanced methods, see e.g. [6,8].

If the singular point is found to be a limit point, then the bifurcation algorithm is turned off and the arc-length method is restarted at the equilibrium state ($\mathbf{u}_k$, λ_k). In the case of a bifurcation point, the directions of secondary equilibrium branches are searched by using the algebraic bifurcation equation, see sec. 3.1, for alternative branch-switch methods, see e.g. [12]. Finally, when the secondary branches are found and continuated far enough the bifurcation algorithm is switched off and the primary path continuation is restarted from the equilibrium state ($\mathbf{u}_k$, λ_k).

3.1 Algebraic bifurcation equations

On a solution path $\mathbf{g}(\mathbf{q}(s),\lambda(s))=\mathbf{0}$ the differentiation with respect to s, marked with $(\,)'$, yields

$$\mathbf{g}_q^0 \mathbf{q}_0' + \mathbf{g}_\lambda^0 \lambda_0' = \mathbf{0}, \tag{11}$$

$$\mathbf{g}_q^0 \mathbf{q}_0'' + \mathbf{g}_\lambda^0 \lambda_0'' = -(\mathbf{g}_{qq}^0 \mathbf{q}_0' \mathbf{q}_0' + 2\mathbf{g}_{q\lambda}^0 \mathbf{q}_0' \lambda_0' + \mathbf{g}_{\lambda\lambda}^0 \lambda_0' \lambda_0'), \tag{12}$$

at a solution point $\mathbf{q}_0 \equiv \mathbf{q}(s_0)$, $\lambda_0 \equiv \lambda(s_0)$. At a limit or bifurcation point the derivative of $\mathbf{g}$ with respect to $\mathbf{q}$, denoted by $\mathbf{g}_q^0 = \mathbf{g}_q(\mathbf{q}_0,\lambda_0)$, is singular. The null space of $\mathbf{g}_q^0$ is spanned by the eigenvectors $\{\varphi_1,\ldots,\varphi_m\}$, $|\varphi_i|=1$. The range of $\mathbf{g}_q^0$ is

$$\mathcal{R}(\mathbf{g}_q^0) = \{\mathbf{x} \in \mathbf{R}^n \,|\, \psi_i^T \mathbf{x} = 0, i = 1,\ldots,m\},$$

where the left eigenvectors ψ_i satisfy $\psi_i^T \varphi_j = \delta_{ij}$. The existence of a solution $\mathbf{q}_0'$ to (11)

reqiures that $\mathbf{g}_\lambda^0 \in \mathcal{R}(\mathbf{g}_q^0)$ or $\lambda_0' = 0$. In the first case the solution point $(\mathbf{q}_0,\lambda_0)$ is a bifurcation point. At a bifurcation point there is a unique solution $\mathbf{v}$ such that

$$\mathbf{g}_q^0\mathbf{v} + \mathbf{g}_\lambda^0 = \mathbf{0} \text{ with } \psi_j^T\mathbf{v} = 0, \quad j = 1,\ldots,m$$

The general solution of (11) can be written in the form

$$\mathbf{q}_0' = \eta\mathbf{v} + \sum_{j=1}^{m} \xi_j\varphi_j; \quad \eta = \lambda_0'. \tag{13}$$

Substituting (13) into (12) yields a necessary condition for the existence of a solution $\mathbf{q}_0''$, the algebraic bifurcation equations (ABE) [9]:

$$\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk}\xi_j\xi_k + 2\sum_{j=1}^{m} b_{ij}\xi_j\eta + c_i\eta^2 = 0, \; i = 1,\ldots,m, \tag{14}$$

where

$$a_{ijk} = a_{ikj} \equiv \psi_i^T\mathbf{g}_{qq}^0\varphi_j\varphi_k,$$
$$b_{ij} \equiv \psi_i^T\mathbf{g}_{qq}^0\mathbf{v}\varphi_j,$$
$$c_i \equiv \psi_i^T\mathbf{g}_{qq}^0\mathbf{v}\mathbf{v}.$$

The homogenous polynomial equations (14) are augmented by a normalization equation

$$\eta^2 + \xi_1^2 + \ldots + \xi_m^2 = 1.$$

In the case m=1 (14) reduces to (denoting $\xi \equiv \xi_1$, $\psi \equiv \psi_1$ and $\varphi \equiv \varphi_1$)

$$a\xi^2 + 2b\xi\eta + c\eta^2 = 0, \tag{15}$$

where

$$a = \psi^T\mathbf{g}_{qq}^0\varphi\varphi,$$
$$b = \psi^T\mathbf{g}_{qq}^0\mathbf{v}\varphi,$$
$$c = \psi^T\mathbf{g}_{qq}^0\mathbf{v}\mathbf{v}.$$

The type of bifurcation depends on the coefficients a, b and c and on the value of the discriminant $d=b^2-ac$. At symmetric bifurcation point the coefficient a equals zero and the eigenvector φ defines the direction of the secondary path. In the opposite case of an asymmetric bifurcation point, the direction can be computed from (15) augmented by a constraint:

$$\eta^2 + \xi^2 = 1. \tag{16}$$

The coefficients a, b and c can be calculated analytically elementwise for some simple finite elements like for the truss element but generally they can be obtained numerically e.g. by using the finite difference method [9]. After finding the direction of the secondary path and switching on it, the path can be continuated by using the arc-length method as usually.

4. SOME PLATE AND SHELL ELEMENTS

4.1 An axisymmetric shell element based on Reissner's theory

Reissner [13] has developed a nonlinear axisymmetric shell theory in which the strain components are defined as follows:

$$\begin{aligned}
\varepsilon &= (r'+u')\cos\varphi + (y'+v')\sin\varphi, \\
\gamma &= (y'+v')\cos\varphi + (r'+u')\sin\varphi, \\
\varepsilon_\theta &= u/r, \\
\kappa_s &= \varphi' - \varphi'_0 \\
\kappa_\theta &= (\sin\varphi - \sin\varphi_0)/r, \\
\kappa_n &= (\cos\varphi - \cos\varphi_0)/r,
\end{aligned} \tag{17}$$

where y is the symmetry axis, r is the radial coordinate, u and v are the displacements in the r- and y-directions, respectively, φ measures the angle between the shell cross section and the y-axis (in the undeformed state $\varphi=\varphi_0$), $(\,)' \equiv d(\,)/ds$, θ means the circumferential direction, s is the arc-length along the shell meridian, ε, γ and ε_θ are the membrane deformations, κ_s, κ_θ, and κ_n are the shell curvature changes, respectively.

In a stress resultant formulation of the shell theory the components conjugate to the membrane and curvature deformation measures are the membrane stress resultants N, Q and N_θ and the moments M_s, M_θ, and M_n. The strain variables and the stress resultants are collected in the respective vectors

$$\mathbf{e}=[\varepsilon, \gamma, \varepsilon_\theta, \kappa_s, \kappa_\theta, \kappa_n]$$

$$\mathbf{s} = [N, Q, N_\theta, M_s, M_\theta, M_n].$$

Restricting to small strains the following constitutive equations can be adopted for an elastic material

$$\begin{bmatrix} N \\ N_\theta \end{bmatrix} = \frac{Eh}{1-\nu^2} \begin{bmatrix} 1 & \nu \\ \nu & 1 \end{bmatrix} \begin{bmatrix} \varepsilon \\ \varepsilon_\theta \end{bmatrix}$$

$$Q = Gh\gamma$$

$$\begin{bmatrix} M_s \\ M_\theta \\ M_n \end{bmatrix} = \frac{Eh^3}{12(1-\nu^2)} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{bmatrix} \begin{bmatrix} \kappa_s \\ \kappa_\theta \\ \kappa_n \end{bmatrix},$$

where E and ν are elastic material constants, G=E/2(1+ν) and h is the shell thickness.

In rubber elasticity the restriction into small strains can be removed by introducing a potential function, e.g. the Green-Rivlin potential, from which the stresses follow by differentiation. For an elasto-plastic material, constitutive models valid for arbitrary strains can be formulated in terms of the Cauchy stress and the logarithmic strain.

From the definitions of the strain components the virtual strains can be derived and written in the form

$$\delta\varepsilon = \mathbf{B}\delta\mathbf{q}, \tag{18}$$

where the vector $\mathbf{q}$ contains the element nodal degrees-of-freedom and at a node i $\mathbf{q}_i^T = [u_i \quad v_i \quad \varphi_i]$. The displacement components u and v and the angle φ are interpolated by Lagrangian polynomials. In the simplest, two-noded element linear polynomial shape functions are used and in order to avoid shear locking in thin shells, integrals over s are performed by a one point Gaussian rule.

The contribution to the internal force vector from an element e is

$$\mathbf{r}_e = \int_{V_e} \mathbf{B}^T s dV \tag{19}$$

By taking a linear increment of (19) yields a formula

$$\mathbf{K}_T \Delta\mathbf{q} = \int_{V_e} \mathbf{B}^T \mathbf{D} \mathbf{B} dV \Delta\mathbf{q} + \int_{V_e} \Delta\mathbf{B}^T s dV$$

from which the element tangent stiffness matrix

$$\mathbf{K}_T = \mathbf{K}_0 + \mathbf{K}_g$$

is obtained. $\mathbf{K}_g$ is the geometric stiffness matrix. Consistent linearization is necessary for obtaining good (quadratic) convergence rate in the Newton iteration method.

4.2 Shallow shell element based on Hermitian bi-cubic shape functions

One of the most effective thin plate elements is the so called BFS-element [3] in which the deflection w(x,y) is interpolated by Hermitian bi-cubic polynomials. The nodal variables at the four nodes of the element comprise w, $w_{,x}$, $w_{,y}$ and $w_{,xy}$. The inclusion of a deformation type measure, $w_{,xy}$, in the nodal variables is a slight drawback of the element.

In a non-linear shallow shell theory the Green-Lagrange strains are

$$\varepsilon_x = u_{,x} + \frac{1}{2} w_{,x}^2 + w_{0,x}\, w_{,x} - z w_{,xx}$$

$$\varepsilon_y = v_{,y} + \frac{1}{2} w_{,y}^2 + w_{0,y}\, w_{,y} - z w_{,yy}$$

$$\gamma_{xy} = u_{,y} + v_{,x} + w_{,x}\, w_{,y} + w_{0,x}\, w_{,y} + w_{0,y}\, w_{,x} - 2 z w_{,xy}$$

where $w_0(x,y)$ is the initial deflection. The membrane displacements u and v can be interpolated by bi-linear shape functions or the same Hermitian bi-cubic polynomials as adopted for the deflection w can be used.

4.3 Semi-loof shell element

The semi-loof shell element was developed by B.Irons [7] for the analysis of general doubly curved thin shells. The element is obtained from an isoparametric degenerated thick shell element by constraining the transverse shear stresses to zero at discrete points and by imposing additional constraint equations in integral form.

At one time the semi-loof element was considered as one of the most effective general thin shell elements and it is still included in the element libraries of some commercial general purpose finite element codes.

The element contains as its nodal degrees-of-freedom the translations u, v and w in the global x, y and z directions at the conventional eight nodes and the normal rotations about the element side tangent vectors at the so called loof nodes. The loof nodes are situated on the element boundaries at $|\xi|=1/\sqrt{3}$ or $|\eta|=1/\sqrt{3}$, i.e. at points corresponding to the integration points of the two-point Gaussian rule, and $\xi,\eta \in [-1,1]$ are the element natural coordinates.

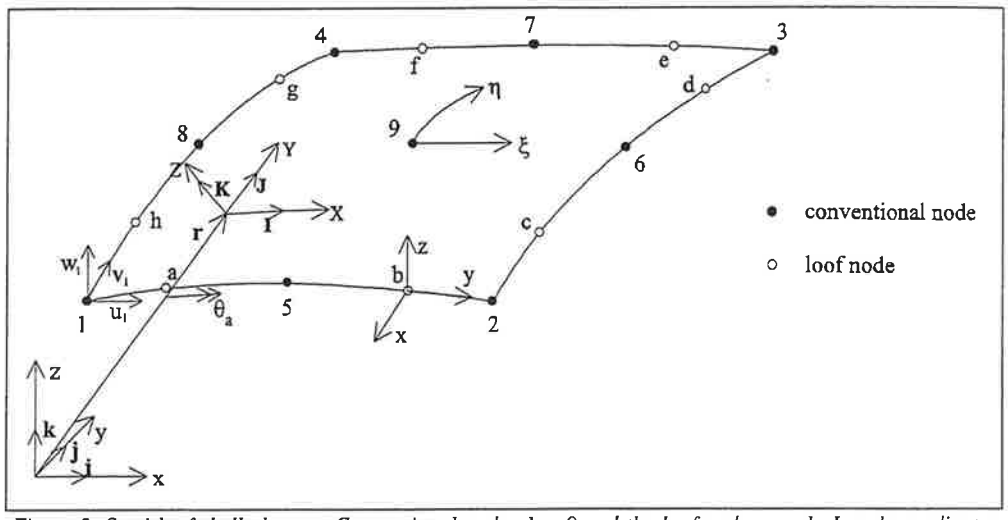


Figure 2. Semi-loof shell element. Conventional nodes 1,...,9 and the loof nodes a,...,h. Local coordinates $\boldsymbol{K}=\boldsymbol{r}_{,\xi}\times\boldsymbol{r}_{,\eta}/|\boldsymbol{r}_{,\xi}\times\boldsymbol{r}_{,\eta}|$, $\boldsymbol{I}=\boldsymbol{r}_{,\xi}/|\boldsymbol{r}_{,\xi}|$ *and* $\boldsymbol{J}=\boldsymbol{I}\times\boldsymbol{K}$, *where* $\boldsymbol{r}$ *is an arbitrary point on the element.*

5. EXAMPLES ON SOME INSTABILITY PHENOMENA

5.1 Snap-through of a shallow spherical dome

A point-loaded shallow axisymmetric shell in Fig. 3 is analyzed in order to test a shell element based on Reissner's theory. The shell is discretized with 22 elements giving the equilibrium path shown in Fig. 3. The path is rather complex having up to ten limit points in the apex deflection - load plane. Thus, the example serves also as a reliability test for the arc-length methods. Both an updated normal-plane method (UNP) and an orthogonal trajectory (OT) arc-length method version are tested with the full Newton-Raphson iteration scheme. By using an iteration terminating criterion tolerance 10^{-6} the maximum valid step size for the both methods is about 0.22.

The computed path is in a good agreement to those reported by Argyris et. al. [1] and Wagner et. al. [18]. In this study, as well as in the papers [1] and [18], only the primary path associated with the symmetric deflection is considered but certainly there are also some bifurcation points associated with secondary path(s) displaying asymmetric buckling mode(s). However, these asymmetric modes can not be analyzed by using axisymmetric elements.

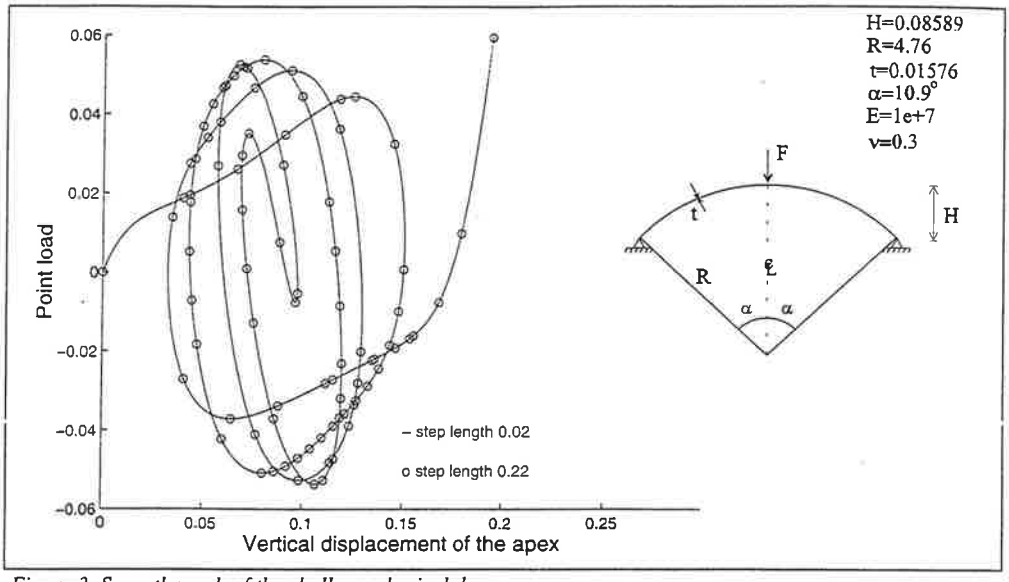


Figure 3. Snap-through of the shallow spherical dome.

5.2 Buckling-mode jumping in a uniaxially compressed square plate

In the papers by Stein [15] and Uemura & Byon [17] a buckling mode jumping phenomenon in plates is analyzed experimentally. In this study we concentrate on the plate introduced in [17] and further considered by Carnoy & Hughes [4]. The initial data of the test plate is depicted in Fig. 4. The plate is discretized by a 10×10 BFS element mesh and Fried's arc-length version with the full Newton-Raphson iteration method is adopted with the step size 1.0 and convergence tolerance 10^{-4}. For robust identification the critical points are searched rather accurately so that the absolute value of the minimum pivot element of the tangent stiffness matrix is forced to be less than 10^{-3}. When compressed uniformly in one direction, the plate shortens linearly until the load reaches its first critical value at $p_{cr,1}$=15.70. At this point two secondary branches emanate

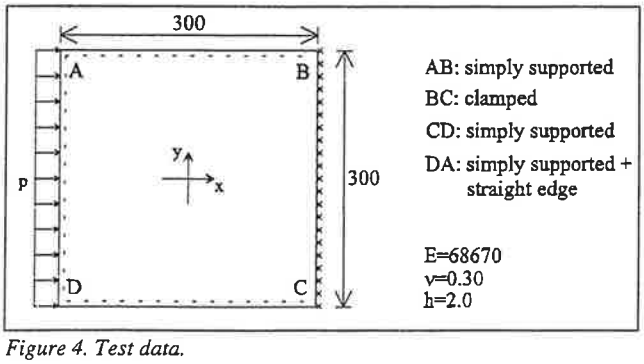


Figure 4. Test data.

from the primary equilibrium path, both of which are associated with the buckling mode consisting of one bubble in the x-y plane (see Figs. 4 and 5).

If the primary path is continued beyond the first critical point, then several more critical points will appear e.g. at $p_{cr,i}$=18.46, 31.15, 38.94, 42.38 and 43.69. Additionally, all the critical points on the primary path are identified as symmetric bifurcation points. When following the secondary path it is found to be stable up to the first secondary bifurcation point at $p_{cr,2}$=35.17. From this point the used branch-switching algorithm ABE leads the solution to an unstable path that changes the deformation shape of the plate from the one-bubble mode to a mode consisting of two waves in the x-direction and one wave in the y-direction. At the point where the deformations of the plate are fully recovered by the two-wave mode another bifurcation point appears. This bifurcation point at $p_{cr,3}$=23.33 connects the previously followed 'mode transition' path to a secondary path associated with the two-wave mode. This path can be numerically followed 'upwards' where the path is stable or 'downwards' when the path leads the solution back to the primary path to the load level $p_{cr,4}$=18.46.

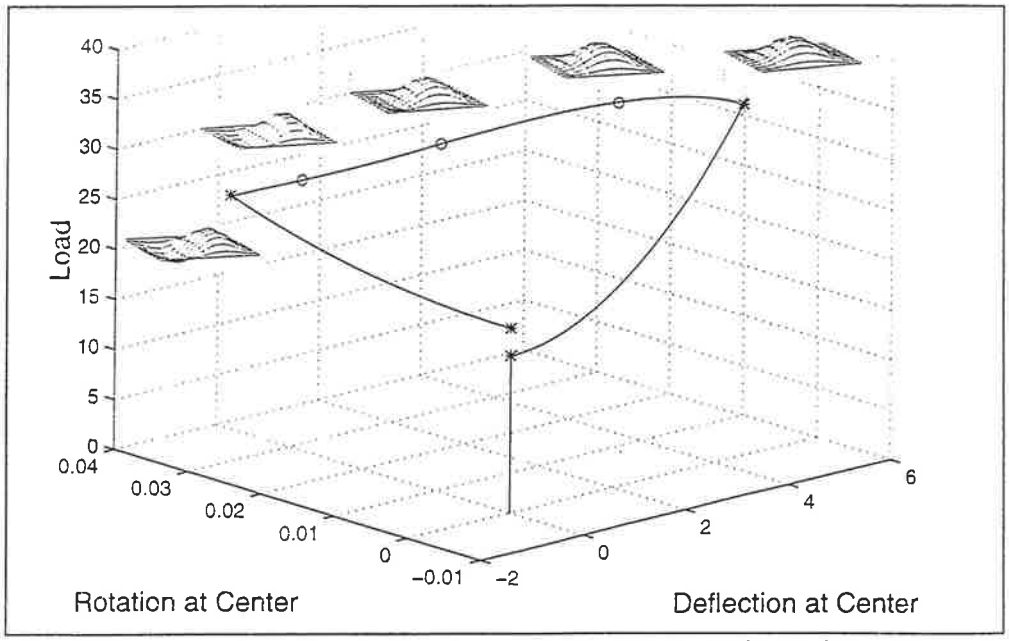


Figure 5. Some equilibrium paths and deformation modes of a plate. The modes are drawn at the points marked with 'o' and at the bifurcation points marked with ''.*

In the next phase we analyze the plate under the influence of small initial imperfections. In Fig. 6. the structural behavior of the plate under symmetric (ε_1) or antisymmetric (ε_2) initial imperfection shape is shown. The symmetric and antisymmetric imperfections are chosen to have the same shape as the one-bubble and two-wave buckling modes, respectively. In the case $\varepsilon_1 \neq 0$, $\varepsilon_2 = 0$ the imperfect equilibrium path follows the perfect one towards the symmetric bifurcation point at $p_{cr,2}=35.17$. It should be noted that also the imperfect plate has now a bifurcation point at $p_{cr,2*}=36.52$.

In order to unfold also the secondary bifurcation point at $p_{cr,2}=36.52$ we analyze the plate with both $\varepsilon_1 \neq 0$ and $\varepsilon_2 \neq 0$. Here the imperfection mode amplitudes ε_i are chosen to be ca. 0.1% of the plate thickness. The resulting path shows once more the mode jumping phenomenon, see Fig. 6. The both secondary bifurcation points have now become limit points ($p_{cr,2*}=34.23$ and $p_{cr,3*}=23.49$).

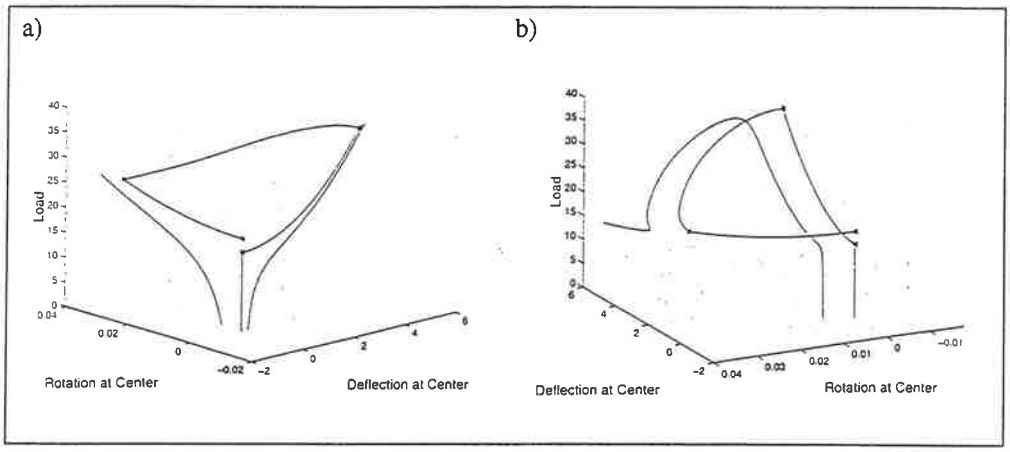


Figure 6. Equilibrium paths of the compressed plate with a) one-bubble shaped imperfection ($\varepsilon_1 \neq 0$, $\varepsilon_2 = 0$), two-wave shaped imperfection ($\varepsilon_1 = 0$, $\varepsilon_2 \neq 0$) and b) combined imperfection ($\varepsilon_1 \neq 0$, $\varepsilon_2 \neq 0$)

5.3 Erosion of the buckling load of a compressed cylindrical panel

The last example highlights the well-known fact that compressed shell structures are exceptionally sensitive to initial imperfections. The test example, depicted in Fig. 7 is taken from [14]. One quarter of a simply supported cylindrical panel is discretised and analyzed using both shallow shell BFS elements and semi-loof elements. Using Fried's arc-length method the equilibrium paths shown in Fig. 7 are obtained.

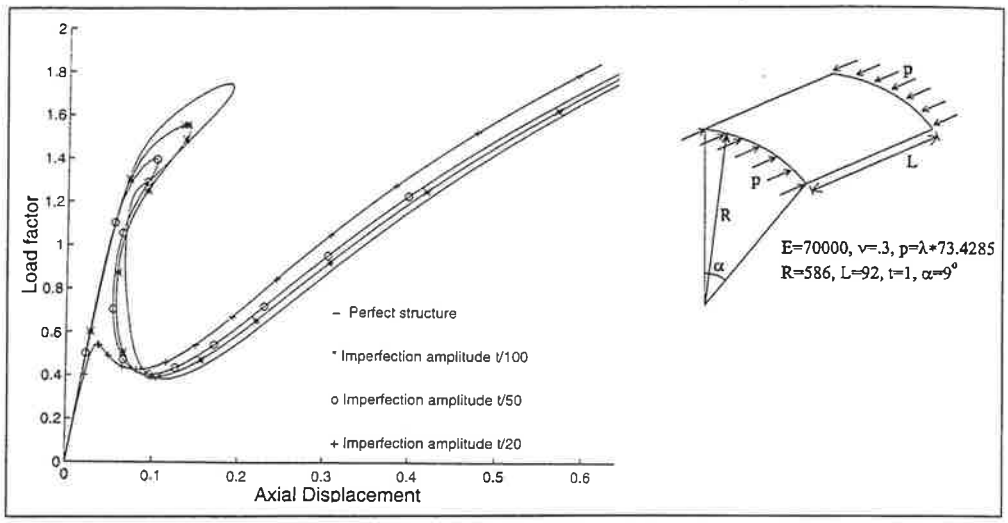


Figure 7. Perfect and imperfect cylinder responses under uniform compression.

Simulating the real behavior of the structure, we analyze the cylindrical panel with an initial geometric imperfection w_0 whose shape is equal to the first buckling mode of the structure:

$$w_0=\alpha\varphi,$$

where the parameter α receives values t/100, t/50 and t/20, in which t is the shell thickness. The associated equilibrium paths, shown in Fig. 7 give the buckling loads 1.55, 1.40 and 0.54, respectively. Compared to the buckling load of the perfect structure, $\lambda_{cr}=1.75$, the structure can be seen to be extremely sensitive to geometric imperfections: the imperfection with amplitude 5% of the cylinder thickness, decreases the buckling load as much as 69%.

ACKNOWLEDGMENT

Concerning the first author, this work is funded by the Finnish Ministry of Education via the national post-graduate school of technical mechanics.

REFERENCES

1. J.H.Argyris & H.Balmer & M.Kleiber & U.Hindenlang, Natural description of large inelastic deformations for shells of arbitrary shape - application of TRUMP element, *Comp. Meth. Appl. Mech. Engng.*, **22** (1980), 361-389.

2. P.G.Bergan & G.Horrigmoe & B.Kråkeland & T.H.Søreide, Solution techniques for non-linear finite element problems, *Int. J. Num. Meth. Eng.*, **12** (1978), 1677-1696.

3. F.K.Bogner & R.L.Fox & L.A.Schmit, The generation of inter-element-compatible stiffness and mass matrices by the use of interpolation formulas. In Matrix Methods in Structural Mechanics, AFFDL-TR-66-80, (1966), 397-443.

4. E.G.Carnoy & T.J.R.Hughes, Finite element analysis of the secondary buckling of a flat plate under uniaxial compression, *Int. J. Non-linear Mechanics*, **18** (1983), 167-175.

5. I.Fried, Orthogonal trajectory accession to the nonlinear equilibrium curve, *Comp. Meth. Appl. Mech. Eng.*, **47** (1984), 283-297.

6. J. Huitfeldt, Nonlinear eigenvalue problems - prediction of bifurcation points and branch-switching, *Numerical analysis report 17*, Department of Computer Sciences, Chalmers University of Technology, Göteborg (1991).

7. B.M. Irons, The Semiloof shell element, in Finite Elements for Thin Shells and Curved Members, Eds. D.G. Ashwell and R.G. Gallagher, Ch. 11, Wiley, Chichester (1976).

8. R.B.Kearfott, Some general bifurcation techniques, *SIAM J. Sci. Stat. Comput.*, **4** (1983), 52-68.

9. H.B.Keller, Numerical solution of bifurcation and nonlinear eigenvalue problems, *Applications of Bifurcation Theory*, Rabinowitz P. (ed.), Academic Press, New York, 359-384 (1977).

10. R.Kouhia & M.Mikkola, Tracing the equilibrium path beyond simple critical points, *Int. J. Num. Meth. Eng.*, **28** (1989), 2923-2941.

11. R.A.F Martins and C.A.M. Oliveira, Semi-loof shell, plate and beam elements - new computer versions: Part 1. Elements Formulation, *Eng. Comput.*, **5** (1988), 15-25.

12. S.Pajunen & M.Tuomala, Calculation of equilibrium paths in nonlinear structural analysis, *Journal of Structural Mechanics (Rakenteiden Mekaniikka)*, **30** (1997), 63-84.

13. E.Reissner, On finite axi-symmetrical deformations of thin elastic shells of revolution, *Comp. Mech.*, **4** (1989), 387-400.

14. E.Riks & F.A.Brogan & C.C.Rankin, Numerical aspects of shell stability analysis, *Proc. International Conference of Computational Engineering Science*, Krätzig W.B. & Oñate E. (eds.), Springer-Verlag, Heidelberg, 125-151 (1990).

15. M.Stein, Loads and deformations of buckled rectangular plates, *NASA TR* R-40 (1959).

16. M.Tuomala, Eräiden yksinkertaisten rakenteiden staattisen ja dynaamisen vasteen analysointi elementtimenetelmällä. *Teknillinen korkeakoulu, rakennetekniikan laitoksen julkaisuja 30,* Otaniemi (1980).

17. M.Uemura & O.Byon, Secondary buckling of a plate under uniaxial compression. Part 2. Analysis of clamped plate by F.E.M. and comparison with experiments, *Int. J. Non-linear Mechanics*, **13** (1978), 1-14.

18. W.Wagner & P.Wriggers & E.Stein, A shear-elastic shell theory and finite-element-postbuckling analysis including contact, *Post-Buckling of Elastic Structures*, Szabó J. (ed.), Akadémiai Kiadó, Budapest, 381-404 (1986).

19. P.Wriggers & J.C.Simo, A general procedure for the direct computation of turning and bifurcation points, *Int. J. Num. Meth. Eng.*, **30** (1990), 155-176.

Using iterative linear solvers in non-linear continuation algorithm

Reijo KOUHIA
Laboratory of Structural Mechanics
Helsinki University of Technology
P.O. Box 2100, 02015 HUT, Finland

ABSTRACT

Some techniques to solve non-linear algebraic equations in finite element analysis are considered. Especially the use of preconditioned iterative linear equation solvers in path-following algorithms is described and the choice of an accelerator iteration and the preconditioner are discussed. Some preliminary results are presented from structural analyses.

1 INTRODUCTION

Solutions of non-linear structural problems under quasi-static loading conditions are usually obtained by using an arc-length type continuation procedure. In these algorithms an additional constraint equation is augmented to the equilibrium equation system. This constraint destroys the symmetry and the banded form of the tangent stiffness matrix. In order to utilize the specific storage format of the tangent matrix, the solution of the constrained system is usually obtained by using the block factorization scheme, where the tangent matrix is factorized and the system is solved with two right-hand side vectors.

In large problems the decomposition time and the storage requirements will be prohibitively high when Gaussian elimination type factorizations are used. Special sparse matrix techniques have been developed which try to minimize the fill in during the decomposition. However, these techniques need reordering of the unknowns and thus are not well parallelizable and vectorizable. Iterative methods seems to be ideal for modern vector and parallel computers to solve systems of linear equations. For large problems they require much less storage than the direct solvers and computing times are also in many cases reduced.

There are at least two main factors which have contributed to the slow spread of iterative linear solvers in non-linear structural problems. Firstly, the computational cost of the incremental and iterative methods is usually so high that the size of the discretized non-linear problems has to be much smaller than it is possible to do in

linear analysis. Therefore storage requirements are not so critical issue. In a practical non-linear analysis the decomposition time is usually comparable to the time needeed to form the global matrices and internal force vectors. Secondly, the stiffness matrix might be ill conditioned and not necessarily positive definite in certain parts of the solution path. In addition, some important control parameters of the continuation process, like the determinant or the lowest eigenvalue of the tangent stiffness matrix, are not easily accessible when iterative solvers are used.

The block factorization strategy is not feasible if the solution of the linear system is obtained with an iterative solver. Usually the solution is carried out directly to the augmented unsymmetric system. Krenk and Hededal [1] have recently introduced an orthogonal or a dual orthogonal residual method where this block factorization type of solution is not needed. In this paper these procedures are coped with iterative linear equation solvers and the performance is compared to that of usual practice. Some example problems both in non-linear heat conduction and solid mechanics are solved.

2 CONTINUATION ALGORITHM

Discretization of the quasi-static equilibrium equations expressing the balance between external and internal forces results in an equation of the form:

$$\boldsymbol{f}(\boldsymbol{q},\lambda) \equiv \lambda \boldsymbol{p}_r - \boldsymbol{r} = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{p}$ is the external load vector. If the finite element method is used in the discretization process, the internal force vector $\boldsymbol{r}$ follows from the assembly operation of the element contributions

$$\boldsymbol{r}^{(e)} = \int_{V^{(e)}} \boldsymbol{B}^T \boldsymbol{s} dV.$$

The vector $\boldsymbol{s}$ contains the stress components. The strain-displacement matrix $\boldsymbol{B}$ is defined by

$$\delta \boldsymbol{e} = \boldsymbol{B} \delta \boldsymbol{q},$$

where the column vector $\boldsymbol{e}$ contains the strain components.

Usually the applied loading is assumed to depend linearily on a single parameter, i.e. the load parameter λ, such that $\boldsymbol{p} = \lambda \boldsymbol{p}_r$, where $\boldsymbol{p}_r$ is the reference load vector.

Solution of the equation system (1) forms a one-dimensional equilibrium curve in an $N+1$ dimensional displacement-load parameter space, where N is the dimension of the state space, i.e. the number of dof's in the vector $\boldsymbol{q}$. Procedures to trace the one dimensional equilibrium path defined by equation (1) are called continuation or path following methods. They are incremental or step-wise algorithms. A typical continuation step includes the predictor and the corrector phases.

To traverse a solution path a proper parametrization is needed. Simple load control is the oldest type of parametrization. It is usually the most efficient one in regular parts of a path, and the adaptation of an iterative linear equation solver in it is straightforward. However, near the so called limit points, where the structure loses its load carrying capacity (at least locally), it might break down. At the limit point the tangent stiffness matrix is singular and the load parameter is decreasing after such a point. A

remedy is to change the control from the load parameter to some of the displacement components. Selecting the controlling displacement (or component from the scaled vector containing both displacements and the load parameter) to be the largest one from the last converged increment, results in a simple and reliable continuation procedure [2]. Non-dimensionalizing of the variables is an essential point of this method. Nevertheless, it is recommendable for all other procedures, too.

A usual setting of a continuation process is to augment the discrete equilibrium equations with a single constraint equation c in the following form:

$$\boldsymbol{g}(\boldsymbol{q},\lambda) = \left\{ \begin{array}{rcl} \boldsymbol{f}(\boldsymbol{q},\lambda) & = & \boldsymbol{0} \\ c(\boldsymbol{q},\lambda) & = & 0 \end{array} \right. . \tag{2}$$

This kind of procedures are also commonly called arc-length methods. A large class of constraint equations can be written in the form

$$c(\boldsymbol{q},\lambda) = \boldsymbol{t}^T \boldsymbol{C}\boldsymbol{n} - c_0 = 0$$

where $\boldsymbol{t}$ and $\boldsymbol{n}$ are $N+1$ dimensional vectors and c_0 is a scalar. For explicit expressons of the vectors $\boldsymbol{t}$ and $\boldsymbol{n}$ see ref. [3]. The weighting matrix $\boldsymbol{C}$ can be partitioned as $\mathrm{diag}(\boldsymbol{W}, \alpha^2)$, where $\boldsymbol{W}$ is a positive definite or semidefinite diagonal matrix corresponding to displacements and α is a scaling factor.

Using the Newton-Raphson linearization on the extended equation system (2) results in

$$\left\{ \begin{array}{rcl} \dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{q}}\delta\boldsymbol{q} + \dfrac{\partial \boldsymbol{f}}{\partial \lambda}\delta\lambda + \boldsymbol{f}(\boldsymbol{q},\lambda) & = & -\boldsymbol{K}\delta\boldsymbol{q} + \boldsymbol{p}_r\delta\lambda + \boldsymbol{f} \;=\; \boldsymbol{0} \\ \dfrac{\partial c}{\partial \boldsymbol{q}}\delta\boldsymbol{q} + \dfrac{\partial c}{\partial \lambda}\delta\lambda + c(\boldsymbol{q},\lambda) & = & \boldsymbol{c}^T\delta\boldsymbol{q} + e\delta\lambda + c \;=\; 0 \end{array} \right. . \tag{3}$$

Usually in structural analyses the tangent stiffness matrix $\boldsymbol{K}$ is symmetric. Therefore, in order to utilize the specific sparsity pattern and symmetry of the tangent stiffness matrix, the solution of the augmented equations (3) is usually performed by using the following three phase block elimination method, also known as bordering algorithm [2], [4], [5]:

1. solve $\boldsymbol{K}\delta\boldsymbol{q}_f = \boldsymbol{f}$ and $\boldsymbol{K}\boldsymbol{q}_p = \boldsymbol{p}_r$,
2. compute $\delta\lambda = -(c + \boldsymbol{c}^T\delta\boldsymbol{q}_f)/(e + \boldsymbol{c}^T\boldsymbol{q}_p)$,
3. update $\delta\boldsymbol{q} = \delta\boldsymbol{q}_f + \delta\lambda\boldsymbol{q}_p$.

In this format the solution of the linear equation system at phase 1 is performed by means of direct solvers. If iterative solvers are used, the nonsymmetric sparse format of the equation system (3):

$$\boldsymbol{H}\delta\boldsymbol{y} = \boldsymbol{h}, \quad \boldsymbol{H} = \left[\begin{array}{cc} \boldsymbol{K} & -\boldsymbol{p}_r \\ \boldsymbol{c}^T & e \end{array} \right], \quad \delta\boldsymbol{y} = \left\{ \begin{array}{c} \delta\boldsymbol{q} \\ \delta\lambda \end{array} \right\}, \quad \boldsymbol{h} = \left\{ \begin{array}{c} \boldsymbol{f} \\ -c \end{array} \right\}, \tag{4}$$

seems to be more appropriate, see refs. [6], [7]. Chan and Saad [8] have studied different preconditioning techniques in the non-linear elliptic second order problem

$\Delta u + \lambda \exp(u) = 0$ on unit square with zero Dirichlet boundary conditions, discretized by a standard five point finite difference formula. One of their conclusions is that, when a preconditioning is available, it seems best to work directly with an iterative method on the unsymmetric form (4).

Usually the question of the "best choice" of the constraint equation is overemphasized in the engineering literature, for example see discussion in ref. [9]. However, the specific form of the constraint equation is a relevant topic in the present context. A procedure which fits well together with the iterative linear equation solvers, is the orthogonal residual procedure by Krenk and Hededal [1], which does not require the block factorization process and thus only one solution of linear equation is needed per iteration.

As argued by Krenk and Hededal, the magnitude of the displacement increment is optimal when an orthogonality condition

$$\Delta \boldsymbol{q}^T \boldsymbol{f} = 0$$

is satisfied. This linear condition is used to determine the current load parameter λ. The algoritm can be described briefly as:

1. compute: $\boldsymbol{r}_i = \boldsymbol{r}(\boldsymbol{q}_0 + \Delta \boldsymbol{q}_i)$, $\Delta \boldsymbol{r}_i = \boldsymbol{r}_i - \lambda_0 \boldsymbol{p}_r$, $\Delta \lambda_{i+1} = \Delta \boldsymbol{q}_i^T \Delta \boldsymbol{r}_i / \Delta \boldsymbol{q}_i^T \boldsymbol{p}_r$,
2. solve: $\boldsymbol{K} \delta \boldsymbol{q}_{i+1} = \boldsymbol{f}_{i+1} = (\lambda_0 + \Delta \lambda_{i+1}) \boldsymbol{p}_r - \boldsymbol{r}_i$,
3. update: $\Delta \boldsymbol{q}_{i+1} = \Delta \boldsymbol{q}_i + \delta \boldsymbol{q}_{i+1}$.

λ_0 and $\boldsymbol{q}_0$ denote the load level and the displacement vector at the beginning of current increment. However, even if the algorithm seems to be ideally suited for the use of an iterative linear equation solver, it has some drawbacks observed in numerical experiments. Since the size of the increment is not restricted during the iteration, the algorithm seems to have some tendency of increasing the size of the displacement increment near limit points.

3 PRECONDITIONED ITERATIVE METHODS

3.1 Krylov subspace methods

In the sequel, a generic linear equation system will be denoted by

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where the coefficient matrix $\boldsymbol{A}$ can be symmetric or unsymmetric. An equivalent preconditioned system is

$$\boldsymbol{M}_1^{-1} \boldsymbol{A} \boldsymbol{M}_2^{-1} \boldsymbol{y} = \boldsymbol{M}_1^{-1} \boldsymbol{b},$$

where $\boldsymbol{M} = \boldsymbol{M}_1 \boldsymbol{M}_2$ is the preconditioning matrix and $\boldsymbol{M}_1, \boldsymbol{M}_2$ are the left- and right preconditioning matrices, respectively. In practice this split form is not always needed. It is usually possible to rewrite the iterative method in a way that only a computational

step: solve $\boldsymbol{u}$ from $\boldsymbol{Mu} = \boldsymbol{v}$, is necessary, so the preconditioner applies in its entirety. The question of preconditioning will be discussed briefly in the next section.

Krylov subspace methods seem to be among the most important iterative techniques available for solving large linear systems [10], [11], [12]. These techniques are based on projections onto Krylov subspaces, which are subspaces spanned by vectors which are obtained recursively by multiplying the previous residual with the matrix: i.e.

$$\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{r}_0) = \text{span}\left\{\boldsymbol{r}_0, \boldsymbol{A}\boldsymbol{r}_0, \boldsymbol{A}^2\boldsymbol{r}_0, ..., \boldsymbol{A}^{m-1}\boldsymbol{r}_0\right\},$$

where $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$. Approximate solution of the system is found from a m-dimensional subspace $\boldsymbol{x}_0 + \mathcal{K}_m$ by imposing the Petrov-Galerkin condition requiring the residual to be orthogonal to another m-dimensional subspace $\mathcal{L}_m$.

The most wellknown Krylov subspace method is the preconditioned conjugate gradient (PCG) method for symmetric positive definite (SPD) matrices. There are many different implementations of the PCG-iteration, but the following algorithm is perhaps the most common: construct $\boldsymbol{M}$, initialize $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$, solve $\boldsymbol{M}\boldsymbol{d}_0 = \boldsymbol{r}_0$, compute $\tau_0 = \boldsymbol{r}_0^T\boldsymbol{d}_0$ and iterate $i = 0, 1, 2, ...$ until convergence:

1. compute: $\boldsymbol{s} = \boldsymbol{A}\boldsymbol{d}_i, \quad \alpha_i = \tau_i / \boldsymbol{d}_i^T\boldsymbol{s},$
2. update: $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \alpha_i\boldsymbol{d}_i, \quad \boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \alpha_i\boldsymbol{s},$
3. solve: $\boldsymbol{M}\boldsymbol{z} = \boldsymbol{r}_{i+1}$ and compute $\tau_{i+1} = \boldsymbol{r}_{i+1}^T\boldsymbol{z}, \quad \beta_i = \tau_{i+1}/\tau_i,$
4. update $\boldsymbol{d}_{i+1} = \boldsymbol{z} + \beta_i\boldsymbol{d}_i.$

It is a Galerkin (orthogonal projection) type Krylov subspace method, i.e. $\mathcal{L}_m = \mathcal{K}_m$. One iterate of the PCG method requires one matrix-vector product, five[1] level-1-operations and one solution of linear equations $\boldsymbol{M}\boldsymbol{z} = \boldsymbol{r}$.

If the matrix $\boldsymbol{A}$ is symmetric but indefinite the PCG-algorithm can become unstable and even break down. Paige and Saunders [13] were the first to devise stable algorithms for symmetric indefinite systems. These two algorithms called SYMMLQ and MINRES are based on Lanczos tridiagonalization, which exists also in indefinite case.

For unsymmetric matrices the situation is much more complex. The CG method for symmetric and positive definite systems has two important properties. It is based on three term recurrence, and it minimizes the error with respect to the energy norm. Unfortunately these two properties can only be fulfilled for nonsymmetric CG-type schemes for a very limited class of matrices, namely the shifted and rotated Hermitean matrices. In this paper only those algorithms are considered which retain the short recurrencies thus being more favourable with respect to memory requirements. Biconjugate gradient (BCG) type algorithms are based on the Lanzcos biorthogonalization algorithm which builds a pair of biorthogonal bases for the two subspaces $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{r}_0)$ and $\mathcal{K}_m(\boldsymbol{A}^T, \tilde{\boldsymbol{r}}_0)$. In the numerical examples the following BCG-type methods are used in this study: biconjugate gradient squared (CGS), biconjugate gradient stabilized (BI-CGSTAB), quasi-minimal residual (QMR) and transpose free QMR (TFQMR). For a unified general description of all these methods with numerous references see ref. [14].

[1]PCG requires an additional norm evaluation if the convergence is checked from the residual $\boldsymbol{r}$.

3.2 Preconditioning

It is well known that the performance of iterative solvers depends on the eigenvalue distribution and on the possible non-normality of the coefficient matrix. These problems can be avoided, at some extent, by employing a preconditioner. It seems to be generally agreed that the choice of the preconditioner is even more critical than the choice of the type of the Krylov subspace iteration [15].

There are two major conflicting requirements in the development of a preconditioned iteration, namely the construction[2] and use of a preconditioner should be cheap and its resemblance with matrix $\boldsymbol{A}$ should be as close as possible. The most general preconditioning strategies can be grouped into classes:

1. preconditioners based on classical iterations like Jacobi, SSOR,
2. incomplete sparse LU-decompositions (ILU or IC for symmetric matrices),
3. polynomial preconditioners,
4. explicit sparse approximate inverse preconditioners,
5. multigrid or multilevel preconditioners.

Incomplete factorization is perhaps the most wellknown strategy. There are many variants of ILU-decompositions differing, for instance on the way how the nonzero pattern of the preconditioner is defined. The simplest strategy is to have the same nonzero pattern for the $\boldsymbol{L}$ and $\boldsymbol{U}$ factors as $\boldsymbol{A}$. This incomplete factorization known as ILU(0) is easy and inexpensive to compute, but often leads to a crude approximation resulting in many iterations in the accelerator to converge. Several alternative ILU factorizations have been developed in which the fill-in is determined by either using the concept of level of fill or by a treshold strategy where the nonzero pattern of the preconditioner is determined dynamically neglecting small elements in the factorization.

Meijerik and Van der Vorst [16] proved existence of the ILU factorization for arbitrary fill patterns if the coefficient matrix is a M-matrix[3]. This is often the case, e.g. matrices arising from discretizations of the heat equation. However, matrices arising from problems in structural mechanics usually do not have this property. In order to circumvent this problem an additional reduction step has been introduced, where an M-matrix is determined from the stiffness matrix and the incomplete factorization scheme is applied to this [17].

Mathematical analysis reveals that for second-order elliptic boundary value problems the ILU(0) approach is asymptotically no better than the unpreconditioned iteration. More precisely, the condition number of the ILU preconditioned operator is of the same order as matrix $\boldsymbol{A}$. Several variants of the basic ILU have been presented in the literature e.g. MILU, RILU and DRILU (modified, relaxed and dynamically relaxed) [17]. However, in real engineering problems these modifield versions does not perform

[2] If the preconditioner is to be used many times more effort can be paid to its construction.

[3] A matrix is a M-matrix if its off-diagonal elements are nonpositive and all the elements of the inverse are positive.

any better than the basic ILU. Ajiz and Jennings [18] proposed the corrected IC factorization (CIC), [4] which guarantees a positive definite preconditioner if the matrix itself is SPD.

It should be remembered that the effectiveness of a preconditioning strategy is highly problem and architecture dependent. For instance, incomplete factorizations are difficult to implement on high-performance computers, due to the sequential nature of the triangular solves. On the other hand, sparse approximate inverse preconditioning needs only matrix-vector products, which are relatively easy to vectorize and parallelize, but they are usually not as robust as ILU-factorization based strategies [15].

For second-order elliptic PDE's discretized by low order finite elements many of the listed preconditioning techniques can be used. However, for finite element models of thin-shells only the corrected incomplete factorization allowing some degree of fill-in [18], [19] or a multilevel preconditioner [20] seems to be the only reasonable choices.

For a certain type of a preconditioning technique, the computational complexity can be reduced. Construction of a preconditioning matrix $\boldsymbol{M}$ in a form

$$\boldsymbol{M} = (\tilde{\boldsymbol{D}} + \boldsymbol{E})\hat{\boldsymbol{D}}(\tilde{\boldsymbol{D}} + \boldsymbol{F}), \tag{5}$$

where $\tilde{\boldsymbol{D}}, \hat{\boldsymbol{D}}$ are diagonal matrices and $\boldsymbol{E}$ and $\boldsymbol{F}$ are the strictly lower and upper parts of $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{A}) + \boldsymbol{E} + \boldsymbol{F}$, allows implementation of the preconditioned CG or Bi-CG-type methods in which the computational labor is comparable to the unpreconditioned case. This strategy is due to Eisenstat [21], and it is commonly called as the Eisenstat trick, see also refs. [10],[12]. Unfortunately the usefulness of this stratgy is somewhat limited. For a very sparse matrices, such as resulting from a low order FE discretizations of the diffusion equation, the triangular solution including short rows is the main bottleneck in a typical supercomputer implementation. Also quality of the split-preconditioners (5), which can be used in the Eisenstat trick is not good enough in shell problems.

4 EXAMPLES

In this section some example problems are solved and the performances of the iterative methods are compared to a direct solution procedure. The direct solver used is a slightly modified version from ref. [22], pages 327-329. All computations have been preformed on Digital Alpha Server 8400 [5] using double precision representation for real numbers. The program is written in Fortran 77 and the level 3 optimization flag is used in the compilation for most, including all linear algebra routines.

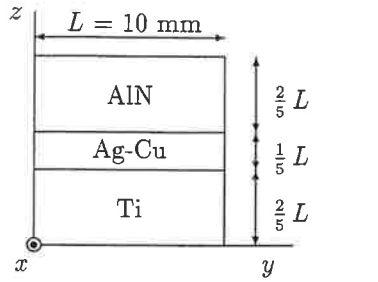
Convergence of the iteration is checked by the relative and absolute residual error and declared if

$$\|\boldsymbol{r}_i\|_2/\|\boldsymbol{b}\|_2 < RTOL \quad \text{or} \quad \|\boldsymbol{r}_i\| < ATOL,$$

except computations where the Eisenstat trick is used. In that case the absolute criteria is used and the measure is the weighted Euclidean norm with the preconditioner as the weight matrix.

[4] The name corrected incomplete Cholesky is adopted from ref. [19].

[5] Center for Scientific Computing, Espoo, Finland.



	Ti	Ag-Cu	AlN
E/GPa	118.6	94.4	315
ν	0.36	0.36	0.25
$10^6\alpha$	6.441	17.995	2.85

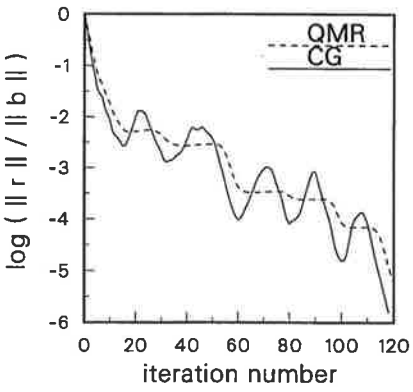


Figure 1: Three material solid block: (a) geometry and material data, (b) convergence behaviour of CG and symmetric QMR iteration with IC(0) preconditioner $L/h = 20$ (h is the sidelength of an element).

4.1 Three material elastic block

An elastic block composed by three material layers and occupying the region (in cartesian coordinates) $0 < (x, y, z) < L$ is considered. The material interfaces are horizontal layers parallel to the xy-plane, and having positions $z = \frac{2}{5}L$ and $z = \frac{3}{5}L$. The stack models a ceramic (AlN) to metal (Ti) joint brazed together with Ag-Cu filler alloy. Constitutive parameters used for these materials are shown in fig. 1.

Uniform meshes with eight node trilinear brick elements are used. The only loading is the temperature change defined by $\Delta T = \Delta T_0 xyz/L^3$. Minimal constraints which prevent the rigid body motion are imposed. Convergence tolerancies used are $ATOL = RTOL = 10^{-6}$.

Some data of the stiffness matrix is shown in table 1 and a comparison of the performance of the preconditioned CG iteration with a direct in-core skyline solver is recorded in table 2. The SSOR preconditioning is implemented by using the Eisenstat trick. In this case the convergence is measured in the weighted norm $\|\boldsymbol{x}\| = (\boldsymbol{x}^T \boldsymbol{M}\boldsymbol{x})^{1/2}$ and thus the tabulated values are not comparable to those of IC(0) preconditioning. Compressed row storage format is used to store the nonzero elements of the matrix (also for the IC(0) preconditioner).

Convergence behaviour for the conjugate gradient method exhibits some oscillations which are absent if the material characteristics are uniform. The symmetric QMR iteration [23] with coupled two term recurrence shows much smoother convergence than the CG method, see fig. 1.

As expected, the solution times for the direct solvers become quickly untolerably high due to the large bandwidth which grows like $B \sim N^{2/3}$. For large 3-D problems iterative solvers are the only possible way to get the solution at reasonable cost.

Table 1: Three material block, stiffness matrix characteristics.

L/h	N	B_{rms}	M	NZ	$2NZ/M$
5	642	117	70552	18615	0.528
10	3987	380	1455352	135915	0.187
20	27777	1355	36805402	1035165	0.056
40	206757	5105	1043093302	8075265	0.015

N number of unknowns
B_{rms} root-mean square bandwidth
M number of elements under envelope
NZ number of nonzero elements

Table 2: Three material block, solution times

	direct solver		CG-IC(0)				CG-SSOR Eisenstat†			
L/h	F time	B time	iter	P time	I time	ratio	ω	iter	I time	ratio
5	0.11	0.01	25	0.02	0.07	1	0.8	22	0.04	3
10	8.3	0.2	57	0.2	1.3	6	1.2	41	0.51	17
20	1660.	6.	120	1.6	39.8	40	1.4	77	15.0	111
40	97 hours*	140.*	257	12.7	697.	490*	1.7	195	324.	1080*

F time = Factorization time in seconds
B time = Backsubstitution + load vector reduction time
P time = Preconditioner construction time
I time = Iteration time
ratio = (F+B)/(P+I)
* = estimated value, † = convergence measured in weighted norm

4.2 Pinched cylinder

A well known shell element test is the pinched cylinder, see e.g. ref. [24]. Length of the shell equals to its diameter ($L = 2R$) and the Poisson's ratio is 0.3. Performance of the conjugate gradient method is studied with respect to the relative thickness and some relevant parameters in the finite element model. As expexted, the problem gets harder when the thickness to radius ratio, i.e. the characteristic thickness gets smaller. Here results of only the cases $t/R = 10^{-2}$ and 10^{-3} are reported.

The shell elements are facet type 3-node triangular or 4-node quadrilateral elements, with drilling rotations using the Hughes-Brezzi formulation [25]. The plate bending part of the element is based on the stabilized MITC theory [26]. In the MITC formulation the stabilization parameter has been 0.4 for both triangular and quadrilateral elements. One octant of the shell is discretized by uniform 30×30 mesh resulting in 5489 unknowns. Strict convergence tolerance is used $RTOL = 10^{-9}$ and only the relative criteria is active.

Value of the regularizing penalty parameter γ used in the formulation of Hughes and Brezzi has a notable effect on the convergence of the conjugate gradient iteration.

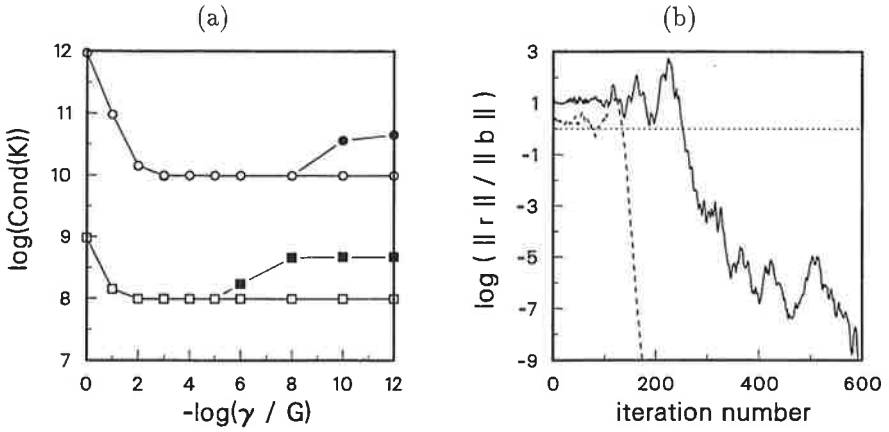


Figure 2: (a) Effect of the regularizing parameter γ on the spectral condition number of the stiffness matrix. Solid markers correspond to stiffness matrices without Allman displacement field. Upper two curves $t/R = 10^{-3}$ and lower ones $t/R = 10^{-2}$. (b) Convergence plot of the CG-IC(0) iteration: $\gamma = G$ (without Allman field), solid line $t/R = 10^{-3}$ and dashed line $t/R = 10^{-2}$. Quadrilateral 30×30 mesh.

It affects the spectral condition number of the stiffness matrix, see fig. 2a. Due to the ill-conditioning of the thinner shell problem the convergence of the PCG iteration slows down at the level of 10^{-4} in the relative error, see fig. 2b. Adding Allman type displacement field [27], [28] to the in-plane interpolation has also an effect on the convergence, however, there seems to be no definite trend on that dependency. The number of iterations needed to convergence are shown in table 3.

The IC factorization does not exist for all values of the γ parameter. A simple remedy is to use the shifting strategy of Manteuffel [29] where the matrix $\boldsymbol{A} + \rho \mathrm{diag}(\boldsymbol{A})$ is factorized instead of $\boldsymbol{A}$. However, the quality of such a preconditioner is not very good as can be seen from table 4.

4.3 Non-linear analysis of cylindrical panel

A shallow cylindrical shell subjected to a central point load on the convex side is a common test problem of path-following algorithms, see e.g. ref. [30]. The longitudinal boundaries are immovable, whereas the curved edges are completely free. The problem data are: radius $R = 2540$ mm, length of the straight hinged edge $L = 508$ mm, Young's modulus $E = 3.10275$ GPa, Poisson's ratio $\nu = 0.3$ and $\theta = 0.1$ rad. Two values for the thickness are used: $t = 12.7$ mm $(R/t = 200)$ and $t = 6.35$ mm $(R/t = 400)$. Uniform 32×32-mesh with DKT-elements is used in the simulations resulting in 6175 dof for a quadrant of the panel. Drilling rotations are included by the Hughes-Brezzi formulation and the γ-parameter has the value $\gamma = 0.026G$. Allman-type interpolation is also included.

First, some comparisons with the IC(0) and the CIC(ψ) preconditioners for the linear problem are performed. Also a pure treshold version of the IC preconditioner is

Table 3: Influence of the regularizing parameter γ on the convergence of the CG-IC(0) iteration, 30×30 mesh with MITC elements. A = Allman type amendment for the in-plane interpolation.

(a) $t/R = 10^{-2}$

γ/G	Q-A	T-A	Q	T
1.0	-	-	175	293
0.5	-	-	153	283
0.25	-	306	142	273
0.2	379	301	138	269
0.1	147	288	129	259
0.01	120	262	117	2337
0.001	118	249	115	1201

(b) $t/R = 10^{-3}$

γ/G	Q-A	T-A	Q	T
1.0	-	-	593	-
0.5	-	-	511	-
0.25	-	-	502	-
0.2	-	-	508	-
0.1	356	-	397	1197
0.01	307	768	330	743
0.001	287	674	284	636

Table 4: Influence of the shift ρ on the convergence, 30 × 30 mesh with Allman type interpolation, $\gamma = G$, quadrilateral MITC elements with $t/R = 10^{-2}$.

shift	0.035	0.040	0.045	0.050	0.055	0.060	0.065	0.075	0.100
iterations	2017	1265	947	825	797	805	825	875	1000

tested, where the diagonal corrections are omitted. This preconditioner is abbreviated as IC(ψ). Behaviour of the CIC(ψ) and IC(ψ) methods are tested with different drop tolerancies ψ and the preconditioner sizes are also recorded in table 5. Convergence tolerancies have been $RTOL = ATOL = 10^{-5}$. It should be noted that the IC(0) factorization needs a small shift ($\rho = 0.005$) as well as the IC(ψ) method, where the optimal shift depend on the drop tolerance ψ. This is an annoying feature, since there is no known method to determine the optimal or near optimal shift a priori.

The computing times for almost all cases shown in table 5 are higher than the solution time needed for the direct solution, worst case almost by factor 10. However, there are some potential of using CIC or IC with a low drop-tolerance in the non-linear analysis, if the preconditioner need not to be computed at every time when the stiffness matrix is formed.

Five continuation strategies are compared. Symmetric formulations use the orthogonal residual method with direct or iterative linear equation solver. For consistently linearized elliptical constraint, the symmetric formulation uses only direct solver and with the augmented (4) nonsymmetric forms both direct and iterative solvers are used. Only the full Newton-Raphson strategy is used in the computations, even it is not necessary for all parts of the continuation paths. Load-deflection curves are shown in fig. 3 and the iteration characteristics from the computations of the thicker shell are recorded in table 6. Same conclusions can also be drawn from the thinner case.

The results in the linear case for the Jacobi and IC(0) strategies for preconditioning can be directly generalized to the non-linear analysis. Only the strategy, where the treshold IC preconditioner (with drop tolerance $\psi = 10^{-3}$ and shift $\rho = 10^{-3}$) is

Table 5: Comparison of the shifted IC(0), Jacobi and CIC(ψ) and IC(ψ) preconditioners on the pinched cylindrical panel. P = preconditioner evaluation time, I = iteration time, t-r = (P+I)-times/direct solution time, m-r = memory ratio: $NZ(\boldsymbol{M})/NZ(\boldsymbol{A})$. Solution time of the direct solver is 2.6 s.

$R/t = 200$

method	shift	iter	P	I	t-r	$NZ(\boldsymbol{M})$	m-r
Jacobi	-	1935	0.0	18.7	7.2	6175	0.05
IC(0)	$5 \cdot 10^{-3}$	184	0.1	4.3	1.7	127095	1.00
CIC(1.0)	-	1995	0.3	22.7	8.8	6175	0.05
CIC(10^{-1})	-	1079	0.4	14.2	5.6	18036	0.15
CIC(10^{-2})	-	299	0.7	6.2	2.7	93349	0.74
CIC(10^{-3})	-	118	1.3	3.9	2.0	200689	1.6
CIC(10^{-4})	-	51	2.7	3.2	2.3	391606	3.1
CIC(10^{-5})	-	23	5.1	3.0	3.1	713691	5.6
CIC(10^{-6})	-	9	7.0	1.8	3.4	1010602	8.0
IC(10^{-1})	$2 \cdot 10^{-2}$	388	0.4	5.6	2.3	40010	0.31
IC(10^{-2})	$5 \cdot 10^{-3}$	78	0.7	1.7	0.9	102486	0.81
IC(10^{-3})	10^{-3}	39	1.4	1.4	1.1	226936	1.8
IC(10^{-4})	10^{-4}	19	2.9	1.5	1.7	455261	3.6

$R/t = 400$

method	shift	iter	P	I	t-r	$NZ(\boldsymbol{M})$	m-r
Jacobi	-	1924	0.0	18.7	7.2	6175	0.05
IC(0)	$5 \cdot 10^{-3}$	181	0.1	4.3	1.7	127095	1.00
CIC(1.0)	-	1957	0.3	23.0	9.0	6175	0.05
CIC(10^{-1})	-	1023	0.4	13.7	5.4	18496	0.15
CIC(10^{-2})	-	304	0.7	6.1	2.6	88230	0.69
CIC(10^{-3})	-	110	1.3	3.7	1.5	195314	1.5
CIC(10^{-4})	-	46	2.7	3.1	2.2	386823	3.0
CIC(10^{-5})	-	21	5.1	2.7	3.0	718339	5.7
CIC(10^{-6})	-	8	7.3	1.7	3.5	1030444	8.1
IC(10^{-1})	$2 \cdot 10^{-2}$	386	0.5	6.3	2.6	38905	0.31
IC(10^{-2})	$5 \cdot 10^{-3}$	76	0.7	1.6	0.9	98030	0.77
IC(10^{-3})	10^{-3}	38	1.4	1.3	1.0	215934	1.7
IC(10^{-4})	10^{-4}	18	2.8	1.3	1.6	434653	3.4

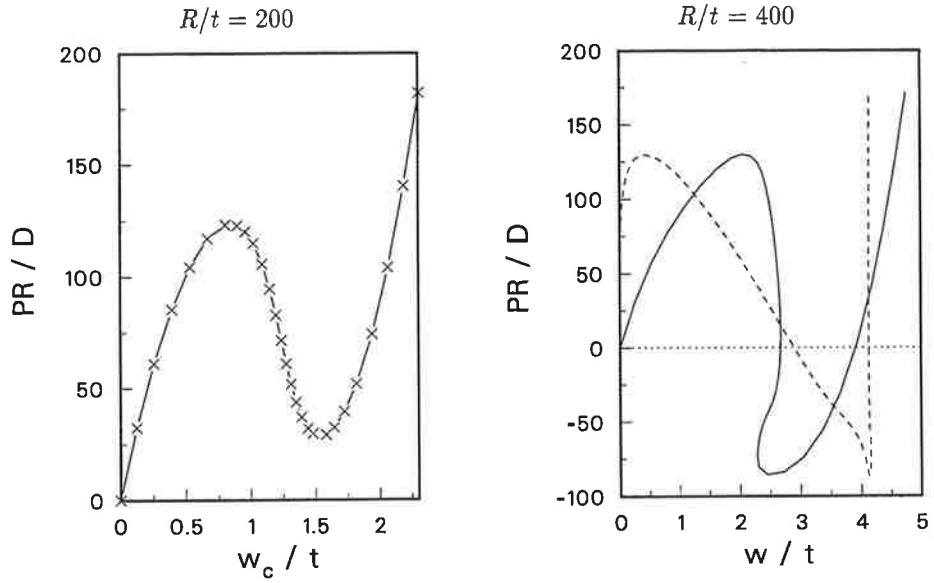


Figure 3: Hinged cylindrical panel, load deflection paths of the load point (solid line) and the free edge (dashed line).

Table 6: Iteration characteristics from the non-linear analysis of the pinched cylinder ($R/t = 200$); OR = Orthogonal Residual, E = Elliptic constraint.

constraint	solver	steps	smu	pu	CG-it	P/F	I/B	T
OR	direct	27	111	-	-	290	15	530
OR	CG-IC(0)	27	111	1	23462	0.1	497	711
OR	CG-IC(0)	27	111	27	20384	3	423	641
OR	CG-CIC(10^{-6})	27	111	1	4725	6	827	1047
OR	CG-CIC(10^{-6})	27	111	27	1037	184	182	580
OR	CG-IC(10^{-3})	27	111	1	9151	1.3	350	567
OR	CG-IC(10^{-3})	27	111	27	4316	35	167	417
OR	CG-IC(10^{-3})	27	111	111	4312	153	168	534
E block f.	direct	46	254	-	-	630	69	1227
E augm.	direct	46	254	-	-	1476	41	2098
E augm.	BI-CGSTAB-ILU(0)	46	254	46	73573	8.6	6287	6842

smu = stiffness matrix updates, pu = precondtitoner updates
CG-it = total number of CG-iterations
P/F = total (incomplete or full) factorization time
I/B = total CG-iteration or backsubstitution time
T = total run time

updated at the beginning of each increment, gives faster solution time than using the direct solver.

Using the augmented nonsymmetric equation system requires much more computing time than the block factorization with direct linear solver. All the tested nonsymmetric iterations CGS, BI-CGSTAB, QMR and TFQMR performed almost identically. Since only the no-fill ILU preconditioner is used in the nonsymmetric case, the figures in table 6 should not be compared to the symmetric iterations with treshold IC preconditioning.

It should be mentioned that a mixed strategy, where the stiffness matrix is updated when necessary, would be in favour of direct solvers.

5 CONCLUDING REMARKS

At present iterative methods for linear systems of equations have reached the level of robustness that they are included in almost every valued commercial FE-code. In despite of the excellent preformance of preconditioned Krylov subspace methods in heat transfer and stress analysis of solid bodies, they cannot be regarded as robust as direct solvers. Especially, for non-linear shell analysis the results with incomplete factorization preconditioners are still far from being satisfactory. Possibly the multilevel approach will change the affirmed state of affairs. Finally, it should be emphasized that the performance of the various procedures is highly problem and computer architecture dependent, for which reason fair comparison of different approaches is difficult.

REFERENCES

[1] S. Krenk and O. Hededal. A dual orthogonality procedure for non-linear finite element equations. *Computer Methods in Applied Mechanics and Engineering*, 123:95–107, 1995.

[2] W.C. Rheinboldt. *Numerical Analysis of Parametrized Nonlinear Equations.* Wiley, 1986.

[3] R. Kouhia and M. Mikkola. Strategies for structural stability analyses. In N.E. Wiberg, editor, *Advances in Finite Element Technology*, pages 254–278, 1995.

[4] H.B. Keller. The bordering algorithm and path following near singular points of higher nullity. *SIAM Journal on Scientific and Statistical Computing*, 4:573–582, 1983.

[5] K.H. Schweizerhof and P. Wriggers. Consistent linearization for path following methods in non-linear FE analysis. *Computer Methods in Applied Mechanics and Engineering*, 59:261–279, 1986.

[6] E.L. Allgower and K. Georg. *Numerical Continuation Methods - An Introduction.* Springer-Verlag, 1990.

[7] E. Barragy and C.F. Carey. A partitioning scheme and iterative solution for sparse bordered systems. *Computer Methods in Applied Mechanics and Engineering*, 70:321–327, 1988.

[8] T.F. Chan and Y. Saad. Iterative methods for solving bordered systems with applications to continuation methods. *SIAM Journal on Scientific and Statistical Computing*, 6(2):438–451, 1985.

[9] E. Riks. On formulatins of path-following techniques for structural stability analysis. paper presented at the European Conf. on New Advances in Computational Structural Mechanics, Giens, France, 1991.

[10] H. Voss. Iterative methods for linear systems of equations. University of Jyväskylä, Department of Mathematics, lecture notes 27, 1993.

[11] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, 1994.

[12] Y. Saad. *Iterative Methods for Sparse Linear Systems.* PWS Publishing, 1996.

[13] C.C. Paige and M.A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12:617–629, 1975.

[14] R. Barrett, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for iterative methods.* SIAM, 1994.

[15] M. Benzi and M. Tůma. Numerical experiments with two approximate inverse preconditioners, 1997. CERFACS TR/PA/97/11.

[16] J.A. Meijerink and H.A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Mathematics of Computation*, 31:148–162, 1977.

[17] P. Saint-Georges, G. Warzee, R. Beauwens, and Y. Notay. High-performance PCG solvers for FEM structural analysis. *International Journal for Numerical Methods in Engineering*, 39:1313–1340, 1996.

[18] M.A. Ajiz and A. Jennings. A robust incomplete Cholesky conjugate gradient algorithm. *International Journal for Numerical Methods in Engineering*, 20:949–966, 1984.

[19] P. Saint-Georges, G. Warzee, Y. Notay, and R. Beauwens. Fast iterative solvers for finite element analysis in general and shell analysis in particular. In B.H.V. Topping, editor, *Advances in Finite Element Technology*, pages 273–282, Edinburgh, 1996. Civil-Comp Press.

[20] V.E. Bulgakov. The use of the multi-level iterative aggregation method in 3-D finite element analysis of solid, truss, frame and shell structures. *Computers and Structures*, 63(5):927–938, 1997.

[21] S.C. Eisenstat. Efficient implementation of a class of preconditioned conjugate gradient methods. *SIAM Journal on Scientific and Statistical Computing*, 2:1–4, 1981.

[22] G. Dhatt and G. Touzot. *Une Présentation de la Méthode des Éléments Finis.* Maloine S.A. Éditeur, 1984. 2^e édition.

[23] R.W. Freund and N.M. Nachtigal. An implementation of the QMR method based on coupled two-term recurrences. *SIAM Journal on Scientific Computing*, 15(2):313–337, 1994.

[24] H. Hakula, Y.Leino, and J. Pitkäranta. Scale resolution, locking and high-order finite element modelling of shells. *Computer Methods in Applied Mechanics and Engineering*, 133:157–182, 1996.

[25] T.J.R. Hughes and F. Brezzi. On drilling degrees of freedom. *Computer Methods in Applied Mechanics and Engineering*, 72:105–121, 1989.

[26] M. Lyly, R. Stenberg, and T. Vihinen. A stable bilinear element for the Reissner-Mindlin plate model. *Computer Methods in Applied Mechanics and Engineering*, 110:343–357, 1993.

[27] D.J. Allman. A compatible triangular element including vertex rotations for plane elasticity analysis. *Computers and Structures*, 19:1–8, 1984.

[28] A. Ibrahimbegovic, R.L. Taylor, and E.L. Wilson. A robust quadrilateral membrane finite element with drilling degrees of freedom. *International Journal for Numerical Methods in Engineering*, 30:445–457, 1990.

[29] T.A. Manteuffel. An incomple factorization technique for positive definite linear systems. *Mathematics of Computation*, 34:473–497, 1980.

[30] G. Horrigmoe and P.G. Bergan. Nonlinear analysis of free-form shells by flat finite elements. *Computer Methods in Applied Mechanics and Engineering*, 16:11–35, 1978.

SUURIKALIIPERISEN AMMUKSEN RASITUSTEN MITTAUS SISÄBALLISTISEN VAIHEEN AIKANA

SEPPO MOILANEN
Puolustusvoimien tutkimuskeskus
Fysiikan osasto
PL 5, 34 111 LAKIALA

TIIVISTELMÄ

Ammuskuoren putkiaikaisten kuormitusten mittaaminen edellyttää mittausantureiden sijoittamista ammukseen, mittausdatan keräämistä sisäballistisen vaiheen aikana ja tiedonsiirtovalmiutta lennon aikana ammuksesta vastaanottoasemalle. Mittaustiedon tallennukseen ja tiedonsiirtoon on kehitetty kotimainen ammuksen sisäinen telemetrialaitteisto, jolla voidaan mitata ammuksen perään vaikuttava ruutikaasun paine, ammuksen aksiaalinen kiihtyvyys ja kuoren venymiä putkiaikana. Tässä artikkelissa kuvataan telemetrialaitteiston keskeiset tekniset ominaisuudet sekä esitetään muutamia koeammuntatuloksia ammuksen sisäballistisen vaiheen mittauksista. Mittausjärjestelmä on osoittautunut käyttökelpoiseksi ja mittaustulokset hyödyllisiksi tykistön ja heittimistön ampumatarvikkeiden tutkimus- ja kehitystyössä.

JOHDANTO

Ammuskuoren ja aseen kuormituksien määrittäminen on perustunut klassiseen ruutikaasun paineen jakautumislakiin, jonka lähtökohtana on oletus lineaarisesta kaasun nopeusjakaumasta etäisyyden funktiona aseen lukkopinnasta mitattuna (Lagrangen oletus) sekä oletus ruutikaasun ja palavien ruutijyvästen tasaisesta jakautumasta palotilassa. Havaituista poikkeuksista huolimatta on lagrangelainen painejakauma katsottu riittävän tarkaksi menetelmäksi kuvaamaan aseiden ja ammusten rasituksia sisäballistisen vaiheen aikana. Uusissa

tykistön pitkänkantaman asejärjestelmissä ammuksen lähtönopeus v_0 on ~1,5-kertainen ja liike-energia E_0 (suuenergia) ~2,5-kertainen perinteisiin aseisiin verrattuna. Koetulosten nojalla on viime aikoina arvioitu, että Lagrangen yksinkertaistuksiin perustuvat sisäballistiset mallit ennustavat väärin sisäballistisen vaiheen tapahtumia suurilla suuenergioilla ja lähtönopeuksilla ammuttaessa. Korkeilla paineilla työskenteleville aseille on ehkä muodostettava uudet suunnitteluperusteet [1], joiden luotettavuus on varmistettava kokeellisin mittauksin.

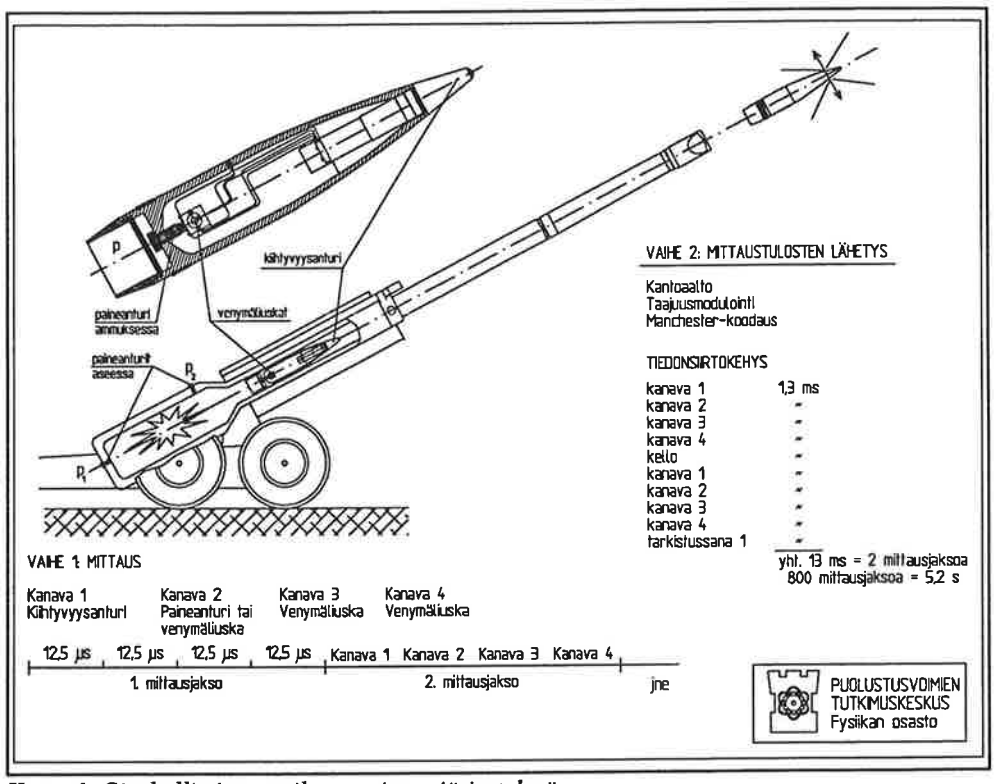
Perinteisesti ammuskuoret on mitoitettu käyttäen lineaarista kimmoteoriaa. Mitoitusperusteena on ollut sallitun jännityksen periaate ja myötölujuuden ylitystä ei ole sallittu tai se on sallittu vain rajoitetulla alueella kuoren geometrisissa epäjatkuvuuskohdissa. Koska ammuskuoren kuormitus putkiaikana on kertaluonteinen, on uusien ammuksien lujuusteknisessä mitoituksessa hyödynnetty myös kuorimetallin plastinen muodonmuutoskyky. Myötäminen sallitaan laajoilla alueilla kuoren seinämässä ja pohjassa mitoituskuormituksilla. Tällöin pienetkin kuormituksien lisäykset aiheuttavat suuria muodonmuutoksia ja kuoren putkikosketuksen vaara kasvaa. Plastisen muodonmuutoskyvyn hyödyntämisen edellytyksenä on, että ammuskuoren kuormitusolosuhteet ja kuorimateriaalin käyttäytyminen sisäballistisen vaiheen aikana tunnetaan entistä tarkemmin [2, 3, 4].

Tavanomaisissa koeammunnoissa aseesta mitataan ruutikaasun paine panoskammiosta $p_1(t)$ pietsokidepaineanturilla sekä ammuksen lähtönopeus v_0 putken suulla lähtönopeustutkalla. Lagrangen mallin mukainen ratkaisu 'sovitetaan' mittaustuloksiin. Laskennallisesti määritetään ammuksen perään kohdistuva paine p, ammuksen putken suuntainen kiihtyvyys a, putken rihlakäyrän avulla kulmanopeus ω ja -kiihtyvyys α [5].

SISÄBALLISTISEN VAIHEEN MITTAUS KOEAMMUKSELLA

Tarkka ammuksen liiketilan määrittäminen sekä sisäballististen mallien luotettavuuden arviointi edellyttävät ammukseen kohdistuvien kuormitusten mittausta putkiaikana [6, 7]. Puolustusvoimien tutkimuskeskuksen Fysiikan osasto, Puolustusvoimien materiaalilaitoksen esikunnan Ampumatarvikeosasto ja Noptel Oy ovat kehittäneet yhteistyöprojektina kuvan 1 mukaisen ammuksen sisäballistisen vaiheen mittausjärjestelmän, jolla voidaan mitata

ammuksen perään kohdistuva ruutikaasun paine p, ammuksen putken suuntainen kiihtyvyys a sekä ammuskuoren muodonmuutoksia ε_i putkiaikana.



Kuva 1. Sisäballistisen vaiheen mittausjärjestelmä.

Tavallinen ammus on instrumentoitu inertiksi sisäballistiseksi mittausammukseksi asentamalla paineanturi ammuksen pohjaan, venymäliuska-anturit (2 kpl) kuoren sisäseinämään ja kiihtyvyysanturi ammuksen kärkeen sytyttimen tilalle asennettuun telemetriaosaan. Ammuksen ulkopuolinen telemetriaosan geometrinen muoto täyttää tykistön sytyttimille asetetut standardivaatimukset ulkoballistiikan osalta, joten mittausammuksen ulkoballistiset ominaisuudet vastaavat normaalia ammusta. Mittausdata talletetaan putkiaikana telemetriaosan muistiin ja lähetetään lennon aikana radioteitse vastaanottoasemalle, jossa tulokset talletetaan mikrotietokoneella.

Keskeiset tiedonsiirtojärjestelmän tekniset ominaisuudet ovat:

a) näytteenottoväli 50 µs/kanava eli f_{mit}= 20 kHz,

b) 800 mittausjaksoa, kokonaismittausaika t=40 ms,

c) neljä mittauskanavaa,

d) radiolähettimen teho 1,5 W ja kantama yli 15 km,

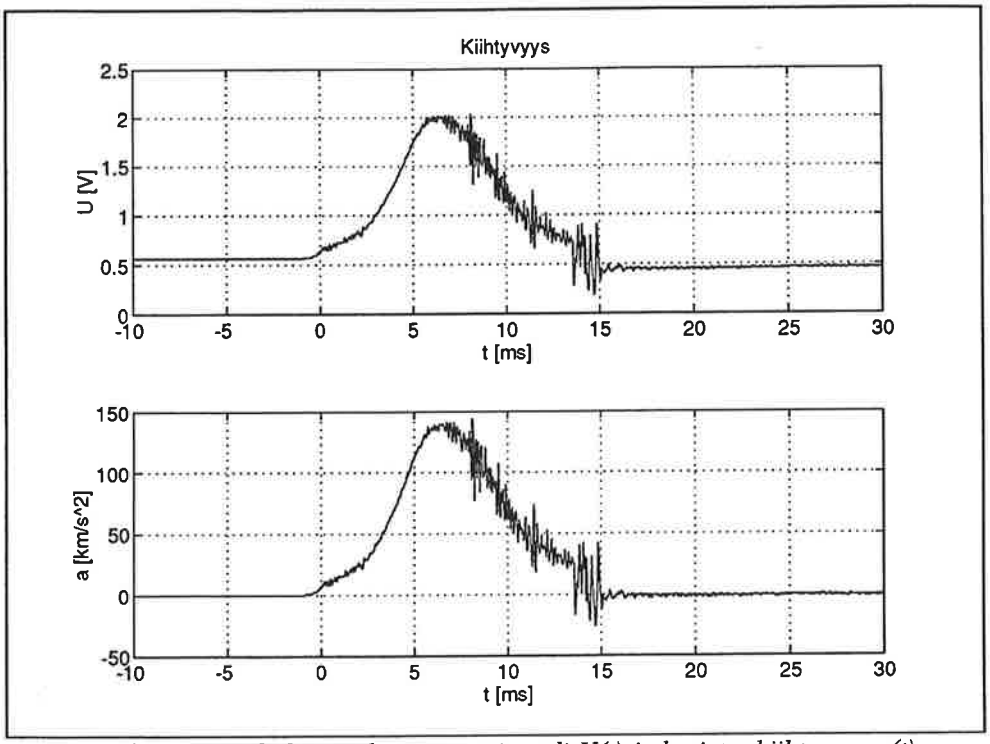
e) mittausjärjestelmä on toteutettu kaupallisilla komponenteilla.

MITTAUSTULOSTEN JÄLKIKÄSITTELY

Mittausammuksen antureilta tulevat signaalit vahvistetaan vahvistinpiireillä, jotka samalla toimivat ylipäästösuodattimina vääristäen mittaussignaaleja. Vääristymä korjataan laskennallisesti mittausdatan jälkikäsittelyllä. Kunkin vahvistimen mitattuun amplitudivasteeseen sovitetaan siirtofunktio, jonka avulla määritetään differenssiyhtälö korjauksen tekemiseksi. Vahvistimen suodattamasta signaalista rekonstruoidaan alkuperäinen mittaussignaali käänteissuodatuksella. Mittausdatan jälkikäsittely on toteutettu MATLAB-ohjelmistolla.

Esimerkki erään kiihtyvyysmittauksen signaaleista on kuvassa 2. Ammuksen korjaamattoman kiihtyvyyssignaalin $U(t)$ kuvasta on havaittavissa signaalin vääristyminen, sillä signaalin loppuosa on selvästi alkuperäistä lähtötasoa alempana. Todellisuudessa ammuksen hidastuvuus ilmalennon aikana on pieni verrattuna sisäballistisessa vaiheessa esiintyvään aksiaalikiihtyvyyteen ja hidastuvuus ei aiheuta mittaussignaalin tason muutosta. Kiihtyvyyssignaalin korjauksen jälkeen mittaussignaali on muunnettu kiihtyvyydeksi $a(t)$ mittayksikkönä km/s^2 kiihtyvyysanturin herkkyyden mukaisesti.

Putkivaiheen loppua kohti voimistuva kiihtyvyyssignaalin värähtely aiheutuu ammuksen kulmanopeuden ω kasvusta, jolloin ammuksen kärkiosa joutuu poikittaisliikkeisiin ammuksen ohjautuessa rihloissa ja kiihtyvyysanturiin kohdistuu voimakkaita iskumaisia rasituksia. Aivan putkiajan lopussa ammuksen kärjen poikittaisliike kasvaa, kun kuoren ohjauspaksunnokset ohittavat putken suun ja suujarrun. Tällöin poikittaisliikkeen amplitudi kasvaa ja kiihtyvyysanturi joutuu poikittaisliikkeeseen, josta aiheutuu voimakkaita häiriöpulsseja mittaussignaaliin.



Kuva 2. Alkuperäinen kiihtyvyyden mittaussignaali U(t) ja korjattu kiihtyvyys a(t).

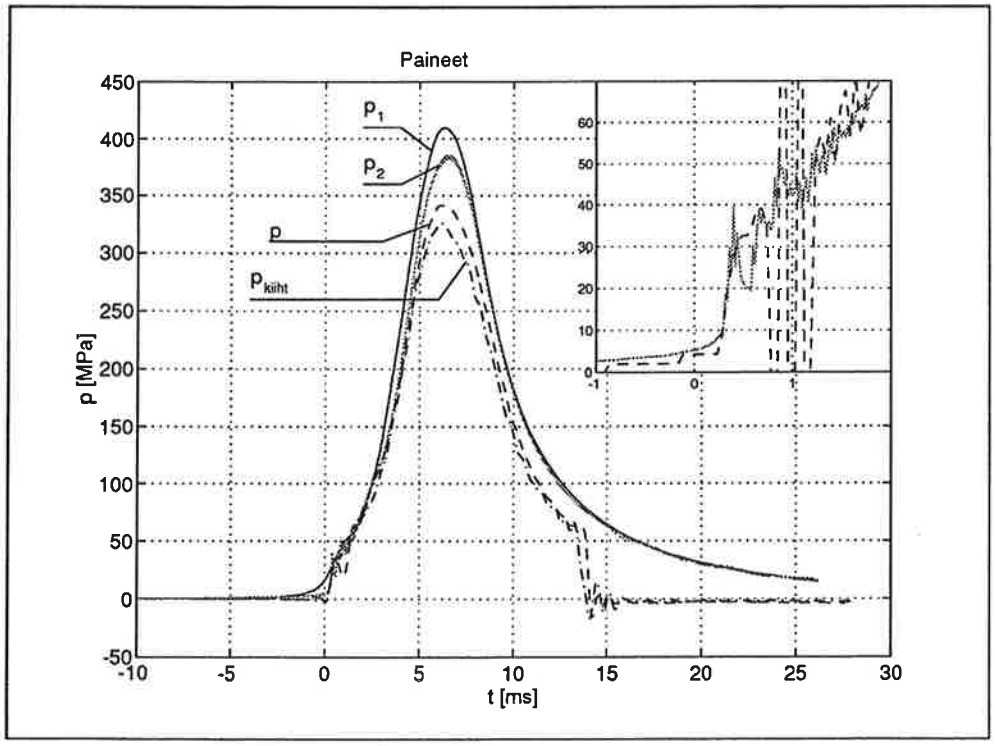
ESIMERKKI MITTAUSTULOKSISTA

Kuvassa 3 on esitetty aseesta mitatut lukkopaine p_1, panoskammion etuosassa vallitseva paine p_2, ammuksen perästä mitattu paine p sekä yhtälön 1 mukaan paineyksiköihin skaalattu kiihtyvyys (-paine) p_{kiiht} eräällä laukausyhdistelmällä ammuttaessa. Ammuksen massa on m ja putken poikkipinta-ala on A.

$$p_{kiiht} = \frac{ma}{A} \qquad (1)$$

Ison kuvan mittaustulokset on alipäästösuodatettu katkaisutaajuudella f_{cut}= 2 kHz. Pikkukuvan painearvot p ja p_2 ovat suodattamattomia. Ylänurkan pikkukuvassa on vertailtu aseesta mitatun panoskammiopaineen p_2 ja ammuksen perästä mitatun paineen p käyttäytymistä

putkiajan alussa, jolloin kyseiset mittausanturit ovat lähekkäin ja voidaan olettaa, että ruutikaasun paine on likimain sama molemmissa antureissa. Sekä panoskammion etuosan paineen p_2 että ammuspaineen p mittaustuloksessa esiintyy voimakasta ei-lagrangelaista painevärähtelyä putkivaiheen alussa, jota ei havaita lukkopaineen p_1 mittauksessa. Värähtelyt vaimenevat kuitenkin nopeasti ja painemittaukset käyttäytyvät "pehmeästi" putkivaiheen loppuajan. Värähtely on saattanut aiheutua panoksen epätavallisesta käyttäytymisestä syntyneestä, ammuksen peräosaan kohdistuneesta, mekaanisesta iskusta, jolloin mittauksessa havaittu värähtely ei välttämättä esitä ruudin palokaasujen paineaaltoja.



Kuva 3. Aseen panoskammiopaineet p_1 ja p_2, ammuksen perästä mitattu paine p sekä paineyksiköihin skaalattu ammuksen aksiaalinen kiihtyvyys p_{kiiht}.

Kunkin mittauksen maksimipaineita vertailtaessa havaitaan, että suurin painearvo saavutetaan lukossa. Lukkopinnasta etäännyttäessä paine pienenee eli $p_1 > p_2 > p > p_{kiiht}$ lukuun-

ottamatta em. mittauksen alkuosassa esiintyvää värähtelyä. Ammuksen perässä vaikuttavan paineen ja paineeksi skaalatun kiihtyvyyden erotus $p_{f\mu} = p - p_{kiiht}$ kuvaa ammuksen aksiaalista liikettä vastustavaa voimaa putkiaikana $F_r(t) = p_{f\mu} A$ tai putkimatkan x funktiona $F_r(x)$, kun ammuksen kulkema matka x on saatu integroimalla aksiaalinen kiihtyvyys a kaksi kertaa ajan suhteen.

YHTEENVETO

Ammuksen sisäinen telemetrialaite mahdollistaa ammuksen laukausrasitusten mittaamisen putkiaikana. Ruutikaasun paineen ja ammuksen kiihtyvyyden mittaustuloksia voidaan käyttää teoreettisten sisäballististen mallien vertailuun, ammuksen ja aseen kuormitusten kokeelliseen määrittämiseen sekä laskennallisten lujuustarkastelujen tulosten luotettavuuden arviointiin. Aineenkoetuskokeiden nopeusalueen valitsemiseksi kuoren venymämittaustuloksista voidaan määrittää laukaustapahtumassa esiintyvät muodonmuutosnopeudet $d\varepsilon_i/dt$ sekä venymähistoriat $\varepsilon_i(t)$ kuoren kriittisiltä alueilta. Venymätuloksia voidaan vertailla kuoren laskennallisten lujuustarkastelujen tuloksiin.

Paitsi sisäballistiikan mittaukseen mittausjärjestelmä soveltuu myös ulkoballistisiin mittauksiin. Mittausjärjestelmää on käytetty ammuksen lennonaikaisten tapahtumien mittauksiin kuten kuoren pintalämpötilan mittaukseen sekä ammusten sähköisten komponenttien tai virtalähteiden toiminnan mittauksiin aina sisäballistisesta vaiheesta iskemään saakka. Sekä sisä- että ulkoballistiset mittaustulokset ovat osoittautuneet käyttökelpoisiksi ja hyödyllisiksi tykistön ja kranaatinheittimistön ase- ja ampumatarvikejärjestelmien tutkimus- ja kehitystyössä.

LÄHTEET

1. Talja, J. Asepaineet ja painejakauman ongelma. Julkaisussa: Kantolahti, E. & Pääkkönen, E. (Toim.). *Eräiden uusien räjähdysaineiden ja ruutien teknologiasta.* Seminaarijulkaisu A/2, Puolustusvoimien tutkimuskeskus, Lakiala, 1993. ss. 65,...,73.

2. Erkkilä, T. Kranaatin metallurgiset ongelmat. Julkaisussa: Salonen, L. & Pääkkönen, E. (Toim.). *Vuosipäiväesitelmät 1992, Materiaalitekniikka.* Julkaisuja B/1. Puolustusvoimien tutkimuskeskus. Lakiala, 1993. ss. 3,...,8.

3. Moilanen, S. Kranaatin kuoren lujuuslaskut elementtimenetelmällä. Julkaisussa: Salonen, L. & Pääkkönen, E. (Toim.). *Vuosipäiväesitelmät 1992, Materiaalitekniikka.* Julkaisuja B/1. Puolustusvoimien tutkimuskeskus. Lakiala, 1993. ss. 9,...,18.

4. Moilanen, S. *Ammuskuoren rasitukset putkiaikana.* Diplomityö. TTKK/TME. Tampere, 1992. 131 s.

5. Tuomainen, A. *The Thermodynamic Model of Interior Ballistics.* Acta Polytechnica Scandinavica, Applied Physics Series No. 205. University of Helsinki. Helsinki, 1996. 87 p.

6. Ewans, J. W. *Measurement of interior ballistic performance using FM/FM radio telemetry techniques.* BRL-TR-2699. U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground. Maryland, 1985. 93 p.

7. Ruth, C. & all. *System Checkout of the 155-mm Short-Barreled Howitzer Using Telemetry Projectiles.* ARL-MR-8. U.S. Army Research Laboratory, Aberdeen Proving Ground. Maryland, 1992. 32 p.

PUUVÄLIPOHJIEN OMINAISVÄRÄHTELYT

M. KILPELÄINEN ja S. PALOLA
Rakennetekniikan laboratorio
Rakentamistekniikan osasto
Oulun yliopisto
PL 191, 90101 Oulu

TIIVISTELMÄ

Liikuttaessa puuvälipohjan päällä voidaan sen värähtely kokea epämiellyttäväksi. Haitallinen värähtely pyritään välttämään laskennallisella värähtelymitoituksella. Mitoituksessa keskeisenä tehtävänä on välipohjarakenteen ominaisjaksolukujen ja niistä erityisesti alhaisimman ominaisjaksoluvun määrittäminen.

Ouluun v. 1996-1997 rakennetussa puukerrostalossa on käytetty kahta erilaista välipohjatyyppiä. Kummallekin rakennetyypille on määritetty laskennallisesti ja rakennuspaikalla mittaamalla ominaisvärähdysluvut ensin ns. raakavälipohjan (pelkän kantavan rakenteen) päältä ja myöhemmin valmiin rakenteen (kun ylä- ja alapuoliset levykerrokset ovat paikallaan) päältä.

Artikkelissa esitetään em. välipohjarakenteiden laskennalliset ominaisjaksoluvut, joita sitten verrataan mitattuihin arvoihin. Tulosten perusteella arvoidaan käytettyjen laskentamenetelmien soveltuvuutta värähtelymitoitukseen ja esitetään arvio ominaisjaksolukuihin vaikuttavista tekijöistä.

1. JOHDANTO

Viime vuosina on Suomeen rakennettu useita puukerrostaloja. Niiden suunnittelussa eräänä keskeisenä haasteena on ollut välipohjien ääni- ja värähtelytekninen suunnittelu. Puuvälipohjien ominaisjaksoluvut ovat verraten matalia ja samaa suuruusluokkaa (n. 2-20 Hz)

kuin välipohjien dynaamisten kuormien värähdysluku (kävely ja muu liikkuminen, pesukoneet jne.). Koska puuvälipohjan jäykkyys on hyvin pieni, voi rakenteen värähtely muodostua häiritsevän voimakkaaksi, mitä vielä mahdollisesti syntyvä resonanssi- ilmiö voi vahvistaa. Värähtelyn voivat kokea epämiellyttäväksi paitsi välipohjan päällä liikkujat myös sen alapuolella olevat henkilöt.

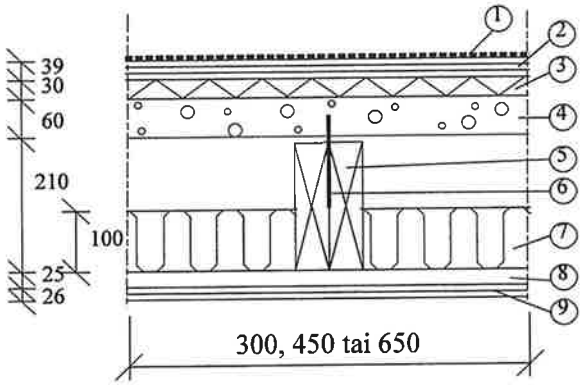
Haitallinen värähtely pyritään välttämään laskennallisella värähtelymitoituksella. Suomen puunormeissa [1], [2] ei ole ohjeita värähtelymitoituksesta. Puurakenteiden eurooppalaisessa esistandardissa EC 5 [3] on ohjeet reunoiltaan vapaasti tuetun suorakaiteenmuotoisen välipohjan värähtelymitoituksesta. Ohjeet soveltuvat parhaiten ns. raakavälipohjan laskennalliseen tarkasteluun, jossa ei kantavan rakenteen yläpuolisen uivan lattian ja alapuolisen jousirangan ja kipsilevyjen muodostamien jousi-massasysteemien vaikutusta oteta huomioon muuten kuin lisämassana. Ohjeet perustuvat Ohlssonin tutkimuksiin [4], joita myös Suomessa on sovellettu kevyiden välipohjien värähtelyanalyysiin.

Mitoituksessa keskeisenä tehtävänä on välipohjarakenteen ominaisjaksolukujen ja niistä erityisesti alhaisimman ominaisjaksoluvun määrittäminen. Laskentatulosten luotettavuuden kannalta on tällöin tärkeää oikean rakenne- ja värähtelymallin valitseminen sekä oikeiden laskentaparametrien (kimmomodulit, jäykkyydet) käyttäminen. Käytännön suunnittelutyön kannalta katsottuna laskelmat eivät saa muodostua liian työläiksi.

Tämän tutkimuksen tavoitteena on ollut arvioida eräiden laskentamallien käyttökelpoisuutta puuvälipohjien ominaisjaksolukujen määrittämiseen. Arviointi tapahtuu vertaamalla laskennallisesti saatuja ominaisjaksolukuja rakenteista mittaamalla saatuihin arvoihin. Ouluun v. 1996-1997 rakennetussa puukerrostalossa on käytetty kahta erilaista välipohjatyyppiä. Kummallekin rakennetyypille on määritetty rakennuspaikalla useissa pisteissä mittaamalla ominaisjaksoluvut ensin ns. raakavälipohjan (pelkän kantavan rakenteen) päältä ja myöhemmin valmiin rakenteen (kun ylä- ja alapuoliset levykerrokset ovat paikallaan) päältä. Tutkimuksen yksityiskohtaiset laskelmat ja tulokset on esitetty lähteessä [5]. Tietoja vastaavista aikaisemmista tutkimuksista ei ole käytettävissä.

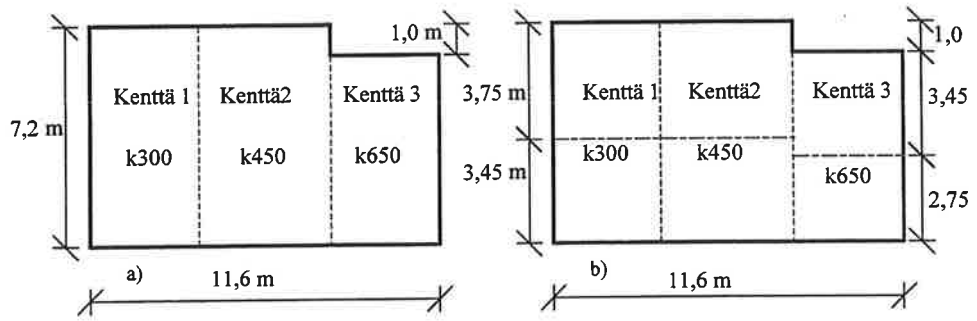
2. TUTKITTAVAT RAKENTEET

Rakennuksen A välipohjissa on käytetty puu-betoniliittorakennetta (RL- laattavälipohja). Sen rakenne käy ilmi kuvasta 1.



KUVA 1: Valmiin RL-laattavälipohjan poikkileikkaus.

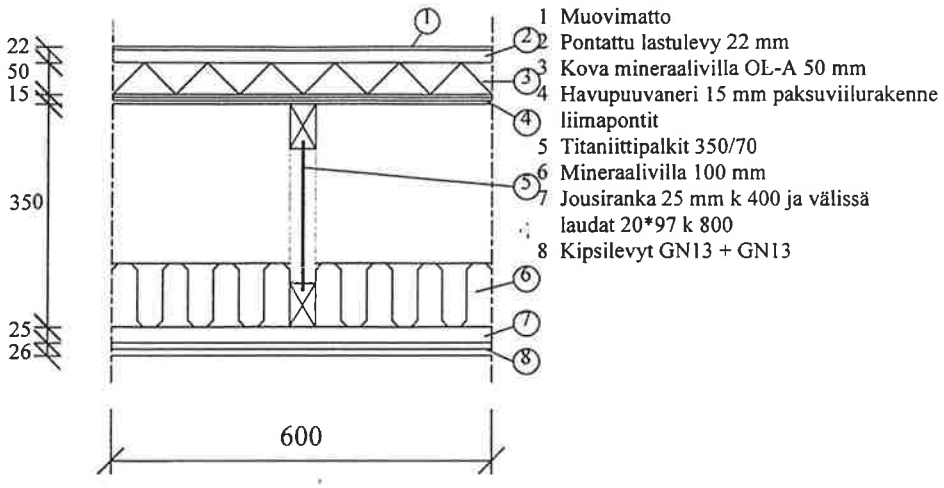
Raakavälipohja toimi mittaushetkellä yksiaukkoisena. Liiallisten taipumien estämiseksi se jouduttiin rakennustyön aikana tukemaan väliseinillä, jonka vuoksi valmis välipohja toimii kaksiaukkoisena (kuva 2).



KUVA 2: RL-laattavälipohjan kentät jännevälin ja palkkijaon perusteella tutkitussa huoneistossa a) raakavälipohja, b) valmis välipohja.

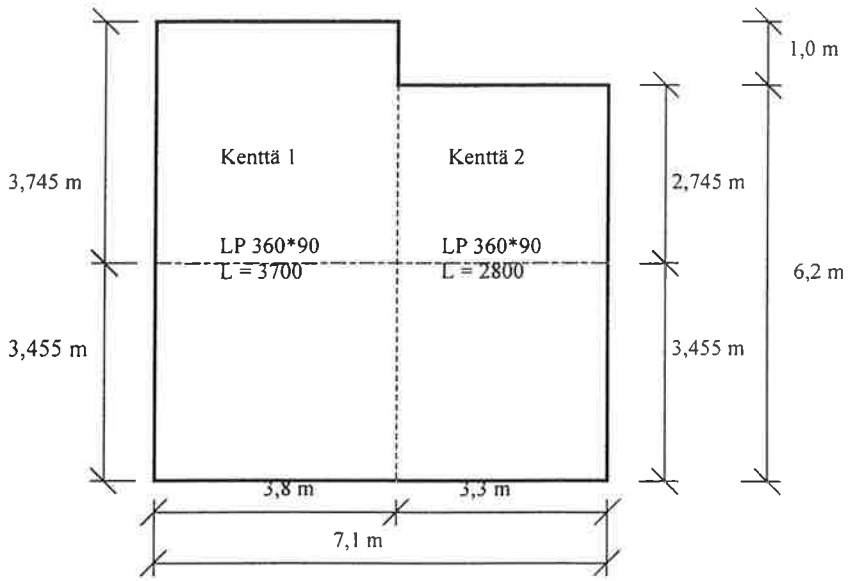
Raakavälipohja on tehty n. 2,5 m leveistä liittorakenne-elementeistä. Elementit on kiinnitetty pitemmiltä sivuiltaan hitsatuilla lattateräskappaleilla kahdesta kohdasta toisiinsa. Lisäksi elementtien välinen sauma on vahvistettu tartuntateräksillä ja valettu betonivalulla umpeen. Poikittaispalkkeja on kussakin palkkivälissä tukipisteiden lisäksi kolme kappaletta.

Rakennuksen C välipohjissa on käytetty uumalevypalkki-vaneriliittorakennetta (Titaniitti-välipohja). Sen rakenne käy ilmi kuvasta 3.



KUVA 3: Valmiin Titaniittivälipohjan poikkileikkaus.

Sekä raakavälipohja että valmis rakenne toimivat kaksiaukkoisina. Välitukena on liimapuusta tehty pilari-palkkilinja (kuva 4).



KUVA 4: Titaniittivälipohjan kentät jännevälin perusteella tutkitussa huoneistossa.

Raakavälipohja on tehty n. 2,5 m leveistä liittorakenne-elementeistä. Elementit on kiinnitetty pitemmältä sivulta naulaamalla toisiinsa. Poikittaispalkkeja on kussakin palkkivälissä tukipisteiden lisäksi kaksi kappaletta.

Kummankin välipohjarakenteen materiaalitiedot ja laskenta-arvot on esitetty lähteessä [5].

3. LASKENTAMALLIT

3.1 Yksiulotteiset eli palkkimallit

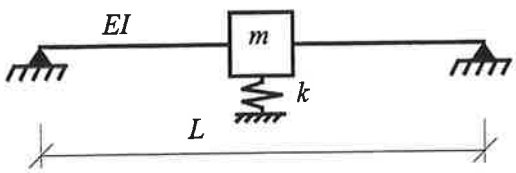
Yksiulotteisessa mallissa oletetaan välipohjapalkkien toimivan toisistaan riippumattomina eli välipohjan jousivakio palkkeja vastaan kohtisuorassa suunnassa jätetään huomioonottamatta. Malli on yksinkertainen ja soveltuu epäsäännöllisille rakennusten pohjaratkaisuille.

Laskentamalli 1

Puolet palkkikaistan massasta keskitetään jännevälin keskelle massaksi m (kuva 5). Palkki on yksiaukkoinen. "Jousen" jousivakio k on sama kuin palkin taipumajäykkyys ja saadaan vapaasti tuetulle palkille kaavasta

$$k = \frac{48EI}{L^3} \tag{1}$$

L on palkin jänneväli [m]
EI on palkin taivutusjäykkyys [N·m^2]



KUVA 5: Välipohjan värähtelymalli laskentamallissa 1.

Tällöin ominaistaajuus saadaan kaavasta

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{2}$$

Laskentamalli 2

Keskitetään edelleen puolet palkkikaistan massasta jännevälin keskelle. Raakavälipohjan osuutta merkitään m_2:lla, uivan lattian massaa m_1:llä ja alapuolisten kipsilevyjen massaa m_3:lla (kuva 6). Uivan lattian ja raakavälipohjan välisen ilmavälin jousivakio on k_1 ja se saadaan kaavasta

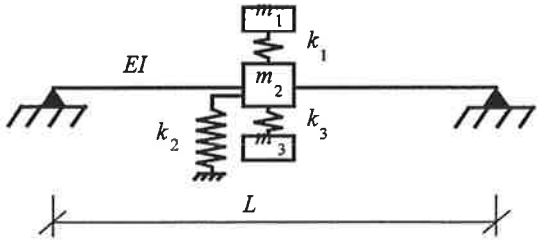
$$k_1 = \frac{c^2 \cdot \rho}{d} \tag{3}$$

c on äänen nopeus ilmassa [m/s]

ρ on ilman tiheys [kg/m^3]

d on ilmavälin paksuus [m]

k_2 on palkin taipumajäykkyys ja k_3 jousirangan ja ilmavälin yhteenlaskettu jousivakio, joka saadaan kokeellisesti.

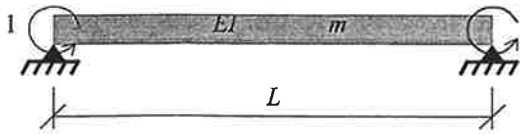


KUVA 6: Välipohjan värähtelymalli laskentamallissa 2.

Kysymyksessä on kolmen vapausasteen värähtelysysteemi. Ominaisjaksoluvut (3 kpl) saadaan ratkaistuiksi matriisimuotoisesta ominaisarvotehtävästä $[K]\cdot\{u\}=\omega^2\cdot[M]\cdot\{u\}$, josta saadaan taajuudet $f = \omega/2\pi$.

Laskentamalli 3

Välipohjalle käytetään kuvan 7 mukaista rakennemallia, joka koostuu yhdestä palkkielementistä.



KUVA 7: Välipohjan värähtelymalli laskentamallissa 3.

Massan m jakautumista kuvataan staattiseen siirtymämalliin perustuvan konsistentin massamatriisin avulla. Rakenteen globaali konsistentti massamatriisi on muotoa

$$[M_c]=\frac{m}{420}\begin{bmatrix}4L^2 & -3L^2\\ -3L^2 & 4L^2\end{bmatrix} \tag{4}$$

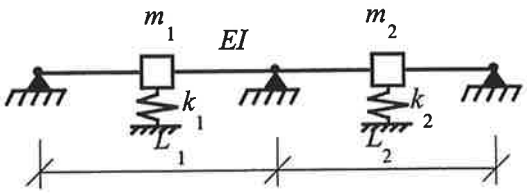
Rakenteen globaali jäykkyysmatriisi on muotoa

$$[K]=\frac{EI}{L^3}\begin{bmatrix}4L^2 & 2L^2\\ 2L^2 & 4L^2\end{bmatrix} \tag{5}$$

Ominaistaajuudet (2 kpl) saadaan ratkaistuiksi matriisimuotoisesta ominaisarvotehtävästä.

Laskentamalli 4

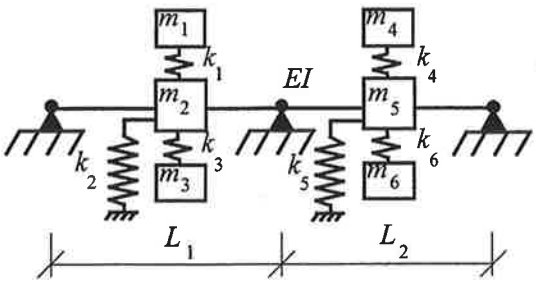
Mallilla kuvataan kaksiaukkoisen jatkuvan palkin värähtelyä. Laskennan yksinkertaistamiseksi välituelle otaksutaan nivel, jolloin palkki jakautuu kahdeksi yksiaukkoiseksi palkiksi, joiden ominaisjaksoluvut voidaan laskea laskentamallia 1 käyttäen. Malli on esitetty kuvassa 8.



KUVA 8: Välipohjan värähtelymalli laskentamallissa 4.

Laskentamalli 5

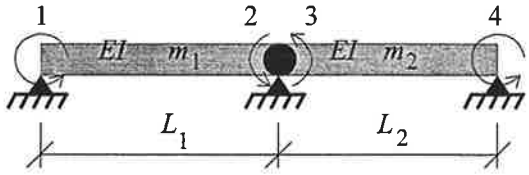
Mallilla kuvataan kaksiaukkoisen jatkuvan palkin värähtelyä. Välituelle otaksutaan jälleen nivel, jolloin palkki jakautuu kahdeksi yksiaukkoiseksi palkiksi. Niiden ominaisjaksoluvut voidaan laskea siten laskentamallia 2 käyttäen. Malli on esitetty kuvassa 9.



Kuva 9: Välipohjan värähtelymalli laskentamallissa 5.

Laskentamalli 6

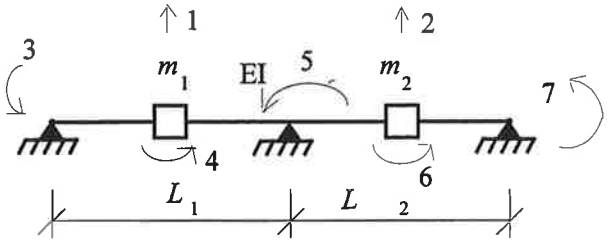
Mallilla kuvataan edelleen kaksiaukkoisen jatkuvan palkin värähtelyä kahden yksiaukkoisen palkin avulla. Niiden ominaisjaksoluvut saadaan laskentamallia 3 käyttäen. Malli on kuvattu kuvassa 10.



KUVA 10: Välipohjan värähtelymalli laskentamallissa 6.

Laskentamalli 7

Tarkastellaan kaksiaukkoista palkkia, jonka massasta keskitetään puolet kummankin jännevälin keskelle kuvan 11 mukaisesti.

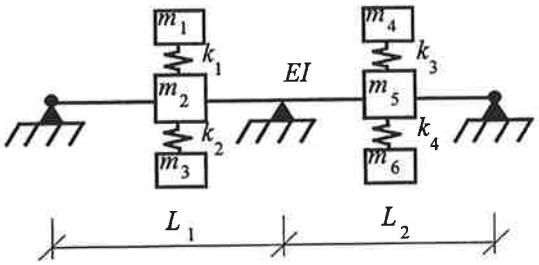


KUVA 11: Välipohjan värähtelymalli laskentamallissa 7.

Palkista muodostetaan neljä palkkielementtiä käsittävä laskentamalli, jossa vapausasteita on 7. Ominaisjaksoluvut (2 kpl) saadaan matriisimuotoisen ominaisarvotehtävän ratkaisuna.

Laskentamalli 8

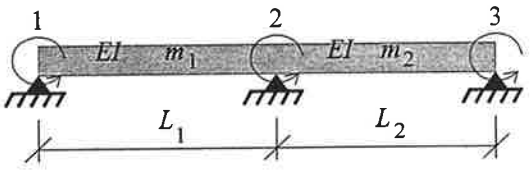
Tarkastellaan edelleen kaksiaukkoista palkkia, jossa raakavälipohjan massasta puolet keskitetään kummankin jännevälin keskelle. Lisäksi raakavälipohjan yläpuolinen uiva lattia ja alapuoliset levykerrokset erotetaan erillisiksi massoiksi samoin jännevälin keskipisteisiin. Malli on esitetty kuvassa 12. Tätä laskentamallia ei käytetä tässä tutkimuksessa.



KUVA 12: Välipohjan värähtelymalli laskentamallissa 8.

Laskentamalli 9

Tarkastellaan vielä kaksiaukkoista jatkuvaa palkkia, jonka massa on tasanjakautunut (kuva 13).



KUVA 13: Välipohjan värähtelymalli laskentamallissa 9.

Rakennemalli koostuu kahdesta palkkielementistä. Niiden muodostama konsistentti massamatriisi on

$$[M_c] = \frac{1}{420}\begin{bmatrix} 4L_1^2 m_1 & -3L_1^2 m_1 & 0 \\ -3L_1^2 m_1 & 4L_1^2 m_1 + 4L_2^2 m_2 & -3L_2^2 m_2 \\ 0 & -3L_2^2 m_2 & 4L_2^2 m_2 \end{bmatrix} \qquad (6)$$

Laskentamallin jäykkyysmatriiksi saadaan

$$[K]=\begin{bmatrix} \frac{4(EI)_1}{L_1} & \frac{2(EI)_1}{L_1} & 0 \\ \frac{2(EI)_1}{L_1} & \frac{4(EI)_1}{L_1}+\frac{4(EI)_2}{L_2} & \frac{2(EI)_2}{L_2} \\ 0 & \frac{2(EI)_2}{L_2} & \frac{4(EI)_2}{L_2} \end{bmatrix} \tag{7}$$

Ominaistaajuudet saadaan ominaisarvoyhtälön ratkaisuna.

3.2 Kaksiulotteiset eli laattamallit

Kaksiulotteisissa malleissa otetaan huomioon välipohjan jäykkyys palkkeja vastaan kohtisuorassa suunnassa. Jäykkyys tässä suunnassa koostuu palkkien varassa olevan laatan (betoni, vaneri tms.) ja poikkipalkkien jäykkyydestä. Erityisesti poikkipalkkien jäykkyyttä on vaikea arvioida.

Mallit soveltuvat säännöllisten, suorakulmaisten välipohjatasojen värähtelyanalyysiin. Ominaisjaksoluvut lasketaan pelkästään raakavälipohjarakenteelle, johon liittyvät ylä- ja alapuoliset levy- ym. kerrokset ovat pelkästään raakavälipohjan lisämassana.

Ohlssonin malli

Sven Ohlssonin esittämä malli on esitetty lähteessä [4]. Sen mukaan alin ominaisjaksoluku määritetään kaavasta

$$f=\frac{\pi}{2\cdot L^2}\sqrt{\frac{EI_x}{m}}\cdot\sqrt{1+\left[2\left(\frac{L}{B}\right)^2+\left(\frac{L}{B}\right)^4\right]\cdot\frac{EI_y}{EI_x}} \tag{8}$$

Tällöin

EI_x on välipohjan taivutusjäykkyys palkkien suunnassa yhden pituusyksikön levyistä kaistaa kohden ($N\cdot m^2/m$]

EI_y on välipohjan taivutusjäykkyys palkkeihin nähden kohtisuorassa suunnassa yhden pituusyksikön levyistä kaistaa kohden [$N\cdot m^2/m$]

m on välipohjan massa pinta-alayksikköä kohden [kg/m^2]

L on palkkien jänneväli eli välipohjan toinen sivumitta [m]
B on välipohjan toinen sivumitta [m]

Palkit ovat yksiaukkoisia ja vapaasti tuettuja.

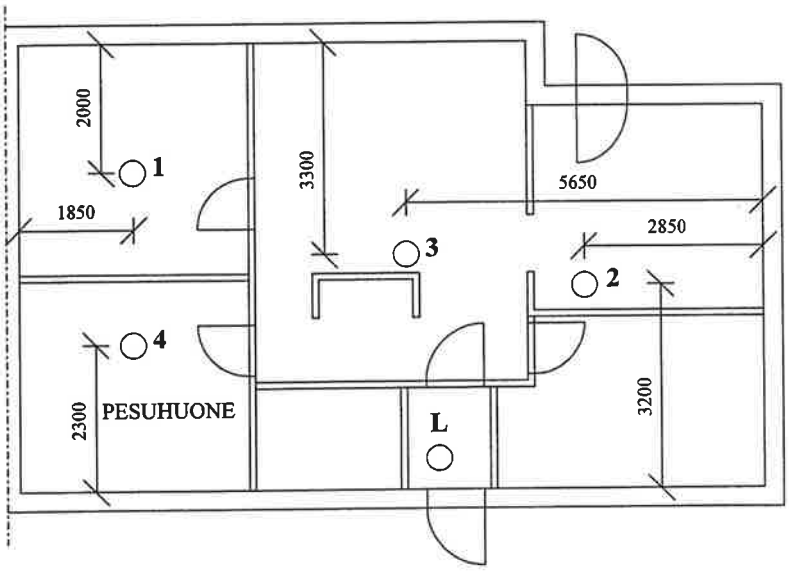
Eurocode 5:n malli
Malli on esitetty lähteessä [3]. Sen mukaan alin ominaisjaksoluku lasketaan kaavasta

$$f = \frac{\pi}{2L^2} \cdot \sqrt{\frac{EI_x}{m}} \tag{9}$$

Myös tässä mallissa oletetaan palkit yksiaukkoisiksi ja vapaasti tuetuiksi.

4. VÄRÄHTELYMITTAUKSET

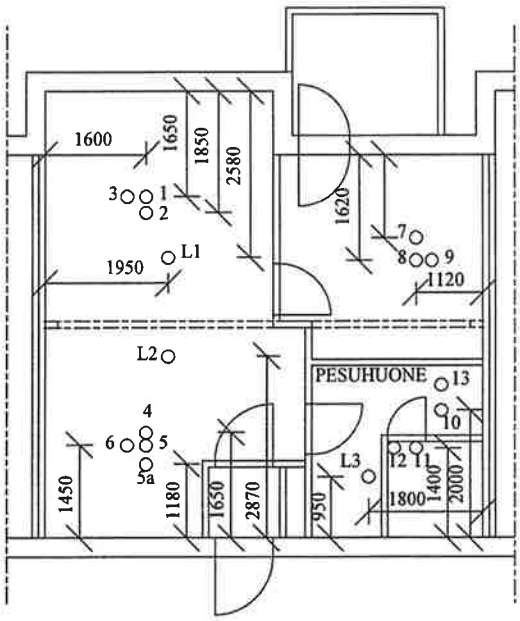
Ominaisjaksoluvut määritettiin mittaamalla rakennusten A ja C 2. kerroksen lattiasta yhdessä huoneistossa kummassakin rakennuksessa. Mittauspisteet 1...4 rakennuksessa A on esitetty kuvassa 14.



KUVA 14: Mittauspisteiden 1...4 sijainti rakennuksessa A. L on lyöntipiste.

Mittauspisteet 1...13 rakennuksessa C on esitetty kuvassa 15. Valmiista välipohjasta ei mittauksia tehty välipohjapalkkien välien kohdilta.

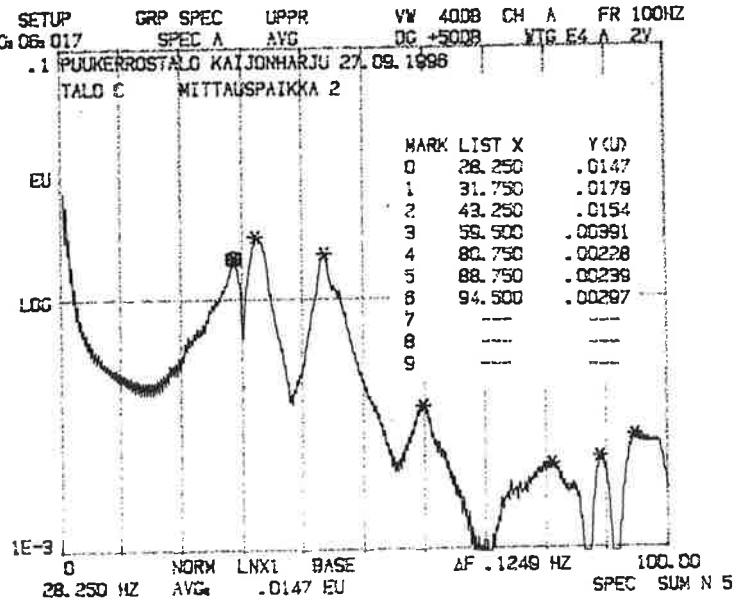
Kaikki mittaukset tehtiin Oulun yliopiston konetekniikan osaston laitteilla laboratorioinsinööri Osmo Väliheikin johdolla,



KUVA 15: Mittauspisteiden 1...13 sijainti rakennuksessa C. L1...L3 ovat lyöntipisteitä.

Välipohjan mittauspisteen kiihtyvyys mitattiin välipohjaan teipillä kiinnitetyn kiihtyvyysanturin avulla. Heräte saatiin aikaan lyömällä vasaralla lyöntipisteeseen. Koska mittauksen avulla haluttiin mitata vain ominaistaajuudet, ei herätteen voimakkuudella ollut merkitystä.

Tuloksina saatiin mittauspisteiden kiihtyvyys taajuuden funktiona. Jokaisesta mittauspisteestä tehtiin 5 mittausta, joiden keskiarvona mittauspisteen lopullinen kiihtyvyyskäyrä tulostettiin. Ominaistaajuudet löytyvät kiihtyvyyskäyrän maksimiarvojen kohdilta. Tyypillinen kiihtyvyyskäyrä jaksoluvun funktiona on esitetty kuvassa 16.



KUVA 16: Rakennuksen C raakavälipohjan kiihtyvyyskäyrä ja ominaisjaksoluvut pisteessä 2.

Mittaamalla saadut ominaisjaksoluvut rakennuksissa A ja C eri mittauspisteissä on esitetty taulukoissa 1 ja 2.

TAULUKKO 1: RL-laattavälipohjan mitatut ominaistaajuudet rakennuksessa A.

Ominais-muoto	OMINAISTAAJUUDET MITTAUSPISTEITTÄIN [Hz]							
	Raaka välipohja				Valmis välipohja			
	1	2	3	4	1	2	3	4
1	8,12	8,00	8,00	8,12	25,00	15,50	9,25	13,75
2	9,75	9,87	9,00	9,75	27,50	24,25	15,00	25,00
3	11,62	13,81	9,81	11,62	40,00	44,75	18,50	31,50
4	13,62	16,37	13,56	13,62	63,50	67,50	38,50	68,50
5	18,00	23,38	14,37	18,00	###	###	48,25	###
6	20,81	28,38	23,50	20,81	###	###	68,00	###
7	24,69	29,94	26,94	24,69	###	###	###	###
8	32,87	###	###	32,87	###	###	###	###
9	45,62	###	###	45,62	###	###	###	###

TAULUKKO 2: Titaniittivälipohjan mitatut ominaistaajuudet rakennuksessa C.

Ominais-muoto	OMINAISTAAJUUDET MITTAUSPISTEITTÄIN [Hz]							
	Raaka välipohja							
	1	2	3	4	5	5A	6	7
1	28,25	28,25	28,00	25,00	25,00	25,00	25,00	22,00
2	31,75	31,75	32,50	31,75	31,75	31,75	31,75	24,50
3	43,00	43,25	43,00	43,00	43,50	43,50	43,50	29,50
4	58,75	59,50	59,75	48,00	48,25	48,25	48,25	35,50
5	73,25	80,75	73,75	80,75	59,00	69,00	69,00	49,75
6	79,75	88,75	88,25	91,00	81,50	83,75	90,75	54,75
7	88,75	94,50	97,50	106,0	92,75	93,50	93,75	69,25
8	94,25	###	106,5	117,0	95,25	95,25	105,5	94,25
9	109,0	###	109,3	128,8	106,5	108,5	130,8	104,5
	7*	8	9	9*	10	11	12	
1	20,75	22,00	22,00	21,00	21,25	21,25	21,25	
2	25,00	24,50	28,00	25,50	30,00	30,00	30,00	
3	28,50	29,50	34,75	32,25	32,75	32,75	32,75	
4	32,50	35,50	49,50	35,00	45,25	45,25	45,25	
5	45,75	49,75	55,50	45,75	50,25	50,25	50,25	
6	51,50	69,00	69,75	53,25	60,00	60,00	60,00	
7	60,50	77,25	73,00	61,25	65,00	65,00	69,50	
8	78,25	97,25	79,00	68,75	79,75	79,75	79,75	
9	87,00	107,5	82,75	79,50	97,25	90,75	94,00	
	Valmis välipohja							
	1	3	5	6	8	9	11	13
1	6,75	13,75	21,25	20,75	24,50	30,00	27,00	27,00
2	21,25	21,75	28,5	28,50	29,75	44,50	61,75	59,25
3	45,12	45,25	51,00	46,25	46,25	55,00	79,50	79,75
4	###	85,00	66,25	66,25	###	77,75	###	121,75
5	###	139,5	###	###	###	150,0	###	###

*) Vertailumittaus herätteen lyöntipiste L1

Voidaan todeta taulukoista 1 ja 2, että raakavälipohjalta ja valmiilta välipohjalta mitatut ominaisjaksoluvut poikkeavat toisistaan huomattavasti.

5. LASKETTUJEN JA MITATTUJEN OMINAISJAKSOLUKUJEN VERTAILU

Eri laskentamalleilla lasketut alimmat ominaisjaksoluvut ja vastaavat mittaamalla saadut alimmat ominaisjaksoluvut eri pisteissä on esitetty taulukoissa 3...6.

TAULUKKO 3: Lasketut ja mitatut alimmat ominaisjaksoluvut RL-laattavälipohjalla eri mittauspisteissä. Raakavälipohja.

Mittauspiste	ALIMMAT OMINAISTAAJUUDET [Hz]				
	Laskentatulokset				Mittaustulos
	LM1	LM3	Ohlsson	EC 5	
1	7,2	8,1	7,5	7,3	8,12
2	7,6	8,5	7,9	7,7	8,00
3	6	7,2	6,6	6,5	8,00
4	7,2	8,1	7,5	7,3	8,12

TAULUKKO 4: Lasketut ja mitatut alimmat ominaisjaksoluvut RL-laattavälipohjalla eri mittauspisteissä. Valmis välipohja.

Mittauspiste	ALIMMAT OMINAISTAAJUUDET [Hz]							
	Laskentatulokset							Mittaustulos
	LM4	LM5	LM6	LM7	LM9	Ohlsson	EC 5	
1	21,5	20,2	24	22,1	24,7	21,7	12,6	25,00
2	19,4	18,4	21,7	22,3	25,7	19,7 (6,2)*	19,6	15,50
3	18,8	17,9	21	20,1	22,6	19,0 (5,3)**	18,9	9,25
4	22,8	20,5	25,4	22,1	24,7	23 (9,7)***	22,9	13,75

*) jänneväli 6,2 m

**) jänneväli 7,2 m

***) 3,45*3,6 m^2 tasausbetonilaatan ominaistaajuus

TAULUKKO 5: Lasketut ja mitatut alimmat ominaisjaksoluvut Titaniitti-välipohjalla eri mittauspisteissä. Raakavälipohja. LP-palkki tarkoittaa Titaniittipalkkien keskitukena olevien liimapuupalkkien alinta ominaisjaksolukua laskettuna laskentamallilla 1.

	ALIMMAT OMINAISTAAJUUDET [Hz]							
Mittauspiste	**Mittaustulokset**							**Mittaustulosten**
	LM4	**LM6**	**LM7**	**LM9**	**LP-palkki**	**Ohlsson**	**EC 5**	**keskiarvo**
1-3	28,9	32,2	21,2	33,8	20,7	29,1	29,1	28,20
4-6	32,6	36,1	21,2	33,8	20,7	32,7	32,7	25,00
7-9	44,4	49,7	21,4	40,5	36,6	44,7	44,7	22,00
10-12	32,6	36,1	21,4	40,5	36,6	32,7	32,7	21,25

TAULUKKO 6: Lasketut ja mitatut alimmat ominaisjaksoluvut Titaniittivälipohjalla eri mittauspisteissä. Valmis välipohja.

	ALIMMAT OMINAISTAAJUUDET [Hz]								
Mittauspiste	**Laskentatulokset**								**Mittaustulosten**
	LM4	**LM5**	**LM6**	**LM7**	**LM9**	**LP-palkki**	**Ohlsson**	**EC 5**	**keskiarvo**
1	13,9	12,9	15,5	9,3	16,3	10,6	14	14	6,75
3	13,9	12,9	15,5	9,3	16,3	10,6	14	14	13,75
5, 6	15,6	14,2	17,4	9,3	16,3	10,6	15,8	15,7	21,00
8, 9	21,3	17,7	23,9	7,6	15,3	15,6	21,5	21,5	27,30
11, 13	11,5	11	12,9	7,6	15,3	15,6	14,3	11,6	27,00

Taulukoiden 3...6 perusteella voidaan todeta seuraavaa:

1. Sekä palkkimallit (laskentamallit 1 ja 3) että laattamallit (Ohlsson ja EC 5) antavat yksiaukkoisen RL-laattavälipohjan alimmalle ominaisjaksoluvulle luotettavan arvion (taulukko 3). Paras vastaavuus laskentatulosten ja mitattujen tulosten välillä saadaan laskentamallilla 3.
2. Valmiin RL-laattavälipohjan ominaistaajuudelle ei mikään käytetyistä laskentamalleista anna kovin hyvää arviota. Syynä lienee se, että välipohjan jälkituennasta johtuen se ei toimi selkeästi yksi- tai kaksiaukkoisena, vaan siltä väliltä (taulukko 4). Mittauspisteessä 1 laskentamallit 6 ja 9 (massa tasan jakautunut) antavat parhaat laskennalliset tulokset.

3. Kaksiaukkoisena toimivan RL-laattavälipohjan ominaistaajuuden laskentaan soveltuvat lähes yhtä hyvin sekä kahden yksiaukkoisen palkin mallit (mallit 4, 5 ja 6) että jatkuvan kaksiaukkoisen palkin mallit (mallit 7 ja 9), (taulukko 4).
4. Titaniittivälipohjan ominaisjaksoluvun määrittämiseen ei mikään laskentamalli anna kovin hyvää tulosta. Eräänä syynä tähän lienee se, että välipohja on tuettu keskituella liimapuupalkkien ja -pilarien varaan. Näiden värähtely vaikuttanee myös välipohjapalkkien mitattuihin ominaisjaksolukuihin, mutta laskentamallissa tätä ei ole otettu huomioon.
5. Raakavälipohjien ja valmiiden välipohjien ominaisjaksoluvut poikkeavat melkoisesti toisistaan. Tämä koskee sekä laskettuja että mitattuja ominaisjaksolukuja.
6. Valmiin välipohjan analysointiin soveltuva kolmikerroksinen jousimassasysteemi (laskentamalli 5) ei anna mittaustuloksiin paremmin soveltuvia ominaisjaksolukuja kuin yksikerroksinen jousi-massasysteemi (mallit 4, 6, 7 ja 9).
7. Valmiin Titaniittivälipohjan lasketut ominaisjaksoluvut ovat selvästi pienempiä kuin raakavälipohjan lasketut ominaisarvot. Mitattujen arvojen suhteen erot menevät ristiin eri mittauspisteissä saatuja arvoja verrattaessa.

6. YHTEENVETO

Tutkimuksessa esillä olleista laskentamalleista ei yksikään osoittautunut muita selvästi paremmaksi välipohjien ominaisjaksolukujen laskentaan. Käytännön suunnittelutyössä yksinkertainen palkkimalli olettaen joko massan puolikas jännevälin keskelle tai koko massa tasan jakautuneeksi koko jännevälille näyttää antavan riittävän tarkan arvion alimmalle ominaisjaksoluvulle. Raakavälipohjan ylä- ja alapuoliset levykerrokset voidaan tällöin ottaa laskennassa huomioon pelkästään lisämassana muodostamatta kolmikerroksista jousi-massasysteemiä. Kaksiaukkoinen välipohjapalkki voidaan laskentaa varten jakaa kahdeksi yksiaukkoiseksi vapaasti tuetuksi palkiksi. Siten yksiaukkoisten välipohjapalkkien alin ominaisjaksoluku voidaan arvioida yksinkertaisimmin joko laskentamallilla 1 tai 3 ja kaksiaukkoisten alin ominaisjaksoluku joko laskentamallilla 4 tai 6.

Laattamalleja voidaan perustellusti käyttää, jos välipohjan pohjamuoto on säännöllinen suorakaide, massa on tasan jakautunut (pesuhuoneen tasausbetoni, huom.) ja välipohjan jäykkyys palkkeihin nähden kohtisuorassa suunnassa on suuri. Tällöin lähteissä [3], [4] ja [6] esitetyt laskentakaavat ovat riittävän yksinkertaisia.

Valmiin välipohjan alimman ominaisjaksoluvun tulisi olla mahdollisimman korkea. Tähän päästään lisäämällä välipohjan taipumajäykkyyttä (lisäämällä palkkien taivutusjäykkyyttä EI ja lyhentämällä jänneväliä) tai vähentämällä välipohjan massaa. Massan vähentäminen heikentää merkittävästi välipohjan ääneneristävyyttä. Alimman ominaisjaksoluvun tulisi välipohjan päällä liikkumisen aiheuttaman värähtelyn vaikutusten minimoimiseksi olla vähintään 8 Hz. Pyykinpesukoneen aiheuttaman tärinän haittojen minimoimiseksi tulisi välipohjan alimman ominaisjaksoluvun olla vähintään 20 Hz. Mittauksilla on voitu osoittaa, että Ouluun rakennetussa puukerrostalossa (Kiinteistö Oy Puukotka) pyykinpesukoneen tärinä- ja ääniefektit eivät juuri erottuneet taustatärinästä ja -melusta eivätkä olleet haitallisia edes koneen linkousvaiheessa.

LÄHTEET

1. Suomen rakentamismääräyskokoelma, osa B10, *Puurakenteet, ohjeet,* Ympäristöministeriö, Helsinki (1990).
2. RIL 120-1986, *Puurakenteiden suunnitteluohjeet*, Suomen Rakennusinsinöörien Liitto RIL r.y., Helsinki (1986).
3. Eurocode 5 - Design of timber structures - Part 1-1, *General rules and rules for buildings*, ENV 1995-1-1:1993, European Committee for Standardization, Brussels (1995).
4. S. Ohlsson, *Springiness and human-induced floor vibration,* A design guide. Swedish Council for Building Research. Publication D12:1988, Stockholm (1988).
5. S. Palola, *Puukerrostalon rakenneosien värähtely ja ääneneristys*, Diplomityö, Oulun yliopisto, rakentamistekniikan osasto, rakennetekniikan laboratorio, Oulu (1997).
6. H. J. Blass ym., *Timber Engineering STEP 1,* Centrum Hout, Hollanti (1995).

ESTIMATION OF SETTLEMENT OF THE HAARAJOKI TEST EMBANKMENT

A. Näätänen, N. Puumalainen, K. Saarelainen, A. Aalto, M. Lojander, P. Vepsäläinen
Helsinki University of Technology
Department of Civil and Environmental Engineering
Rakentajanaukio 4 A, 02150 Espoo, FINLAND

ABSTRACT

The calculation of settlements is a basic task for geotechnical engineers. This article deals with several different methods for estimating the behaviour of a test embankment. The Haarajoki test embankment will be constructed in the summer 1997. The Finnish National Road Administration is organizing an international competition to calculate settlements of the test embankment. All the calculations presented in this article are made before the construction of the embankment. The suitability of the calculation methods is discussed.

1. INTRODUCTION

The calculation of the settlement of embankments constructed on soft soils is often based on the results of oedometer tests. In most cases this kind of classical analysis is good enough. In some cases (complicated geometry, low factor of safety, horizontal movements...) the classical one-dimensional analysis is not acceptable. A simple modification may be the assumption of two- or three-dimensional waterflow (elastic Terzaghi- or Biot-type consolidation analysis). The parameters are derived from vertical (and horizontal - if available) oedometer test results. The next step may be the use of the Finite Element Method, and then the parameters are derived from the oedometer and triaxial test results. FEM-

analyses gives also better understanding on horizontal displacements and pore pressures in the ground.

The Finnish National Road Administration is organising an international competition to calculate settlements concerning the Haarajoki Test Embankment. The goals of the competition are to improve the standard of geotechnical calculations, and to test the usability of new calculation methods, and to improve the mode of presentation of calculations [1]. The test embankment will be constructed as part of the noise barrier during July and August 1997. The construction of the test embankment is part of the Road Administration's strategic project Road Structure Research Programme (TPPT). The Laboratory of Soil Mechanics and Foundation Engineering of Helsinki University of Technology has taken part in the soil sampling and laboratory testing. Calculations are also part of the TPPT-project (together with the Technical Research Centre of Finland) for developing the settlement calculation method which is based on the water content of the ground.

2. THE HAARAJOKI TEST EMBANKMENT

The embankment is 100 meters long and it will be constructed on clay ground in part without ground improvement and partly onto vertically drained area [1]. The geometry of the embankment is shown in Fig. 1. This article deals with that part of the embankment which is made without ground improvement.

Classification properties of the ground are shown in Fig. 2. The subsoil consist of over 20 meters of overconsolidated clay and silt. Overconsolidation is so low that in some layers the increase of the vertical pressure of the embankment will reach the pre-consolidation pressure. The ground water table is on the ground surface, and excess pore pressure is measured in the ground (-3...10kPa). Most compressible layers are at the depth of 2...10m.

Numerous measuring devices (settlement plates, piezometers, inclinometers, extensometers, total stress gauges) are installed under test embankment for monitoring the vertical and the horizontal displacements and the pore pressures.

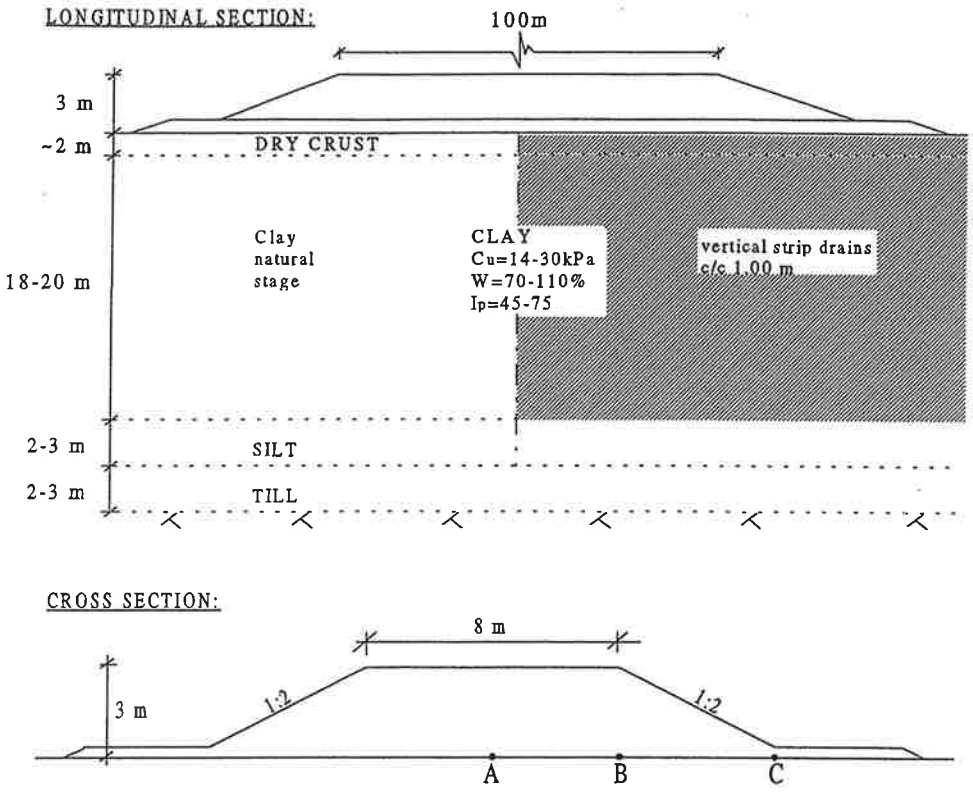


Figure 1. The Haarajoki Test Embankment [1].

The construction of the embankment will take three weeks in six stages of 0,5m filling. The fill material is gravel...sandy gravel. The bulk density of the fill material is 20...22 kN/m^3 (in the calculations the value is 21 kN/m^3).

3. CALCULATION METHODS

3.1 Calculation programs and parameters

Six calculation programs were used for analysing the stress-strain-time behaviour of the Haarajoki test embankment : RAKPA, KONSOL, EMBANKCO, Sage CRISP, ZSOIL and PLAXIS.

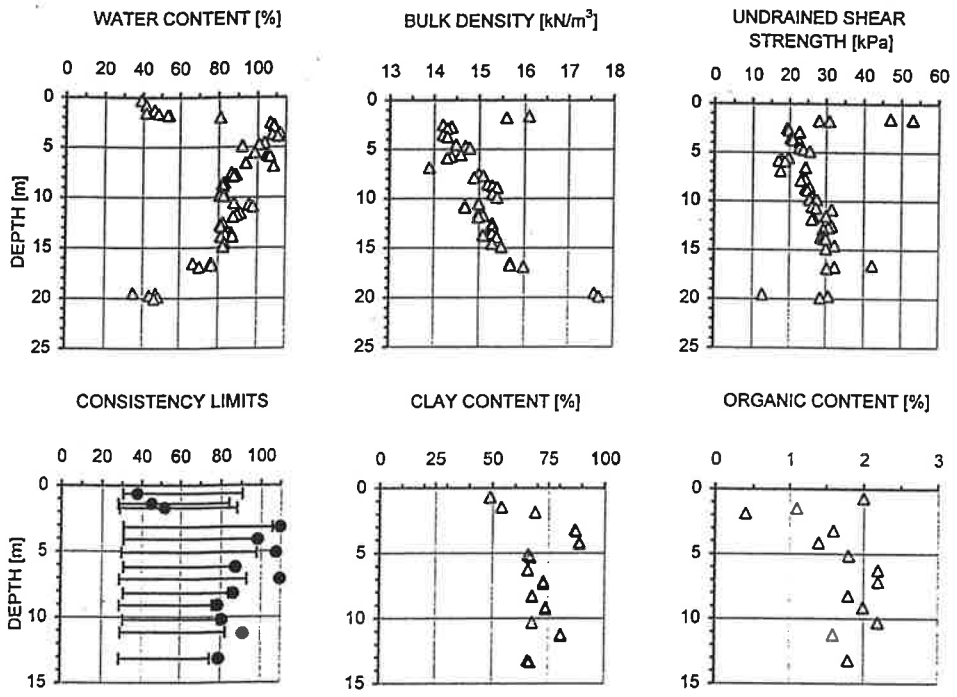


Figure 2. The Haarajoki Test Embankment. The classification properties.

Program RAKPA is based to the conventional Finnish tangent-modulus method for total primary consolidation settlement calculation. KONSOL is based to the modified Terzaghi one-dimensional time-settlement calculation with the finite element method. Programs were developed by Pauli Vepsäläinen in 1980's and 1990's.

EMBANKCO is a swedish one-dimensional settlement calculation program which is based on finite difference method (Användarhandbook [2], Larsson et al SGI 13 [3]. EMBANKCO contains also secondary consolidation.

Sage CRISP [4]. is a Finite Element Program which is based on the Critical State Model. Consolidation analysis is based on Biot's two-dimensional and three-dimensional consolidation theory. Program CRISP was originally developed in the Cambridge University

in England. Sage CRISP was developed from the program CRISP by adding graphical pre- and post-processors.

ZSOIL [5] is a Swiss Finite Element Program which is suitable for stress-strain-time analysis and stability calculations in undrained, consolidation and drained state. Primary consolidation is based on Biot's theory. Secondary consolidation model (Kelvin) was not used here.

PLAXIS [6] is a Dutch Finite Element Program which is suitable for stress-strain-time analysis and stability calculations in undrained, consolidation and drained state (like programs Sage CRISP and ZSOIL). Primary consolidation is based on Biot's theory.

The parameters for the calculations are shown in Table 1. RAKPA and KONSOL use parameters which are derived from oedometer test results. Constant Rate of Strain (CRS) oedometer tests are most suitable for determining parameters for EMBANKCO. FEM-programs need also triaxial test results. Main differences in giving the input parameters are concentrated to the pre-consolidation pressure or overconsolidation ratio.

3.2 RAKPA and KONSOL

RAKPA is a conventional program for total primary consolidation settlement calculation. It needs the usual settelement parameters of the tangent modulus method used in Finland. Program RAKPA is acting as an input program for the terrace, material and loading data for the program KONSOL.

KONSOL is a FEM-program for time-settlelement calculations of layered ground. It is based to the one-dimensional primary consolidation theory of Terzaghi. The theory is modified so that it includes also the over-consolidated part of the time-settlement behaviour. Besides of the data from RAKPA, program KONSOL needs the coefficients of consolidation for over- and normally consolidated stages, seepage boundary conditions and the load-time history of

Table 1. The material parameters for the calculation programs.

a) RAKPA and KONSOL b) EMBANKCO c) CRISP, ZSOIL and PLAXIS

(in program PLAXIS σ_p from table a is used instead of p_c in table c)

a)

layer	depth m	γ kN/m^3	σ_p kPa	m_1	β_1	m_2	$c_{vNK\ STD}$ m^2/a	$c_{vYK\ STD}$ m^2/a	$c_{vNK\ CRS}$ m^2/a	$c_{vYK\ CRS}$ m^2/a
1	0-1	17	80	28	0,25	105	1,0	5,3	1,0	5,3
2	1-2	17	60	26	0,46	57	1,0	1,9	2,5	2,5
3	2-5	14	52	4,2	-1,1	50	0,10	9,0	0,20	9,0
4	5-7	14	52	4,9	-1,0	59	0,10	1,7	0,20	1,7
5a	7-8	15	60	4,0	-0,75	58	0,05	7,9	0,24	7,9
5b	8-9	15	70	4,0	-0,75	58	0,05	7,9	0,24	7,9
5c	9-10	15	82	4,0	-0,75	58	0,10	6,5	0,42	6,5
6	10-12	15	95	2,0	-1,26	80	0,10	2,7	0,50	2,7
7	12-15	15	99	2,7	-0,86	72	0,10	4,8	0,10	4,8
8	15-18	16	129	3,7	-0,60	58	1,3	12	1,3	12
9	18-22,2	17	105	30	0,5		10	10	10	10

b)

layer	depth m	point m	γ kN/m^3	σ_c' kPa	M_0 kPa	M_L kPa	σ_L' kPa	M'	a	$\alpha_{s\ max}$ *10^{-2}	β_{as}	k_i *10^{-10} m/s	β_k
1	0-1	0,60	17	80	6940	3167	180	12	1000	0,12	0	20,2	9,87
2	1-2		17	80	6940	3167	180	12	1000	0,12	0	20,2	9,87
3	2-5	2,30	14	67	6150	89	77	27,6	100	1,76	0	21,1	4,64
		4,33	14	76	5950	28	77	18,6	100	1,15	0	9,46	3,77
4	5-7	6,33	14	72	6280	62	77	18,6	100	1,03	0	11,7	4,26
5a	7-8	7,24	15	60	7490	350	92	18,0	100	1,48	0	7,24	3,94
5b	8-9	8,22	15	87	6810	220	103	22,2	100	1,48	0	7,83	4,14
5c	9-10	9,24	15	92	7140	514	114	12,5	100	2,26	0	8,59	4,12
6	10-12	10,22	15	109	6250	67	114	18,3	100	0,58	0	15,5	4,57
7	12-15	14,10	15	130	4640	335	167	18,6	100	0,43	0	6	4,18
8	15-18	15,10	16	137	6100	470	200	13,4	100	0,99	0	6	3,63
		17,60	16	129	6780	560	180	15,0	100	0,30	0	9	3,52
9	18-22,	18,00	17	129->90					100		0		
		19,69	17	104	11480	3300	225	23,8	100	0,30	0	50	3,52
		22,20	17	121,4	11480	3300	225	23,8	100	0,30	0	100	3,52

c)

layer	depth m	γ kN/m^3	M	ϕ' °	ν	E kPa	κ	λ	e_0	e_{cr}	k_x *10^{-4} m/d	k_y *10^{-4} m/d	p_c kPa
1	0-1	17	1,5	36,9	0,38	6410	0,007	0,1	1,4	1,78	13	13	80
2	1-2	17	1,5	36,9	0,38	3900	0,013	0,1	1,4	1,76	13	13	65
3	2-5	14	1,5	36,9	0,38	3150	0,028	1,08	2,9	6,68	1,56	1,3	65
4	5-7	14	0,93	23,7	0,1	4330	0,068	1,86	2,8	9,51	1,21	0,86	72
5a	7-8									4,68			72
5b	8-9	15	1,07	27	0,1	4890	0,075	0,65	2,3	4,84	1,38	0,69	91,5
5c	9-10									4,96			111
6	10-12	15	1,07	27	0,28	5600	0,066	1,23	2,2	7,22	2,59	1,3	114
7	12-15	15	1,15	28,8	0,28	12000	0,027	1,09	2,2	6,63	2,59	1,3	114
8	15-18	16	1,5	36,9	0,28	5870	0,061	0,48	2	3,98	8	1,12	114
										4,11			150
9	18-22,2	17	1,5	36,9	0,28	5870	0,009	0,1	1,4	1,84	80	80	150

distributed surface loads. The initial excess pore pressure condition is calculated through excess vertical stresses to simulate typical Finnish normally consolidated clay conditions. Excess vertical stress influence values are calculated already in RAKPA by the Boussinesq method. The finite element method and the time integration by the implicit method is used to calculate excess pore pressures, effective vertical stresses, vertical strains and settlements with time. The program makes the one-dimensional finite element net automatically.

3.3 EMBANKCO

EMBANKCO is a swedish program for settlement calculations of road embankments on fine grained soils. EMBANKCO contains one-dimensional primary and secondary consolidation model. The stress distribution is calculated by using the theory of elasticity. The parameters are determined from the results of constant rate of strain (CRS) oedometer tests. The coefficient of secondary consolidation is determined from long duration incremental oedometer tests. EMBANKCO is based on empirical observations of changing values of tangent modulus (M), permeability (k) and coefficient of secondary consolidation (α_s) (Fig. 3).

3.4 Sage CRISP

Sage CRISP is developed from the program CRISP (CRitical State Program, Univ. of Cambribge). The soft layers are simulated by the Modified Cam Clay Model and the embankment material by ideally elastic-plastic Mohr-Coulomb model. Sage CRISP uses the displacement FEM-scheme coupled with Biot's consolidation and implicit time integration.

The finite element net consists only half of the symmetric embankment and soft soil layers with 575 eight-noded quadrilateral elements and 630 nodes. Input data is needed for net geometries and topologies, material parameters for soil layers and in-situ conditions for effective vertical and horizontal stresses, pore pressures and hydrostatic overconsolidation pressures (the latter for MCC-model). Furthermore there is need for displacement and seepage boundary conditions, the vertical boundaries of the net are assumed impervious.

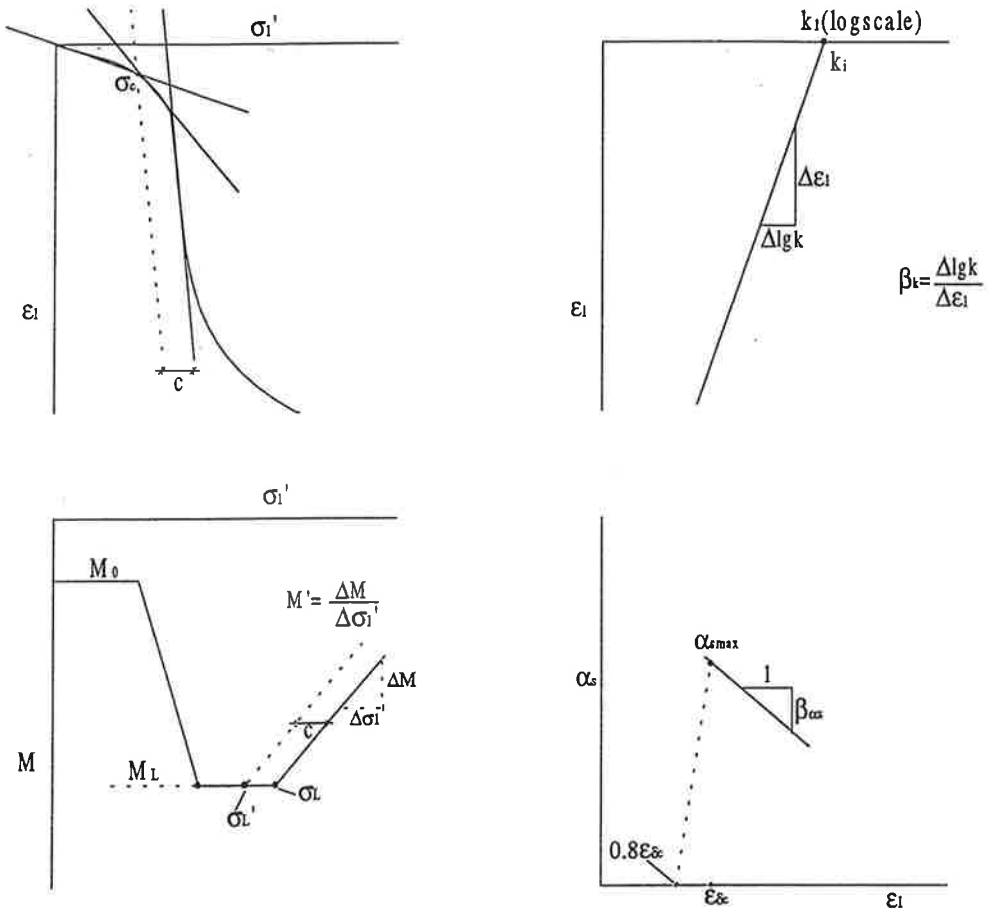


Figure 3. The material parameters for EMBANKCO.

The construction of the test embankment is simulated by adding embankment elements with the supposed time schedule (35 days), so the consolidation process is happening already in the construction stage.

3.5 ZSOIL

The nonlinear finite element scheme incorporated in ZSOIL (version 3.1) uses the four-node isoparametric quadrilateral element with bilinear interpolation function. The problem was handled in plane strain. To reduce the number of elements, the shape of the embankment was

assumed symmetric and only half of it was included in the model. The vertical sides and the bottom of the mesh were impermeable, the bottommost soil layer modelled consists permeable silt. The finite element net consists of 700 elements and 760 nodes.

The initial stress conditions were automatically evaluated. The isotropic hardening elastoplastic model chosen for the problem was the Extended Drucker-Prager yield criteria with cap-closure 3N. 3N closure adjusts the size of the Drucker-Prager criterion to the Mohr-Coulomb criterion with the following assumptions: Failure under plain-strain conditions and non-associative plastic potential (plane strain default).

The cap closure require the specification of the following parameters: Initial void ratio e_0 , pre-consolidation pressure p_c , modulus E and compression index λ. The parameters were evaluated from triaxial test results or incrementally loaded oedometer test results. The value of pre-consolidation pressure is constant in the calculation layer. In thick normally consolidated layers the pre-consolidation pressure is increasing when the depth is increasing. This error is corrected (z = 5...7 m) by dividing the layer to three thin layers. In the analysis the value of cohesion was assumed to be zero. The primary consolidation is calculated by using Biot's consolidation theory.

The construction of the embankment was modelled by Mohr-Coulomb model. The embankment was constructed on 6 steps during 35 days. Subsoil was consolidated also during the construction of the embankment.

ZSOIL calculates automatically initial horizontal and vertical stresses before further analysis. K_0 is calculated by using the Poisson's ratio. Small values of Poisson's ratio affect incorrectly very small horizontal stress. In this case this effect was avoided by using Poisson's ratio which was calculated from the value of effective friction angle.

The initial excess pore pressure was determined in all the nodes of macro elements. Instead of hydrostatic pore pressure and total bulk densities the effective bulk densities were used.

3.6 PLAXIS version 6.1 (Professional version)

PLAXIS is a non-linear finite element program, which is used for different geotechnical problems. The problem is handled either in plane-strain or axi-symmetric conditions.

The geometry is modelled by a suitable finite element mesh. The user devides the whole mesh into mesh blocks and every mesh block into smaller quadrilaterals. The mesh is generated automatically by dividing each quadrilateral in two triangular finite elements. Every triangular element is either 6- or 15-noded, depending on the user. Displacements are calculated at each individual node. Stresses are calculated at Gaussian integration points ('stress points'). A 15-noded triangular element contains 12 and 6-noded 3 stress points.

The consolidation analysis is based on Biot's theory and only the primary consolidation is possible. Darcy's law for fluid flow is also assumed. The permeabilities of soil layers can be given both in horizontal and in vertical directions.

The shape of the test embankment was assumed symmetric and only a half of it was included in the model. The problem was handled in plane-strain state. The whole profile (mesh) was devided into five mesh blocks and it included 780 triangular finite elements. These elements were 6-noded because of the great number of elements.

The material model used for clay layers was isotropically hardening elasto-plastic Modified Cam-Clay Model (MCCM). In this model the yield surface represents an ellipse in p'-q-plane. The material is behaving elastically within the yield surface, whereas stress paths that tend to cross the boundary generally give both elastic and plastic strain increments. Parameters needed by using Modified Cam-Clay Model were determined by oedometer and triaxial tests.

The initial stress state in lightly overconsolidated subgrade layers is defined by using Pre-Overburden Pressure, POP ($= \sigma'_c - \sigma'_{v0}$). This value gives a more homogenous stress state in

clay layers than the normally used Over-Consolidation Ratio, OCR (= σ'_c / σ'_{v0}). To trying to have as homogenous stress state in every single layer as possible the third layer (z = 2-5 m) is devided into three thinner layers (Fig. 4). After determining these POP-values and giving K_0 -values (K_0 -values were calculated by using effective friction angle) the initial stress conditions were evaluated automatically.The difference from static pore water pressure was handled by changing unit weights so that the effective vertical stress σ'_{v0} equals the real in-situ stress.

The material model chosen for the embankment was the Mohr-Coulomb model. The embankment elements were activated at six stages so that the total building time was 35 days. The settlements at building time were due to both undrained and consolidation settlement.

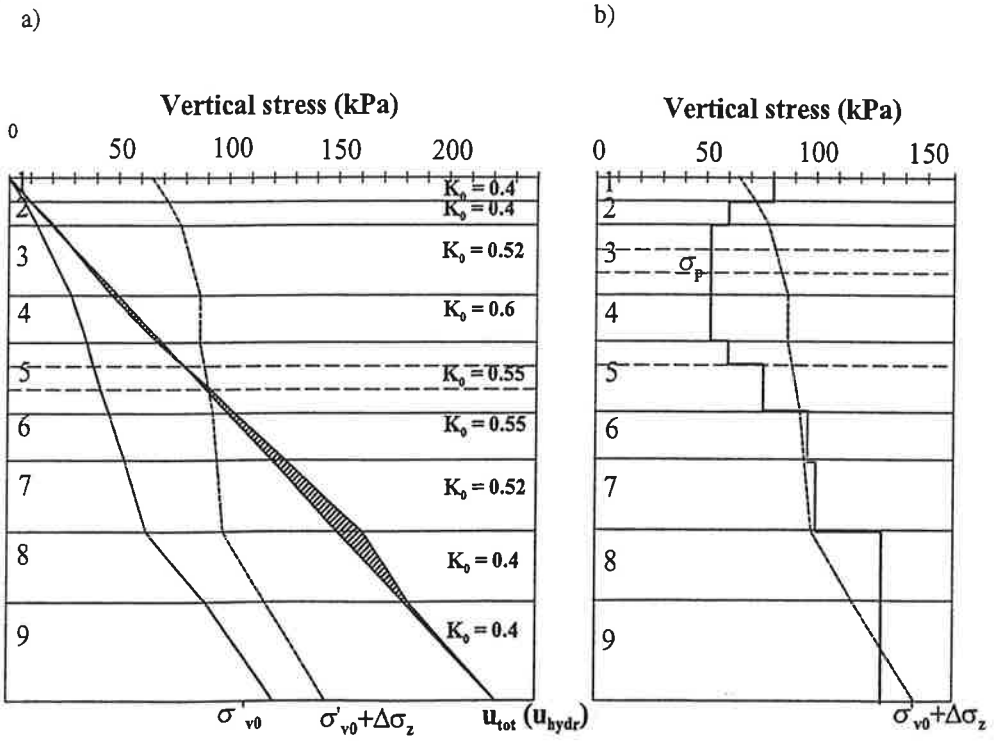


Figure 4. The Haarajoki Test Embankment.

a) Effective vertical stresses. Pore pressure b) Pre-consolidation pressure.

4. RESULTS AND DISCUSSIONS

The summary of settlements from different calculations is presented in table 2. It is seen that the differences are quite remarkable depending on the calculation method or program at the settlement-time scale. Moreover, the calculated long-time settlements (30 years) are from 172 mm to 831 mm, which anticipates that even the estimation of the total primary consolidation settlement correctly may be very difficult.

Table 2. The Haarajoki Test embankment.

a) The calculated settlements. b) The calculated horizontal movements.

a)

	0 months			6 months			12 months			18 months			24 months			10 years			30 years		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Sage CRISP	117	103	40	156	138	63	168	149	72	177	158	80	183	164	85	224	203	117	244	222	133
ZSOIL	121	96	30	134	108	42	139	113	47	142	117	50	145	119	52	165	139	70	172	145	76
PLAXIS	151	128	44	241	199	72	279	230	87	309	255	98	333	275	108	462	384	155	477	397	160
EMBANKCO	91	82	39	192	130	75	219	150	89	241	165	97	265	179	103	470	292	123	633	384	124
KONSOL CR	42	41	19	93	87	43	120	113	58	140	132	69	156	146	78	448	329	140	831	689	183
KONSOL ST	42	40	19	89	84	42	116	109	55	135	127	66	150	141	74	359	264	137	656	543	182

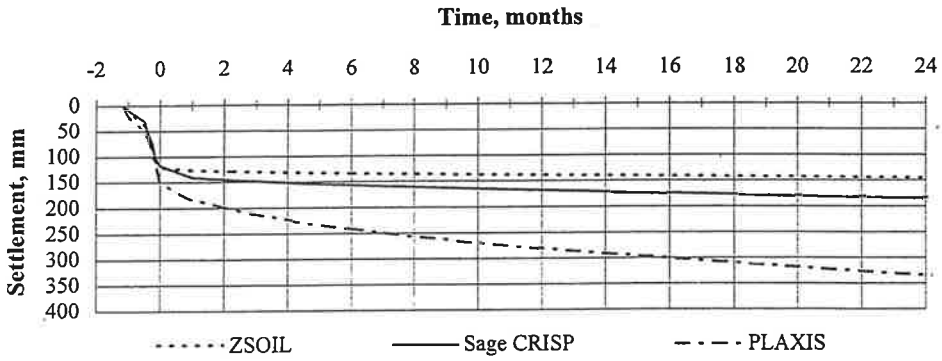
b)

	0 months		6 months		12 months		18 months		24 months		10 years		30 years	
	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Sage CRISP	9	18	11	20	11	20	11	20	11	20	11	20	12	21
ZSOIL	18	25	18	24	18	24	18	24	18	24	18	24	18	24
PLAXIS	23	40	36	52	40	55	41	58	44	59	51	66	52	67

The graphical presentation of calculated settlements at the center line of the test embankment during the first two years is shown in figures 5 and 6. In figure 5 are the results of FEM-calculations (programs Sage CRISP, ZSOIL and PLAXIS) and in fig. 6 the results of perhaps more conventional methods (programs KONSOL and EMBANKCO). The KONSOL calculations were made by two ways: KONSOL STD means that all the settlement parameters (including the coefficients of consolidation) used are from standard oedometer tests, and in KONSOL CRS-calculations the coefficients of consolidation are from CRS-oedometer tests and all other settlement parameters from standard oedometer tests (table 1).

The results of EMBANKCO-calculations in figure 5b include the effect of secondary consolidation.

a)



b)

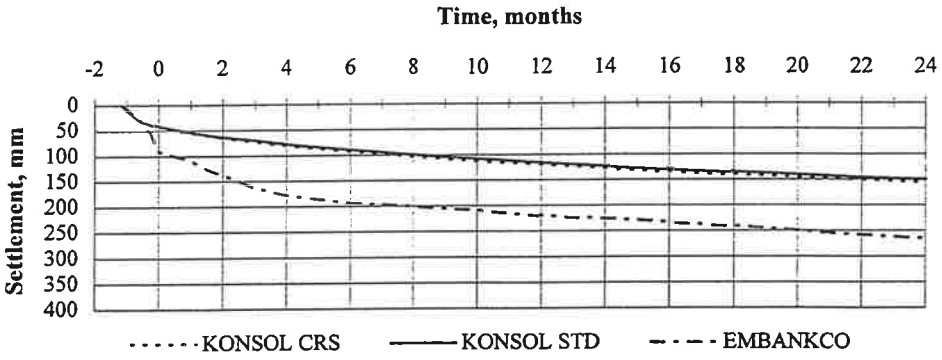


Figure 5. The Haarajoki test embankment. Calculated settlements 0 - 2 years.

Calculated settlements during a longer time period (30 years) are presented in fig. 6. The graphical presentation includes the results of EMBANKCO-calculations with and without secondary consolidation, KONSOL STD and KONSOL CRS-results and results from PLAXIS-calculations. Settlements presented are from the center line of the embankment. As is seen from the table 2a, calculated settlements by programs Sage CRISP and ZSOIL are much lower than those presented in fig. 6.

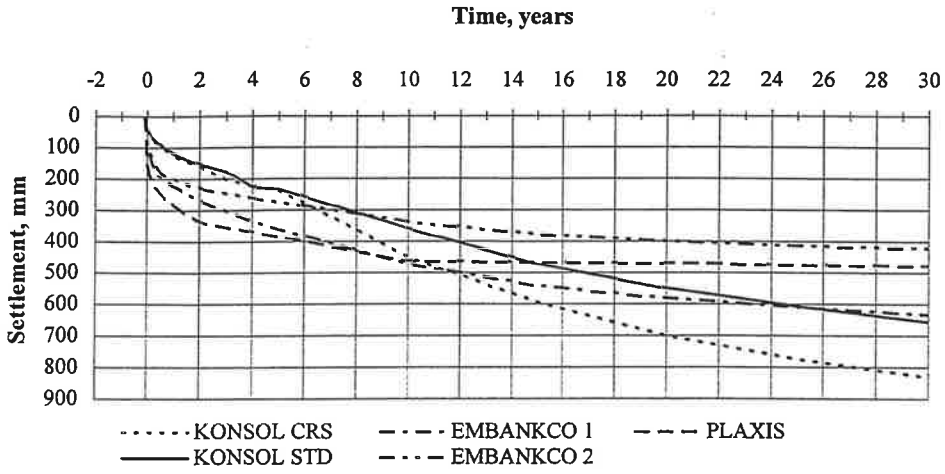


Figure 6. The Haarajoki Test embankment. The calculated settlements 0-30 years.

Results of the calculated horizontal movements at the frontier line between the test embankment and the natural ground are presented in the table 2b. Point B means the distance of 4 meters and point C the distance of 9 meters from the center line. The horizontal movements were calculated only by the FEM-programs Sage CRISP, ZSOIL and PLAXIS (the estimation is not possible by one-dimensional programs KONSOL and EMBANKCO). Differences in the results are remarkable even in this case, but the usually accepted tendency - the larger vertical settlements mean the larger horizontal movements - is best seen from the PLAXIS-results.

5. CONCLUSIONS

The Finnish National Road Administration has arranged an international competition to calculate settlements of Haarajoki test embankment. The settlement during the competition time (2 years) will be only 15....33 cm with different calculation methods. The calculations were made also for 30 years, and then differences in calculated settelements are significant (17....83cm). All the programs used here are suitable for analyzing the behaviour of an embankment constructed on soft ground.

RAKPA and KONSOL are especially made for this kind of analyses. KONSOL is used for estimating primary settlement. The parameters are determined with standard oedometer tests or with continuous loading oedometer tests. The need for calculating secondary consolidation also could be seen from pore pressure data.

EMBANKCO is especially made for this kind of analyses. EMBANKCO contains also a secondary consolidation model. Primary and secondary consolidation are occuring at the same time. The values of the calculation parameters are dependent on stress or strain. The parametres are determined with standard oedometer tests (secondary consolidation) and with continuous loading oedometer tests. All the parameters can be determined also with standard oedometer tests. The program contains also empirical parameters from swedish data for preliminary analyses before laboratory testing.

Sage CRISP is suitable for analyzing stresses, strains and pore pressures with time. The determination of the parameters from oedometer and triaxial test results is laborous and difficult. The determination of the pre-consolidation pressure is most important for getting reliable results. The anisotopic initial yield surface of Haarajoki clay (and most Finnish clays) is totally different than the Modified Cam Clay yield surface which is used in calculations. The adjustment of the input value of the pre-consolidation pressure may cause underestimation on the settlements.

ZSOIL is originally not made for analyzing soft clay behaviour, but the modern version of the program is suitable for this type of analyses. ZSOIL contains also model for secondary consolidation. In this case only primary consolidation was calculated. The use of ZSOIL needs special knowledge on the program and the determination of calculation parameters. Then it is a powerful tool for analyzing also complicated problems.

PLAXIS is suitable for primary consolidation analyses of the embankment. The program uses the pre-consolidation pressure from oedometer tests for solving the overconsolidation ratio of soil layers. This is maybe the main reason for larger settlements than in other FEM analyses.

The project for analyzing the behaviour of the Haarajoki test embankment will now continue with efforts to approximate the settlements by using water content data. Another task is analyzing the embankment with vertical drains. Then the most interesting stage will be the comparison of the measured values to the calculated values guring the years 1997-1999.

REFERENCES

1. Finnish National Road Administration (FinnRA). *Competition to Calculate Settlements at the Haarajoki Test Embankment.* Competition Programme. 1997
2. *Användarhandbok. Program EMBANKCO. Version 1.02. Program för sättningberäkning för vägbank på finkornig jord.* 1994.
3. Larsson, R., Eriksson, L., Bengtsson, P-E. *Sättningsprognoser för vägbankar på lös finkornig jord.* SGI Information 13. 1993.
4. *Sage CRISP User Guide and Technical Reference Guide. Version 3.02 Update.* 1995.
5. *ZSOIL V 3.11. User's guide.* Zace Services Ltd, Lausanne. 1995.
6. *PLAXIS Finite Element Code for Soil and Rock Analyses. Version 6.* P.A. Vermeer, R.B.J. Brinkgreve (eds). A.A. BALKEMA. Rotterdam 1995.

Shape calculus and Babuska's paradox

T. Tiihonen
Department of Mathematics, University of Jyväskylä
Box 35, FIN–40351 Jyväskylä, Finland

Abstract

Treatment of domains with curved boundaries in finite element method leads often to questions concerning the continuous dependence of the solution on the geometry. We propose here to use the techniques arising from optimal shape design to estimate the error due to approximation of the geometry. We provide examples where shape differentiability leads to useful error estimates. Likewise, some examples are given where the lack of shape differentiability indicates also lack of continuous dependence on geometric data.

1 Introduction

This paper is motivated by the dilemma of 'smooth polygonal domains' related to error analysis of the Finite Element Method. The dilemma is that the finite element methods are naturally formulated in polygonal (or more generally in piecewise polynomial) geometries. On the other hand, the abstract error estimates rely on interpolation error estimates that require smoothness of the solution that is typically achieved only in regular geometries.

In the literature this question has been treated in several ways. Perhaps the most popular approach is that of the above mentioned smooth polygonal domains. That is, the analysis is carried out assuming that the domain is polygonal and hence the grid fits exactly to the domain, and at the same time the solution is regular enough for the optimal interpolation estimates. At the other extreme are the works where the curved boundary is captured more or less exactly by introducing corresponding curved elements, [2], [3], [9], [10]. In this case the analysis can be made rigorous but the price to pay is more complicated local analysis and implementation. Finally, there exists quite a number of papers where the smooth domain is approximated by a polygonal one which is then triangulated. In many of the papers the error analysis is based on some specific feature, like the possibility to extend the FE-solution outside the polygonal domain. Also the analysis of the approximation of geometry is generally interwoven with the analysis of the FE approximation properties.

In this paper we introduce an approach where the approximation of geometry is detached from the FE-analysis. This means that we shall analyze the error due to approximating the original problem in smooth domain by an auxiliary problem in (polygonal) approximate domain. Then, the error for the finite element approximation of the auxiliary problem is analyzed, bearing in mind that the auxiliary solution is close to the original, smooth one. The first error is estimated using the techniques familiar from shape optimization [5] where the question of continuous dependence with respect to variations in geometry is a key issue. The question of continuous dependence of the solution on the geometric data dates back to at least Hadamard's times. Most of the analysis has been qualitative. Some quantitative works were published in early 70's, [6], [8] with the motivation arising from finite element error estimates. Then, it seems, the issue was forgotten.

The contents of the paper can be briefly summarized as follows. In chapter 2 we introduce the strategy in abstract framework. Then in chapter 3 we consider a model second order elliptic problem for which we show the continuous dependence of the solution under polyhedral approximation of the boundary. Corresponding finite element error estimates are also formulated. In the subsequent chapters we consider second order systems and a fourth order problem where polyhedral approximation of the boundary turns out to be much more delicate issue. In particular we comment Babuska's famous counter-example, [1], [4] from the point of view of shape derivatives. This paper is an abridged version of [7].

2 Abstract formulation

The aim is to study the dependence between the solution u of the variational problem defined in a smooth domain Ω and its finite element approximation u_h defined in approximate domain Ω_h. To be able to compare u and u_h we have to be able to prolongate one of the two to the domain of definition of the other. As prolongation of FE- functions is difficult (as FE-functions) in the general case we choose to prolongate the original solution from Ω to Ω_h by some prolongation operator. The idea behind error estimation is to introduce an auxiliary problem defined in Ω_h to be able to separate the approximation of geometry from approximation by FE spaces.

Thus, let the original problem be given as

$$a(u, w) = \langle f, w \rangle \qquad \forall w \in V$$

for $f \in V'$, $u \in V = V(\Omega)$. We introduce an auxiliary problem

$$\hat{a}(\hat{u}, w) = \langle \hat{f}, w \rangle \qquad \forall w \in \hat{V}$$

with $\hat{V} = \hat{V}(\hat{\Omega})$. The bilinear forms a and $\hat{a}$ are assumed to be continuous and V (resp. $\hat{V}$) elliptic. The auxiliary problem is defined so that its FE discretization gives the discrete problem

$$\hat{a}(u_h, w_h) = \langle \hat{f}, w_h \rangle \qquad \forall w_h \in V_h \subset \hat{V}.$$

We want now to estimate the error between u and u_h. Because of different domains of definition we have to consider an extension of u, $\tilde{u}$. Let now $\|\cdot\|$ be some norm for functions defined in Ω_h. Then by triangle inequality

$$\|\tilde{u} - u_h\| \leq \|\tilde{u} - \hat{u}\| + \|\hat{u} - u_h\|.$$

Here the first part corresponds to the error due to geometry, whereas the second part is standard FE-approximation error (but now in polygonal domain).

Assume now that we have a quasi-optimality result with respect to $\|\cdot\|$ norm

$$\|\hat{u} - u_h\| \leq C \inf_{w_h \in V_h} \|\hat{u} - w_h\|. \tag{1}$$

Then using triangle inequality again we can estimate

$$\|\hat{u} - w_h\| \leq \|\hat{u} - \tilde{u}\| + \|\tilde{u} - w_h\|.$$

Hence,

$$\|\tilde{u} - u_h\| \leq C_1\|\tilde{u} - \hat{u}\| + C_2 \inf_{w_h \in V_h} \|\tilde{u} - w_h\|. \tag{2}$$

In order to derive useful concrete error estimates from (2) we have to make sure that the following conditions are satisfied.

1. $\hat{f}$, $\hat{a}$ and the extension operator are chosen so that we get an appropriate estimate for $\|\tilde{u} - \hat{u}\|$.
2. Extension is defined so that $\tilde{u}$ has the regularity needed for interpolation estimates. That is, $\tilde{u}$ should be at least piece-wise smooth.
3. $\hat{f}$ and $\hat{a}$ should be natural extensions of f and a as they will be used to construct the discrete problem.
4. The quasi-optimality estimate (1) is valid uniformly for all Ω_h.

The condition 1 above is best satisfied when the auxiliary problem is defined simply by mapping Ω to $\hat{\Omega}$ and defining the 'extensions' using the same transformation. Then, of course $\|\tilde{u} - \hat{u}\| = 0$. On the other hand the mapping ends up to the coefficients of the auxiliary problem and its discretization complicating thus both error analysis and implementation. This is what happens in practice with curved elements.

3 Model problem

Let us consider the following model problem

$$-\Delta u = f, \qquad \text{in } \Omega, \tag{3}$$

$$u = u_0, \qquad \text{on } \Gamma_D, \tag{4}$$

$$\frac{\partial u}{\partial n} = g, \qquad \text{on } \Gamma_N, \tag{5}$$

where $\Omega \subset D \subset R^n$ is a smooth domain with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\bar{\Gamma}_D \cap \bar{\Gamma}_N = \emptyset$.

We assume that $f \in L^2(D)$, $g \in H^1(D)$ and $u_0 \in H^2(D)$. Then it is well known that the problem (3) has a unique solution $u \in H^2(\Omega)$, $u - u_0 \in H^1(\Omega; \Gamma_D)$.

We shall study the dependence of u on the shape of Ω. For that we construct a family of domains Ω_t as follows: let us choose a vector field $V \in C^1(D : R^n)$ and introduce a deformation map $T_t : x \to x + V(x)$ which is injective for small t. We denote $\Omega^t = T_t(\Omega) = \{x + tV(x) \mid x \in \Omega\}$. In Ω^t we define the state problem as follows:

$$-\Delta u_t = f, \qquad \text{in } \Omega^t, \tag{6}$$

$$u_t = u_0, \qquad \text{on } \Gamma_D^t, \tag{7}$$

$$\frac{\partial u_t}{\partial n} = g, \qquad \text{on } \Gamma_N^t. \tag{8}$$

We address the question of continuous dependence using the concept of shape derivative. We define the shape derivative of u to direction V as the limit

$$u'_V = \lim_{t \to 0+} \frac{\tilde{u}_t - u}{t}\bigg|_\Omega$$

where $\tilde{\cdot}$ denotes any regularity preserving extension from Ω^t to D. The limit does not depend on the choice of the extension.

Theorem 3.1. *Under the above assumptions there exists a shape derivative $u'_V \in H^1(\Omega)$. Moreover, u'_V is the unique solution of the problem*

$$-\Delta u'_V = 0, \qquad \text{in } \Omega, \tag{9}$$

$$u'_V = -\frac{\partial u - u_0}{\partial n}\langle V, n\rangle, \qquad \text{on } \Gamma_D, \tag{10}$$

$$\frac{\partial u'_V}{\partial n} = -\nabla_\Gamma \cdot (\langle V, n\rangle \nabla_\Gamma u) + (f + Hg + \frac{\partial g}{\partial n})\langle V, n\rangle, \qquad \text{on } \Gamma_N. \tag{11}$$

Here ∇_Γ denotes the tangential part of ∇, that is, $\nabla_\Gamma u = \nabla u - \partial u/\partial n n$.

The proof of this theorem can be obtained by combining the proofs of Propositions 3.1 and 3.3 of [5]. This result is almost what we need to analyze the error due to polyhedral approximation of the geometry. Namely, let Ω_h be a polyhedral approximation of Ω. If h is small enough, we can write $\partial\Omega_h = \{x + hV_h(x) \mid x \in \partial\Omega\}$ where $\|V_h\|_{L^\infty} \le Ch$, $\|V_h\|_{W^{1,\infty}} \le C$. Clearly, the problem (9) is well defined for such V_h. Moreover, we can estimate the norm of u' as follows:

$$\|u'\|_{1,\Omega} \le C(\|u'\|_{1/2,\Gamma_D} + \|\frac{\partial u'}{\partial n}\|_{-1/2,\Gamma_N})$$

Now,

$$\begin{aligned}
\|u'\|_{1/2,\Gamma_D} &\le \|-\frac{\partial(u-u_0)}{\partial n}\langle V,n\rangle\|_{0,\Gamma_D}^{1/2}\|-\frac{\partial(u-u_0)}{\partial n}\langle V,n\rangle\|_{1,\Gamma_D}^{1/2} \\
&\le \|-\frac{\partial(u-u_0)}{\partial n}\|_{1,\Gamma_D}\|\langle V,n\rangle\|_{L^\infty}^{1/2}\|\langle V,n\rangle\|_{W^{1,\infty}}^{1/2} \\
&\le Ch^{1/2}
\end{aligned}$$

provided that u and u_0 belong to $H^{5/2}(\Omega)$. Similarly,

$$\|\frac{\partial u'}{\partial n}\|_{-1/2,\Gamma_N} \leq Ch^{1/2}$$

if $u \in H^{5/2}(\Omega)$, $f \in H^{1/2}(\Omega)$ and $g \in H^{3/2}(\Omega)$.

This suggests the estimate $\|u_{\Omega_h} - u_\Omega\|_1 \leq Ch^{3/2}$ in the case of sufficiently smooth data. However, the above development is so far only formal as the velocity field V does not satisfy the conditions of the shape differentiability proof of [5]. Repeating the argument in the case of $V \in W^{1,\infty}$ one can, however, prove the following result.

Theorem 3.2. *[7] Let $V \in W^{1,\infty}(D)$. Then if u and u_0 belong to $H^{5/2}$, $f \in H^1(\Omega)$ and $g \in H^1(\Omega)$ and $\tilde{u}$ is an $H^{5/2}$ extension of u to D, it holds*

$$\|u_t - \tilde{u}\|_{1,\Omega^t} \leq Ct\|V\|_{L^\infty}^{1/2}\|V\|_{W^{1,\infty}}^{1/2}.$$

If we now apply the above result to a smooth domain and its polyhedral approximation, we obtain an $O(h^{3/2})$ estimate for the energy norm of the error in the solution. Similar estimates have been obtained already by Strang and Berger, [6] and Thomee, [8] in the early seventies under the assumption that the domain to be approximated was convex.

For the purpose of finite element error estimates in energy norm an $O(h)$ estimate is sufficient. It can be obtained quite easily with weaker regularity of the data.

Theorem 3.3. *Let $\tilde{u}, u_0 \in H^2(D)$, $f \in L^2(D)$ and $g \in H^1(D)$. Then*

$$\|u_t - \tilde{u}\|_{1,\Omega^t} \leq Ct\|V\|_{W^{1,\infty}}.$$

The above estimate can be obtained by writing

$$\|u_t - \tilde{u}\|_{1,\Omega^t} \leq \|u_t - u \circ T_t^{-1}\|_{1,\Omega^t} + \|u \circ T_t^{-1} - \tilde{u}\|_{1,\Omega^t}$$

and estimating the terms separately. Note that $u \circ T_t^{-1}$ defines in fact another 'extension' of u in the sense of section 2.

We conclude this section by discussing the finite element error estimates related to above geometric estimates. In the case of the estimate in the energy norm the situation is very simple. Thanks to Poincare's inequality and Cea's lemma we obtain easily that the quasi-optimality result holds with uniform constant. Hence the developments of section 2 lead immediately to

Theorem 3.4. *Let Ω_h be a polygonal approximation of Ω. Define the discrete problem as standard piecewise linear finite element approximation of (6) for $t = h$. Then for the error between the discrete solution u_h and the prolongated solution $\tilde{u}$ it holds*

$$\|\tilde{u} - u_h\|_{H^1(\Omega_h)} \leq Ch\|f\|_{L^2(\Omega)}.$$

4 Second order systems

Let us next consider the case where the state problem is a system like the elasticity problem or the (Navier) Stokes equations. It turns out that the question of continuous dependence on geometry is more delicate in this case. To illustrate this we shall consider a model problem in elasticity with several different boundary conditions.

Let us write the problem in abstract setting as

$$a(u,\phi) = F(\phi) \qquad \forall \phi \in V. \tag{12}$$

Here

$$V = \{\phi \in (H^1(\Omega))^d \mid \phi = 0 \text{ on } \Gamma_0,\ \phi \cdot n = 0 \text{ on } \Gamma_1\}.$$

The bilinear form a is defined as

$$a(u,\phi) = \int_\Omega \sigma_{ij}(u)\epsilon_{ij}(\phi)$$

where

$$\epsilon(\phi) = \frac{1}{2}(D\phi + D^\top \phi)$$

and the stress tensor σ is defined as

$$\sigma_{ij}(u) = C_{ijkl}\epsilon_{kl}(u).$$

Above and in the sequel we use the standard convention of summation over repeated indices. The fourth order tensor of elastic coefficients, C, satisfies the standard symmetry conditions

$$C_{ijkl} = C_{jikl} = C_{klij} \tag{13}$$

and the coercivity condition

$$C_{ijkl}\psi_{ij}\psi_{kl} \geq c\psi_{ij}\psi_{ij} \tag{14}$$

for some $c > 0$ and for all symmetric second order tensors ψ. Finally, the data F is given by

$$F(\phi) = \int_\Omega f_i\phi_i + \int_{\Gamma_2} g_i\phi_i$$

for some $f \in (L^2(D))^d$, $g \in (L^2(D))^d$. The boundary of Ω is decomposed as $\partial\Omega = \bar{\Gamma}_0 \cup \bar{\Gamma}_1 \cup \bar{\Gamma}_2$. If $|\Gamma_0|_{d-1} > 0$ the problem (12) is coercive and, consequently, its solvability is obvious.

Let us now address the question of continuous dependence on geometry. As in the previous section we start by stating a shape differentiability result. First we formulate sufficient conditions for shape differentiability

- The velocity field $V \in C^2(D)$ with $V = 0$ on $\bar{\Gamma}_0 \cap \bar{\Gamma}_1$, $\bar{\Gamma}_0 \cap \bar{\Gamma}_2$, $\bar{\Gamma}_1 \cap \bar{\Gamma}_2$.
- $f \in (C^1(D))^d$, $g \in (C^1(D))^d$, $C \in (C^1(D))^{d^4}$.
- $Du \cdot V \in (H^1(\Omega))^d$.

Under these conditions the solution of (12) is shape differentiable to direction V and the shape derivative u' solves the problem

$$\begin{aligned}
\nabla \cdot \sigma &= 0, \qquad \text{in } \Omega, && (15)\\
u' &= -\langle V, n\rangle \frac{\partial u}{\partial n}, \quad \text{on } \Gamma_0, && (16)\\
u' \cdot n &= u_\tau \cdot (DV^\top \cdot n), \quad \text{on } \Gamma_1, && (17)\\
\sigma(u')_\tau &= \langle V, n\rangle f_\tau, \qquad \text{on } \Gamma_1, && (18)\\
\sigma(u') \cdot n &= \langle V, n\rangle (f + Hg) - \nabla_\Gamma \cdot (\langle V, n\rangle \sigma_\tau), \quad \text{on } \Gamma_2. && (19)
\end{aligned}$$

Here $\sigma_\tau = \sigma \cdot n - (\sigma \cdot n \cdot n)n$ and u_τ, f_τ denote the tangential components of u and f respectively. For the proof see, [5], theorem 3.11.

If $\Gamma_1 = \emptyset$ we can extend the above results for $W^{1,\infty}$ velocity fields as well.

Theorem 4.1. *Let $\Gamma_1 = \emptyset$, $V \in W^{1,\infty}(D)$, $V = 0$ on $\bar{\Gamma}_0 \cap \bar{\Gamma}_2$. Then if u and u_0 belong to $H^{5/2}$, $f \in H^1(\Omega)$ and $g \in H^1(\Omega)$ and $\tilde{u}$ is an $H^{5/2}$ extension of u to D, it holds*

$$\|u_t - \tilde{u}\|_{1,\Omega^t} \leq Ct \|V\|_{L^\infty}^{1/2} \|V\|_{W^{1,\infty}}^{1/2}.$$

If Γ_1 is present, we can not obtain shape differentiability in H^1 for $W^{1,\infty}$ velocity fields. This follows directly from the fact that

$$u' = u_\tau \cdot (DV^\top \cdot n)$$

does not belong to $H^{1/2}(\Gamma_1)$ anymore if V has only the $W^{1,\infty}$ regularity.

This result implies that the finite element approximation of this problem has to be made more carefully than just by approximating the domain by a polyhedric one.

5 A fourth order model problem

In this section we consider the continuous dependence on geometry for higher order problems. As a model example we take the linear Kirchhoff plate. Thus, let $w \in H^4(\Omega)$ be the transversal deflection of the plate $\Omega \subset R^2$, which satisfies the equation

$$(b_{ijkl} w_{,kl})_{,ij} = f \qquad \in \Omega$$

with the boundary conditions

$$\begin{aligned}
w &= 0, \quad \frac{\partial w}{\partial n} = 0 \qquad \text{on } \Gamma_0,\\
w &= 0, \quad M_n = 0 \qquad \text{on } \Gamma_1,\\
M_n &= 0, \quad Q = 0 \qquad \text{on } \Gamma_2,
\end{aligned}$$

where $\partial\Omega = \bar{\Gamma}_0 \cup \bar{\Gamma}_1 \cup \bar{\Gamma}_2$,

$$M_n = M_{ij} n_i n_j$$

with $M_{ij} = b_{ijkl} w_{,kl}$, denotes the moment and

$$Q = -M_{kl,l} n_k - \frac{\partial}{\partial \tau}(M_{kl} n_l \tau_k)$$

is the shear force. The coefficient tensor $b \in R^{2^4}$ satisfies the standard symmetry and coercivity conditions analogous to (13) and (14).

The corresponding weak form can be written as

$$a(w, \phi) = F(\phi) \qquad \forall \phi \in V, w \in V \tag{20}$$

where

$$V = \{\phi \in H^2(\Omega) \mid \phi = 0 \text{ on } \Gamma_0 \cup \Gamma_1, \frac{\partial \phi}{\partial n} = 0 \text{ on } \Gamma_0\} \tag{21}$$

and

$$\begin{aligned} a(w, \phi) &= \int_\Omega b_{ijkl} w_{,ij} \phi_{,kl}, \\ F(\phi) &= \int_\Omega f \phi. \end{aligned}$$

If $\Gamma_0 \neq \emptyset$ the solvability is guaranteed for all $f \in V'$.

Next we state without proof a shape differentiability result from [5]: If the data (Ω, b, f), the deformation velocity V and the solution w are smooth enough, then the solution w is shape differentiable to direction V and the shape derivative w' is the unique solution of the problem

$$\begin{aligned} (b_{ijkl} w'_{,kl})_{,ij} &= 0 \qquad \text{in } \Omega, \\ w' &= 0, \quad \frac{\partial w'}{\partial n} = \langle V, n \rangle \frac{\partial^2 w}{\partial n^2} \quad \text{on } \Gamma_0, \\ w' &= -\langle V, n \rangle \frac{\partial w}{\partial n} \qquad \text{on } \Gamma_1, \\ M'_n &= \nabla_\Gamma(\langle V, n \rangle M_{n\tau}) + \nabla_\Gamma(\langle V, n \rangle M_\tau) \cdot n \quad \text{on } \Gamma_1 \cup \Gamma_2, \\ Q' &= -\nabla_\Gamma \cdot (\nabla_\Gamma \cdot (\langle V, n \rangle M_\tau)_\tau) + \langle V, n \rangle f \quad \text{on } \Gamma_2, \end{aligned}$$

where

$$\begin{aligned} M'_n &= b_{ijkl} w'_{,kl} n_i n_j, \\ (M_\tau)_{ij} &= M_i j - M_n n_i n_j. \end{aligned}$$

As before we shall use the equation for the shape derivative to get error estimates for the solution under perturbation of the geometry. First we notice that conforming

approximation of fourth order problems requires the use of C^1-elements. Hence, if we use similar approximation for the geometry as well, we are lead to study deformation velocities in $W^{2,\infty}$. In this case we can show that the leading term in the difference between w_t, solution in the perturbed domain, and $\tilde{w}$, H^2-extension of w to Ω_t, is tw'. Hence, to estimate the error due to change of domain, it is sufficient to get a bound for w'.

Now, to get an estimate for w' in $H^2(\Omega)$ we have to bound w' in $H^{3/2}(\Gamma_0 \cup \Gamma_1)$, $\frac{\partial w'}{\partial n}$ in $H^{1/2}(\Gamma_0)$, M_n' in $H^{-1/2}(\Gamma_1 \cup \Gamma_2)$ and Q' in $H^{-3/2}(\Gamma_2)$. For w' we have

$$\|w'\|_{3/2,\Gamma_1} \leq \|w'\|_{1,\Gamma_1}^{1/2} \|w'\|_{2,\Gamma_1}^{1/2} \leq C\|V\|_{1,\infty}^{1/2} \|V\|_{2,\infty}^{1/2} \|\frac{\partial w}{\partial n}\|_{2,\Gamma_1}.$$

Similarly, for the other terms

$$\|\frac{\partial w'}{\partial n}\|_{1/2,\Gamma_1} \leq C\|V\|_{L^\infty}^{1/2} \|V\|_{1,\infty}^{1/2} \|\frac{\partial^2 w}{\partial n^2}\|_{1,\Gamma_1},$$

$$\|M_n'\|_{-1/2,\Gamma_1 \cup \Gamma_2} \leq C\|V\|_{L^\infty}^{1/2} \|V\|_{1,\infty}^{1/2} \|M\|_{1,\Gamma_1 \cup \Gamma_2},$$

and

$$\|Q'\|_{-3/2,\Gamma_2} \leq C\|V\|_{L^\infty}^{1/2} \|V\|_{1,\infty}^{1/2} \|M\|_{1,\Gamma_2} + C\|V\|_{L^\infty} \|f\|_{-1,\Gamma_2}.$$

Thus summarizing the above estimates we get

$$\begin{aligned} \|w'\|_{2,\Omega} \quad \leq \quad & C\|V\|_{1,\infty}^{1/2} \|V\|_{2,\infty}^{1/2} \|w\|_{7/2,\Omega} + C\|V\|_{L^\infty(\Gamma_1)}^{1/2} \|V\|_{1,\infty,\Gamma_1}^{1/2} \|w\|_{7/2,\Omega} \\ & + C\|V\|_{L^\infty} \|f\|_{-1/2,\Omega}. \end{aligned}$$

Now, if Ω_h is an approximation of Ω resulting from C^1 cubic spline approximation of the boundary $\partial\Omega$ and if Ω is a C^4-surface we have that $|\Omega - \Omega_h| = O(h^4)$ and that Ω_h can be obtained as the image of a $W^{2,\infty}$ deformation map $T_h = I + hV$, where $\|V\|_{2-k,\infty} \leq Ch^{1+k}$, $k = 0, 1, 2$. This results further to the estimate

$$\|w_h - \tilde{w}\|_{2,\Omega} \approx h\|w'\|_{2,\Omega} \leq Ch^{5/2}$$

in the general case and to

$$\|w_h - \tilde{w}\|_{2,\Omega} \leq Ch^{7/2}$$

in the case where $\Gamma_1 = \emptyset$. One observes thus that the solution is more sensitive to approximation of Γ_1 than to approximation of other boundaries. In both cases, however, the geometric error is asymptotically smaller than the $O(h^2)$ error related to finite element approximation by C^1-elements.

In the case of polyhedral approximation of Ω one observes a situation analogous to the elasticity case. For $V \in W^{1,\infty}$ the problem for w' admits a solution only if $\Gamma_1 = \emptyset$. However, unlike in the case of second order problems we can not deduce continuous dependence on geometry from this result as the proof of shape differentiability and hence the derivation of the corresponding problem requires higher derivatives of the deformation velocity. In the case where Γ_1 is present, the solution is not continuous with respect to polyhedral approximation of the geometry. This is demonstrated in the classical counter-example of [1], the so-called Babuska's plate paradox.

References

[1] Babuska I., Czechoslovak Math. J. **11**, 1961, pp. 76–105, 165-203.

[2] Ciarlet P.G., Raviart P.A., *Interpolation theory over curved elements with applications to finite element methods*, Comp. Meth. Appl. Mech. Eng., **1**, 1972, pp. 217–249.

[3] Lenoir M., *Optimal isoparametric finite elements and error estimates for domains involving curved boundaries*, SIAM J. Num. Anal., 1986, pp. 562–580.

[4] Necas J., *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[5] Sokolowski J., Zolésio J.P., *Introduction to shape optimization, shape sensitivity analysis*, Springer, Berlin, 1992.

[6] Strang G., Berger A. E., *The change in solution due to change in domain*, Proceedings of symposia in pure mathematics, XXIII, Ed. D.C. Spencer, AMS, vol. 23, pp. 199–205, 1973.

[7] Tiihonen T. *Shape calculus and FEM in smooth domains*, in preparation.

[8] Thomée V., *Polygonal domain approximation in Dirichlet's problem*, Journal of IMA, pp. 33–44, **11**, 1973.

[9] Zlamal M., *Curved elements in the finite element method, I*, SIAM J. Num. Anal., **10**, pp. 229–240, 1973.

[10] Zlamal M., *Curved elements in the finite element method, II*, SIAM J. Num. Anal., **11**, pp. 347–362, 1974.

ON CONJOINT INTERPOLATION IN TWO PATCH RECOVERY METHODS

J. AALTO and M. PERÄLÄ
University of Oulu
Department of Civil Engineering
PL 191
90101 Oulu
FINLAND

ABSTRACT

So called conjoint interpolation for expressing the final smoothed derivatives is applied in connection with two novel patch recovery methods. A numerical comparison is made between this interpolation technique and standard C^0–continuous finite element interpolation, which has been applied in connection with these patch recovery methods earlier. The new technique seems to give better results.

1. INTRODUCTION

Patch recovery methods [1]–[7] combined with so called Zienkiewicz–Zhu error estimate [8] have proved out to be efficient tools for *a posteriori* error analysis of finite element results.

Two different techniques of expressing the final smoothed derivative quantities (such as fluxes or stresses) have been used in connection with existing patch recovery methods. The first technique uses standard C^0–continuous finite element approximation, whose nodal values are obtained using patch by patch extrapolation to the nodes and averaging. This technique has been proposed and explained in some detail in reference [1]. It has also been used in references [2],[4]–[7]. The second technique uses so called conjoint interpolation and has been proposed in reference [3].

The purpose of this paper is to combine the conjoint interpolation technique of reference [3] with two novel patch recovery methods [5]–[7] developed by the authors. This paper also compares numerical results obtained using this second interpolation technique with earlier results of the first interpolation technique. Typical diffusion and plane elasticity problems are considered as numerical examples.

2. TWO PATCH RECOVERY METHODS

This chapter introduces two novel patch recovery methods. The speciality of these methods is how they use information from the original boundary value problem to improve the quality of the recovered solution. Information from the field equations is included in advance to the local polynomial representation of the basic unknown functions, which is used within the recovery patch. The unknown parameters of this polynomial are obtained by least squares fitting at the nodes of the patch. In patches, whose assembly nodes are located on the boundary of the domain, information from the boundary conditions cause additional constraint equations between the unknown parameters. These constraint equations are included into the least squares fitting procedure using Lagrange multipliers. These two patch recovery methods have many common features, but they differ in the way how the constraint equations from the field equations and boundary conditions are formed.

2.1 Local polynomial representation with "built-in" field equations

Consider a boundary value problem in two dimensions governed by n second order linear partial differential equations

$$\begin{aligned}\mathcal{R}^f(\mathbf{u}) \equiv & \mathbf{A}_{xx}\frac{\partial^2\mathbf{u}}{\partial x^2} + 2\mathbf{A}_{xy}\frac{\partial^2\mathbf{u}}{\partial x\partial y} + \mathbf{A}_{yy}\frac{\partial^2\mathbf{u}}{\partial y^2} \\ & + \mathbf{A}_x\frac{\partial\mathbf{u}}{\partial x} + \mathbf{A}_y\frac{\partial\mathbf{u}}{\partial y} + \mathbf{A}\mathbf{u} + \mathbf{f} = \mathbf{0} \quad \text{in } \Omega,\end{aligned} \tag{1}$$

where $\mathbf{u}(x,y)$ is $n \times 1$ vector of unknown functions, $\mathbf{A}_{xx}(x,y), \ldots, \mathbf{A}(x,y)$ are $n \times n$ matrices and $\mathbf{f}(x,y)$ is $n \times 1$ vector of known functions.

Let us represent the unknown $\mathbf{u}(x,y)$ locally within the recovery patch using a complete polynomial of degree p as

$$\tilde{\mathbf{u}}(\lambda,\mu) = \sum_{i=0}^{p}\sum_{j=0}^{i}\lambda^{i-j}\mu^j\mathbf{u}_{ij}, \tag{2}$$

where $\mathbf{u}_{ij}$ are $n \times 1$ vectors of unknown parameters,

$$\lambda = \frac{x - x_0}{h}, \qquad \mu = \frac{y - y_0}{h} \tag{3}$$

are dimensionless coordinates, x_0 and y_0 are coordinates of the patch assembly node and h is a charcteristic length of the patch. The dimensionless coordinates λ and μ have been adopted here in order to shorten the notation of the paper and to keep the parameters $\mathbf{u}_{ij}$ dimensionally homogeneous. A suitable value of the length h is

$$h = \frac{1}{2}[(x_{\max} - x_{\min})(y_{\max} - y_{\min})]^{\frac{1}{2}}, \tag{4}$$

where $x_{\max}$, $x_{\min}$, $y_{\max}$ and $y_{\min}$ are maximum and minimum values of the coordinates of the nodes of the patch considered. Equation (2) can also be expressed in matrix form as

$$\tilde{\mathbf{u}}(\lambda,\mu) = \mathbf{P}^p(\lambda,\mu)\mathbf{U}^p, \tag{5}$$

where

$$\mathbf{P}^p = [\mathbf{I}, \lambda\mathbf{I}, \mu\mathbf{I}, \lambda^2\mathbf{I}, \lambda\mu\mathbf{I}, \mu^2\mathbf{I}, \ldots, \lambda^p\mathbf{I}, \lambda^{p-1}\mu\mathbf{I}, \ldots, \lambda\mu^{p-1}\mathbf{I}, \mu^p\mathbf{I}] \tag{6}$$

and

$$\mathbf{U}^p = \left[\mathbf{u}_{00}^T, \mathbf{u}_{10}^T, \mathbf{u}_{11}^T, \mathbf{u}_{20}^T, \mathbf{u}_{21}^T, \mathbf{u}_{22}^T, \ldots, \mathbf{u}_{p0}^T, \mathbf{u}_{p1}^T, \ldots, \mathbf{u}_{p(p-1)}^T, \mathbf{u}_{pp}^T\right]^T. \tag{7}$$

To form constraint equations between the parameters $\mathbf{U}^p$, we demand that the right-hand-side (residual) $\mathcal{R}^f(\mathbf{u})$ of the field equation (1), with $\mathbf{u} = \tilde{\mathbf{u}}$, should approximarely vanish. This can be done using (i) power series method or (ii) weighted residual method. These methods are briefly described in the following.

(i) Power series method: We construct a complete polynomial representation

$$\tilde{\mathcal{R}}^f(\lambda,\mu) = \sum_{i=0}^{p-2}\sum_{j=0}^{i} \lambda^{i-j}\mu^j \mathcal{R}_{ij}^f \tag{8}$$

of degree $p-2$ of the residual $\mathcal{R}^f(x,y)$. Based on two dimensional Taylor expansion of the function $\mathcal{R}^f(x,y)$, the corresponding coefficients can be obtained from

$$\mathcal{R}_{ij}^f = \frac{h^i}{(i-j)!j!}\frac{\partial^i \mathcal{R}^f}{\partial x^{i-j}\partial y^j}(x_0,y_0). \tag{9}$$

With the help of equations (9), using rules of differentiation and using further equation

$$\mathbf{u}_{ij} = \frac{h^i}{(i-j)!j!}\frac{\partial^i \mathbf{u}}{\partial x^{i-j}\partial y^j}(x_0,y_0), \tag{10}$$

which holds for the unknown parameters $\mathbf{u}_{ij}$, the coefficients $\mathcal{R}_{ij}^f$ of the residual can be expressed as (linear) functions $\mathcal{R}_{ij}^f(\mathbf{U}^p)$ of the unknown parameters $\mathbf{U}^p$. (More details of this procedure is given in references [5] and [6]). Demanding that the polynomial representation (8) should vanish with all values of x and y (or λ and μ) results to linear equations $\mathcal{R}_{ij}^f(\mathbf{U}^p) = \mathbf{0}$ $(j = 0,\ldots,i,\ i = 0,\ldots,p-2)$, which can be expressed in matrix form as

$$\mathbf{C}^f\mathbf{U}^p + \mathbf{d}^f = \mathbf{0}. \tag{11}$$

(ii) Weighted residual method: We write weighted residual equations of form

$$\int_{\Omega^*} (\mathbf{P}^{p-2})^T \mathcal{R}^f(\tilde{\mathbf{u}})d\Omega = \mathbf{0}, \tag{12}$$

where the integration domain Ω^* is simply $x_0-h \le x \le x_0+h$ and $y_0-h \le y \le y_0+h$ and it is demonstrated in Fig. 1. The number of weighted residual equations (12)

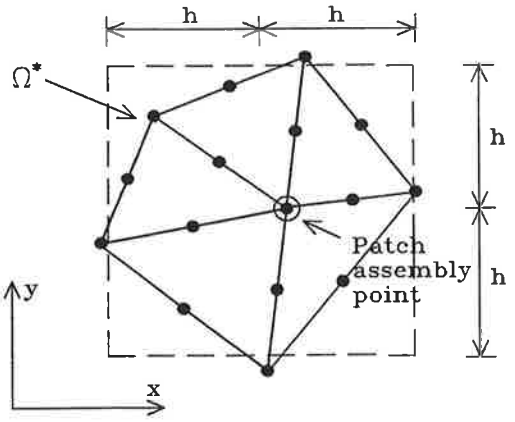


Fig. 1: Domain Ω^* of integration.

has been chosen to be $n(p-1)p/2$, which is exactly equal to the number of similar equations in the power series method. Equations (12) can be easily reduced to

$$\mathbf{C}^f \mathbf{U}^p + \mathbf{d}^f = \mathbf{0}. \tag{13}$$

(More details of this procedure is given in reference [7]).

Equations (11) or (13) are $n(p-1)p/2$ linear constraint equations between the $n(p+1)(p+2)/2$ unknown parameters $\mathbf{U}^p$. They demand that the representation (2) or (5) satisfies the field equation (1) in an average sense. It is thus possible to choose $n(p+1)(p+2)/2 - n(p-1)p/2 = n(2p+1)$ of these parameters as independent parameters

$$\mathbf{a}^p = [\mathbf{a}_1^T, \ldots, \mathbf{a}_{2p+1}^T]^T. \tag{14}$$

It is reasonable to do the choice of these independent ($n \times 1$ vectors of) parameters $\mathbf{a}_i$ so that first $\mathbf{a}_1 = \mathbf{u}_{00}$ and then corresponding to each degree i (≥ 1) independent parameters $\mathbf{a}_{2i}$ and $\mathbf{a}_{2i+1}$ are equal to two of the original parameters $\mathbf{u}_{i0}, \ldots, \mathbf{u}_{ii}$. By substituting the independent parameters $\mathbf{a}^p$ into the constraint equations (11) or (13) we get

$$\mathbf{C}_I \boldsymbol{\mathcal{U}}^p + \mathbf{C}_{II} \mathbf{a}^p + \mathbf{d}^f = \mathbf{0}, \tag{15}$$

where $\boldsymbol{\mathcal{U}}^p$ is vector of the remaining dependent parameters $\mathbf{u}_{ij}$, $\mathbf{C}_I$ and $\mathbf{C}_{II}$ are matrices containing those columns of matrix $\mathbf{C}^f$, which correspond to $\boldsymbol{\mathcal{U}}^p$ and $\mathbf{a}^p$, respectively. Solving equation (15) for $\boldsymbol{\mathcal{U}}^p$ results to

$$\boldsymbol{\mathcal{U}}^p = \boldsymbol{\mathcal{S}} \mathbf{a}^p + \boldsymbol{\mathcal{T}}, \tag{16}$$

where

$$\boldsymbol{\mathcal{S}} = -\mathbf{C}_I^{-1} \mathbf{C}_{II}, \quad \boldsymbol{\mathcal{T}} = -\mathbf{C}_I^{-1} \mathbf{d}^f. \tag{17}$$

The chosen relations between the dependent and independent parameters together with equations (16) can be written as

$$\mathbf{U}^p = \mathbf{S}\mathbf{a}^p + \mathbf{T}. \tag{18}$$

These equations express the original unknown parameters $\mathbf{U}^p$ in terms of the new independent parameters $\mathbf{a}^p$.

Substituting the result (18) into the original polynomial representation (5) finally gives

$$\tilde{\mathbf{u}}(\lambda,\mu) = \mathbf{N}^p(\lambda,\mu)\mathbf{a}^p + \mathbf{u}_0^p(\lambda,\mu), \tag{19}$$

where

$$\mathbf{N}^p(\lambda,\mu) = \mathbf{P}^p(\lambda,\mu)\mathbf{S}, \quad \mathbf{u}_0^p(\lambda,\mu) = \mathbf{P}^p(\lambda,\mu)\mathbf{T}. \tag{20}$$

Equation (19) is local polynomial representation of the unknown function $\mathbf{u}(x,y)$, which contains "built-in" approximate solution of the field equations (1). The degree of this representation is p and the number of unknown parameters a is $n(2p+1)$. Corresponding polynomial representation for the derivative quantities $\boldsymbol{\gamma}(x,y) \equiv \nabla\mathbf{u}(x,y)$ is obtained straightforwardly and is

$$\tilde{\boldsymbol{\gamma}}(\lambda,\mu) = \mathbf{B}^p(\lambda,\mu)\mathbf{a}^p + \boldsymbol{\gamma}_0^p(\lambda,\mu), \tag{21}$$

where

$$\mathbf{B}^p(\lambda,\mu) = \nabla\mathbf{N}^p(\lambda,\mu), \quad \boldsymbol{\gamma}_0^p(\lambda,\mu) = \nabla\mathbf{u}_0^p(\lambda,\mu). \tag{22}$$

The degree of this representation is $p-1$.

2.2 Constraint equations based on boundary conditions

The n boundary conditions of our second order boundary value problem are expressed as

$$\boldsymbol{\mathcal{R}}^b(\mathbf{u}) \equiv \mathbf{B}_x\frac{\partial\mathbf{u}}{\partial x} + \mathbf{B}_y\frac{\partial\mathbf{u}}{\partial y} + \mathbf{B}\mathbf{u} + \mathbf{g} = \mathbf{0} \quad \text{on } \Gamma, \tag{23}$$

where $\mathbf{B}_x(s),\ldots,\mathbf{B}(s)$ are $n\times n$ matrices and $\mathbf{g}(s)$ is $n\times 1$ vector of known functions.

We want to form constraint equations between the parameters $\mathbf{U}^p$ by demanding that the right-hand-side (residual) $\boldsymbol{\mathcal{R}}^b(\mathbf{u})$ of the boundary conditions (23), with $\mathbf{u} = \tilde{\mathbf{u}}$, should approximately vanish. This is done by using either (i) power series method or (ii) weighted residual method.

(i) Power series method: We construct polynomial representation

$$\tilde{\boldsymbol{\mathcal{R}}}^b(\sigma) = \sum_{i=0}^{p}\sigma^i\boldsymbol{\mathcal{R}}_i^b, \tag{24}$$

of degree p in the boundary coordinate s for the residual of the boundary conditions. In equation (24)

$$\sigma = \frac{s-s_0}{h} \tag{25}$$

is corresponding dimensionless coordinate and s_0 is the coordinate s of the patch assembly point. Based on one dimensional Taylor expansion of function $\mathcal{R}^b(s)$, the corresponding coefficients can be obtained from

$$\mathcal{R}_i^b = \frac{h^i}{i!}\frac{d^i\mathcal{R}^b}{ds^i}(s_0). \tag{26}$$

With the help of equation (26), using the chain rule and other rules of differentiation and using further equation (10), the coefficients of the residual can be expressed as (linear) functions $\mathcal{R}_i^b(\mathbf{U}^p)$ of the unknown parameters $\mathbf{U}^p$. (More details of this procedure is given references [5] and [6].) Demanding that the polynomial representation (24) should vanish with all values of s (or σ) results to $n(p+1)$ equations $\mathcal{R}_i^b(\mathbf{U}^p) = \mathbf{0}$, $i = 0, \ldots, p$, which can be written in matrix form as

$$\mathbf{C}^b\mathbf{U}^p + \mathbf{d}^b = \mathbf{0}. \tag{27}$$

(ii) Weighted residual method: We write weighted residual equations of form

$$\int_{\Gamma^*} (\mathbf{Q}^p)^T \mathcal{R}^b(\tilde{\mathbf{u}}) d\Gamma = \mathbf{0}, \tag{28}$$

where

$$\mathbf{Q}^p = [\mathbf{I}, \sigma\mathbf{I}, \sigma^2\mathbf{I}, \ldots, \sigma^p\mathbf{I}]. \tag{29}$$

The integration domain Γ^* on the boundary curve is simply $s_0 - h \leq s \leq s_0 + h$ and it is demonstrated in Fig. 2. The number of weighted residual equations (29) has been chosen to be $n(p+1)$. This is exactly equal to the number of similar equations in the power series method. Equations (28) can be easily reduced to

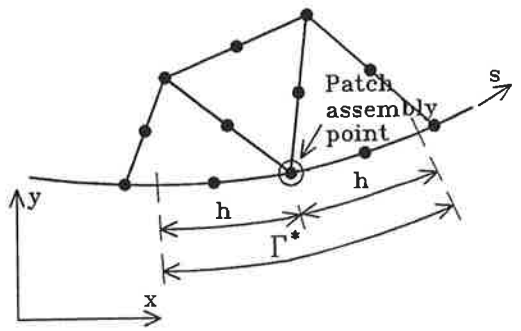


Fig. 2: Domain Γ^* of integration.

$$\mathbf{C}^b\mathbf{U}^p + \mathbf{d}^b = \mathbf{0}. \tag{30}$$

(More details of this procedure is given in reference [7].)

Equations (27) or (30) are $n(p+1)$ linear constraint equations between the $n(p+1)(p+2)/2$ unknown parameters $\mathbf{U}^p$, which demand that the representation (2) or (5) should satisfy the boundary conditions (23) in an average sense. These equations are additional constraint equations (in addition to equations (11) or (13)) between the original parameters $\mathbf{U}^p$, which should hold, if the patch assembly node is on the boundary of the domain. The corresponding equations between the new parameters $\mathbf{a}^p$ are obtained with the help of equation (18) and they can be written as

$$\bar{\mathbf{C}}\mathbf{a}^p + \bar{\mathbf{d}} = \mathbf{0}, \tag{31}$$

where

$$\bar{\mathbf{C}} = \mathbf{C}^b\mathbf{S}, \quad \bar{\mathbf{d}} = \mathbf{d}^b + \mathbf{C}^b\mathbf{T}. \tag{32}$$

2.3 Least squares fitting in a recovery patch

The unknown parameters $\mathbf{a}^p$ (of the patch under consideration) are obtained by least squares fitting of the polynomial representation $\tilde{\mathbf{u}}(x,y)$ to the corresponding nodal values $\mathbf{u}_i$ of the original finite element solution at the nodes of the patch. The corresponding least squares function is thus

$$\Pi = \sum_i [\tilde{\mathbf{u}}(x_i,y_i) - \mathbf{u}_i]^T [\tilde{\mathbf{u}}(x_i,y_i) - \mathbf{u}_i]. \tag{33}$$

If the patch assembly node is on the boundary of the domain, this least squares function is modified to

$$\Pi = \sum_i [\tilde{\mathbf{u}}(x_i,y_i) - \mathbf{u}_i]^T [\tilde{\mathbf{u}}(x_i,y_i) - \mathbf{u}_i] + (\boldsymbol{\lambda})^T(\bar{\mathbf{C}}\mathbf{a}^p + \bar{\mathbf{d}}), \tag{34}$$

where $\boldsymbol{\lambda}$ is vector of corresponding Lagrange multipliers.

3. FINAL INTERPOLATION OF THE DERIVATIVES

Our purpose is to find an interpolation formula for getting representative values of the derivative quantities $\boldsymbol{\gamma}$ at certain point (ξ,η) within an element of the finite element grid. Typically such points are the integration points. There are two possibilities to perform this final interpolation after the parameters $\mathbf{a}^p$ corresponding to each recovery patch have been solved.

3.1 C^0–continuous finite element interpolation

In the first procedure the derivative quantities $\boldsymbol{\gamma}$ are first extrapolated patch by patch to appropriate system nodes using the representation (21). Unique system nodal values $\boldsymbol{\gamma}_i$ are then calculated by averaging these values at the system nodes. Finally local finite element approximation

$$\bar{\boldsymbol{\gamma}}(\xi,\eta) = \sum_{i=1}^{m} N_i(\xi,\eta)\boldsymbol{\gamma}_i^e \tag{35}$$

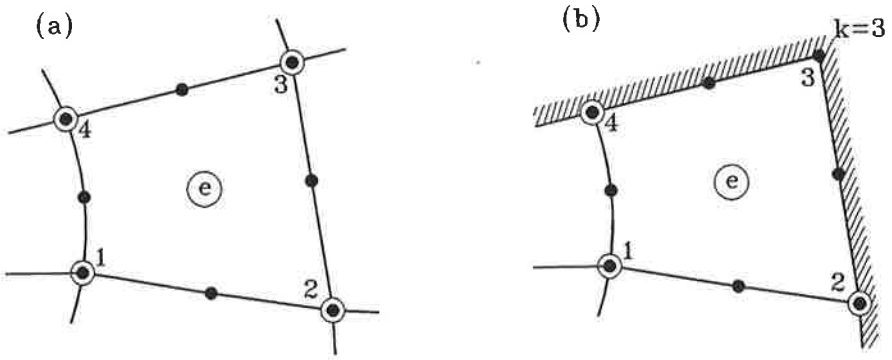


Fig. 3: A quadrilateral element with (a) 4 and (b) 3 patch assembly nodes.

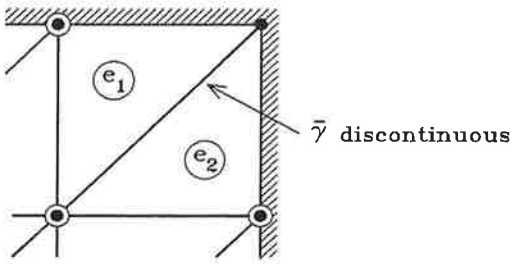


Fig. 4: Two triangular elements in a corner of the domain.

is used, where $N_i(\xi,\eta)$ are the element shape functions, which were used to approximate the basic unknown function $\mathbf{u}(x,y)$ in the original finite element analysis, γ_i^e are element nodal values corresponding to the system nodal values γ_i and m is the number of element nodes. (More details of this procedure are given in reference [1].)

3.2 Conjoint interpolation

The second procedure is explained here in more detail. To do this the polynomial representation (21) of the derivatives $\gamma(x,y)$ corresponding to recovery patch i is here expressed as

$$\tilde{\gamma}^i(\lambda^i,\mu^i) = \mathbf{B}^i(\lambda^i,\mu^i)\mathbf{a}^i + \gamma_0^i(\lambda^i,\mu^i), \tag{36}$$

where

$$\lambda^i = \frac{x - x_0^i}{h^i}, \quad \mu^i = \frac{y - y_0^i}{h^i}. \tag{37}$$

The superscript p referring to polynomial degree has been dropped out and the superscript i referring to the patch under consideration has been added.

Consider an isoparametric triangular or quadrilateral finite element, whose each corner node i is a patch assembly node (see Fig. 3a). The geometrical mapping of the element is

$$x = \sum_{j=1}^{m} N_j(\xi,\eta)x_j, \quad y = \sum_{j=1}^{m} N_j(\xi,\eta)y_j, \tag{38}$$

where x_j and y_j are the coordinates of the element nodes. The interpolation formula we are looking for should weight the the derivatives $\tilde{\gamma}^i$ of each patch corresponding to the corner nodes i of the element suitably. A rather natural choise could be to use the shape functions of the corner nodes as weighting functions. The interpolation formula is thus

$$\bar{\gamma}(\xi,\eta) = \sum_{i=1}^{m^*} N_i^*(\xi,\eta)\tilde{\gamma}^i(\lambda^i,\mu^i), \tag{39}$$

where $N_i^*(\xi,\eta)$ are the shape functions of linear triangle or bilinear quadrilateral, respectively, and m^* is the number of corner nodes of the element (3 for triangles and 4 for quadrilaterals). Equations (36)–(39) can be used to perform conjoint interpolation of reference [3].

The calculation process could be as follows: (i) Physical coordinates x and y corresponding to the natural coordinates ξ and η are evaluated using equations (38). (ii) Patch coordinates λ^i and μ^i corresponding x and y are evaluated using equations (37). (iii) The patchwise derivatives $\tilde{\gamma}^i$ are obtained using equation (36). (iv) The final interpolated derivative $\bar{\gamma}$ is obtained using equation (39).

A patch recovery node at the corner of the boundary of the domain is rather complicated to handle even though a singularity of the solution does not exist (see [6]). Therefore a good engineering solution could be to leave the corners of the domain without patch recovery nodes and we are going to do so in this paper. Consequently we have to be able to interpolate within an element, whose one corner node k is not a patch assembly node (see Fig. 3b). One possibility is to use instead of equation (39) the formula

$$\bar{\gamma}(\xi,\eta) = \sum_{\substack{i=1 \\ i\neq k}}^{m^*} [N_i^*(\xi,\eta) + \frac{1}{m^*-1}N_k^*(\xi,\eta)]\tilde{\gamma}^i(\lambda^i,\mu^i). \tag{40}$$

At the corner node k this interpolation gives

$$\bar{\gamma}(\xi_k,\eta_k) = \frac{1}{m^*-1}\sum_{\substack{i=1 \\ i\neq k}}^{m^*} \tilde{\gamma}^i(\lambda_k^i,\mu_k^i), \tag{41}$$

which is the mean value of the corresponding patch approximations. This seems at first glance, to be quite acceptable. There is, however, a small limitation: If two elements are connected to a corner node, which is not a patch assembly node (see Fig. 4), the approximation $\bar{\gamma}$ on the boundary of these two elements is not continuous. In spite of a better idea we, however, accept formula (40).

4. NUMERICAL EXAMPLES

4.1 Example problems and typical grids

Two example problems with known, nonpolynomial, analytical solutions were used to demonstrate the presented methods:

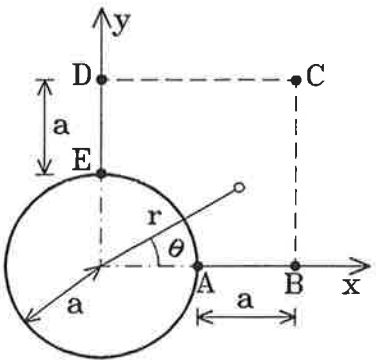


Fig. 5: Problem A.

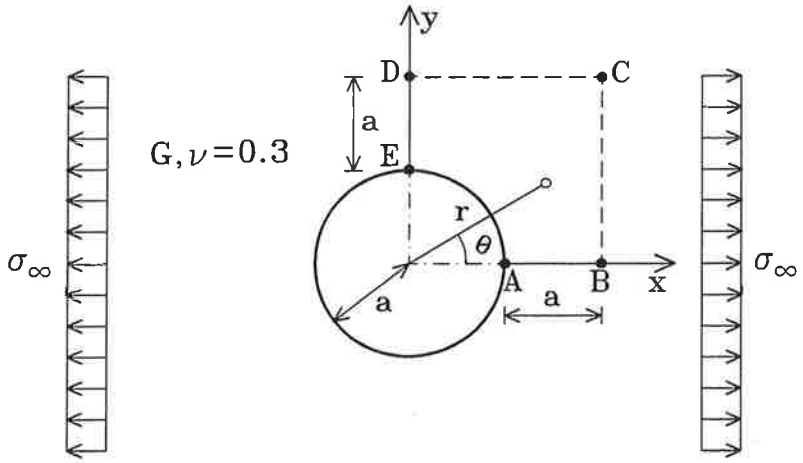


Fig. 6: Problem B.

Problem A: Potential flow around a circular cylinder. Figure 5 shows the domain and demonstrates the problem. It is governed by equation (1) with $n = 1$, $\mathbf{u} = \phi$, $\mathbf{A}_{xx} = \mathbf{A}_{yy} = 1$, $\mathbf{A}_{xy} = \mathbf{A}_x = \mathbf{A}_y = \mathbf{A} = 0$, and $\mathbf{f} = 0$, where ϕ is the unknown potential.

Problem B: Stretching of an infinite plate with a circular hole. Figure 6 shows the domain and demonstrates the problem. It is governed by equation (1) with

n=2,

$$\mathbf{u} = \begin{Bmatrix} u \\ v \end{Bmatrix}, \quad \mathbf{A}_{xx} = \frac{G}{\eta - 1} \begin{bmatrix} \eta + 1 & 0 \\ 0 & \eta - 1 \end{bmatrix}, \quad \mathbf{A}_{xy} = \frac{G}{\eta - 1} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$
$$\mathbf{A}_{yy} = \frac{G}{\eta - 1} \begin{bmatrix} \eta - 1 & 0 \\ 0 & \eta + 1 \end{bmatrix}, \quad \mathbf{f} = \begin{Bmatrix} f_x \\ f_y \end{Bmatrix} \tag{43}$$

and $\mathbf{A}_x = \mathbf{A}_y = \mathbf{A} = \mathbf{0}$, where u and v are the unknown displacements, G is shear modulus, $\eta = (3 - \nu)/(1 + \nu)$, ν is Poisson's ratio and f_x and f_y are volume forces. Here $f_x = f_y = 0$ and thus $\mathbf{f} = \mathbf{0}$.

Figure 7 shows typical finite element grids ($h/a = 0.5$) used in experimental convergence study of the two example problems.

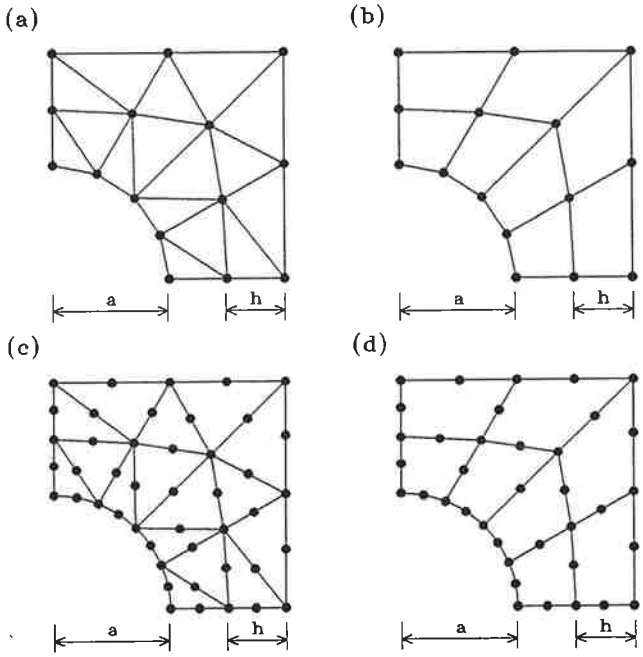


Fig. 7: Typical grids of (a) linear triangles, (b) bilinear quadrilaterals, (c) quadratic triangles and (d) quadratic Serendip quadrilaterals.

4.2 Numerical results

Figs. 8–11 present experimental convergence study of the relative error in energy (η_E) of the two example problems A and B. Four different element types: linear triangles, bilinear quadrilaterals, quadratic triangles and quadratic Serendip quadrilaterals were used in the analysis. Each of the figures show error of the smoothed solution obtained with either of the two patch recovery methods of this paper using C^0–continuous finite element interpolation (FEI) and conjoint interpolation (COI). Error of the smoothed

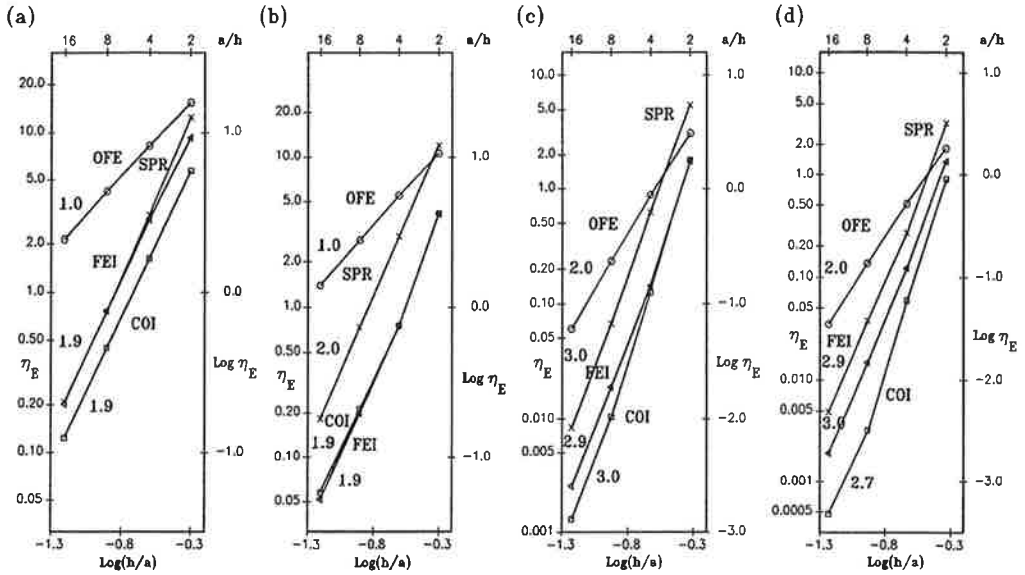


Fig. 8: Results of problem A using power series method:
(a) linear triangles ($p = 2$), (b) bilinear quadrilaterals ($p = 3$),
(c) quadratic triangles ($p = 4$) and (d) quadratic quadrilaterals ($p = 5$).

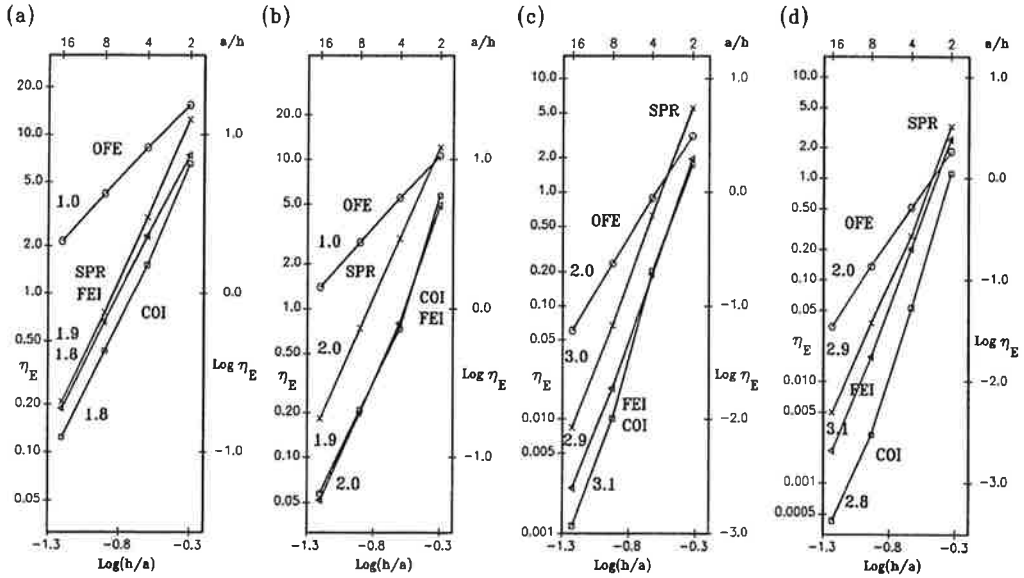


Fig. 9: Results of problem A using weighted residual method:
(a) linear triangles ($p = 2$), (b) bilinear quadrilaterals ($p = 3$),
(c) quadratic triangles ($p = 4$) and (d) quadratic quadrilaterals ($p = 5$).

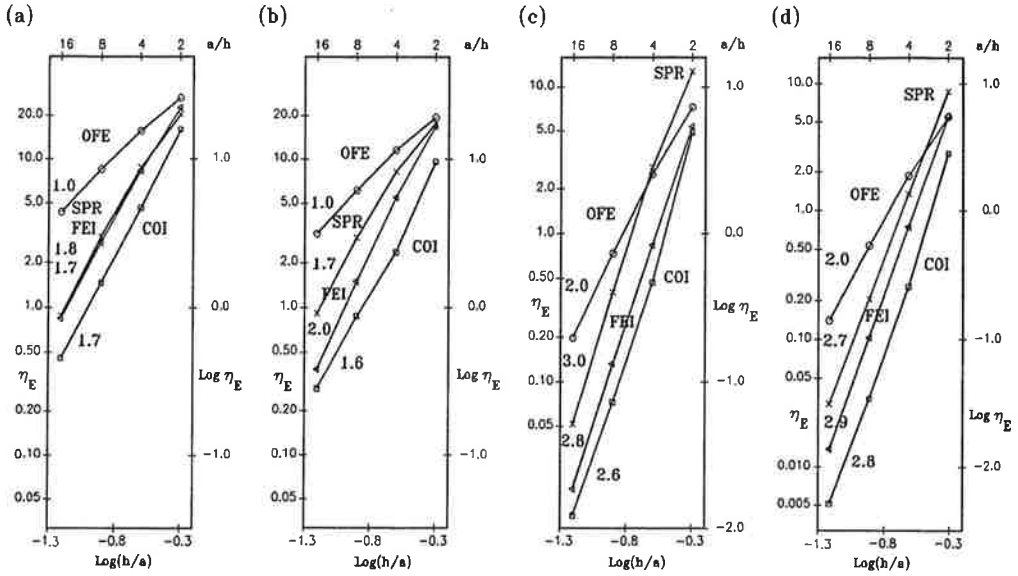


Fig. 10: Results of problem B using power series method:
(a) linear triangles ($p = 2$), (b) bilinear quadrilaterals ($p = 3$),
(c) quadratic triangles ($p = 4$) and (d) quadratic quadrilaterals ($p = 5$).

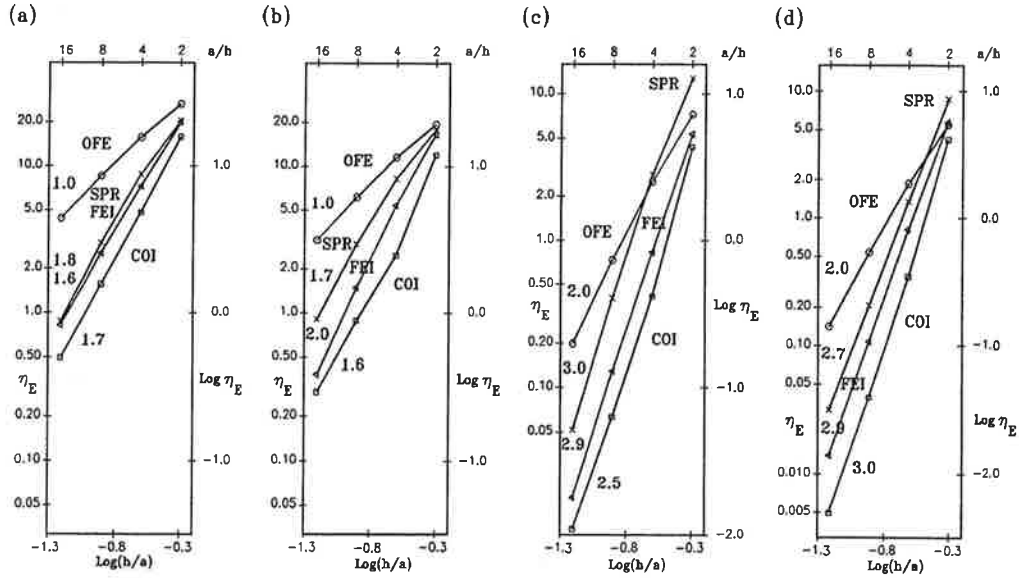


Fig. 11: Results of problem B using weighted residual method:
(a) linear triangles ($p = 2$), (b) bilinear quadrilaterals ($p = 3$),
(c) quadratic triangles ($p = 4$) and (d) quadratic quadrilaterals ($p = 5$).

Table 1: Effectivity indices θ with coarse grid ($h/a = 0.5$); problem A.

Element type	SPR Ref. [1]	Power series: FEI	Power series: COI	Weighted residuals: FEI	Weighted residuals: COI
Linear triangles	1.06	1.16	1.05	1.05	1.01
Bilinear quadrilaterals	1.26	1.11	1.10	1.10	1.18
Quadratic triangles	1.95	1.23	1.21	1.27	1.20
Quadratic quadrilaterals	1.86	1.18	1.04	1.61	1.06

Table 2: Effectivity indices θ with dense grid ($h/a = 0.0625$); problem A.

Element type	SPR Ref. [1]	Power series: FEI	Power series: COI	Weighted residuals: FEI	Weighted residuals: COI
Linear triangles	0.994	1.004	0.999	1.002	0.998
Bilinear quadrilaterals	0.991	1.002	1.000	1.002	1.000
Quadratic triangles	0.989	1.001	0.998	1.001	0.998
Quadratic quadrilaterals	1.015	1.004	1.003	1.000	1.003

solution obtained with the superconvergent patch recovery (SPR) method of reference [1] and error of the original finite element solution (OFE) are also shown for comparison. Results of problem A obtained using patch recovery based on power series method and weighted residual method are shown in Figs. 8 and 9, respectively. Results of problem B obtained using patch recovery based on power series method and weighted residual method are shown in Figs. 10 and 11, respectively. Both patch recovery methods of this paper give clearly better smoothed solution than superconvergent patch recovery method (SPR) of reference [1]. Conjoint interpolation (COI) gives better smoothed solution than finite element interpolation (FEI), but the rate of convergence in connection with bilinear and quadratic quadrilaterals slightly reduces with dense grids.

Tables 1–4 show a comparison of the effectivity indices

$$\theta = \frac{\eta_E^{esti}}{\eta_E}, \tag{44}$$

where η_E is relative error in energy and η_E^{esti} is the corresponding Zienkiewicz-Zhu

Table 3: Effectivity indices θ with coarse grid ($h/a = 0.5$); problem B.

Element type	SPR Ref. [1]	Power series: FEI	Power series: COI	Weighted residuals: FEI	Weighted residuals: COI
Linear triangles	0.93	1.41	1.14	1.28	1.07
Bilinear quadrilaterals	0.96	1.36	1.07	1.32	1.11
Quadratic triangles	1.93	1.33	1.26	1.33	1.23
Quadratic quadrilaterals	1.50	1.38	1.18	1.42	1.40

Table 4: Effectivity indices θ with dense grid ($h/a = 0.0625$); problem B.

Element type	SPR Ref. [1]	Power series: FEI	Power series: COI	Weighted residuals: FEI	Weighted residuals: COI
Linear triangles	0.971	1.017	1.000	1.011	0.996
Bilinear quadrilaterals	0.977	1.006	0.999	1.005	0.999
Quadratic triangles	0.952	0.999	0.986	0.999	0.987
Quadratic quadrilaterals	1.010	1.007	1.003	1.006	1.000

[8] estimate, obtained for the two example problems A and B. Tables 1 and 2 show the results of problem A using coarse ($h/a = 0.5$) and dense ($h/a = 0.0625$) grids, respectively. Tables 3 and 4 show the results of problem B using coarse ($h/a = 0.5$) and dense ($h/a = 0.0625$) grids, respectively. Both patch recovery methods of this paper give clearly better error estimates than superconvergent patch recovery method (SPR) of reference [1]. Conjoint interpolation (COI) gives better error estimates than finite element interpolation (FEI) especially with coarse grid.

5. CONCLUSIONS

The purpose of this paper was to introduce two novel patch recovery methods [4]-[7] and to combine them to a new final interpolation technique of the derivative quantities, so called conjoint interpolation [3].

The obtained numerical results show, that this new interpolation technique works well and improves the results of these two patch recovery methods compared to standard C^0–continuous finite element interpolation, which has been applied in connection

with these patch recovery methods earlier, especially with coarse grids.

REFERENCES

1. O.C. Zienkiewicz and J.Z. Zhu, The superconvergent patch recovery and a posteriori error estimates. Part 1: The recovery technique, *Int. J. Num. Meth. Engng.* **33** (1992), 1331–1364.

2. N.-E. Wiberg and F. Abdulwahab, An efficient postprocessing technique for stress problems based on superconvergent derivatives and equilibrium, in *Numerical methods in engineering '92*, Elsevier, Amsterdam, 1992, 25–32.

3. T. Blacker and T. Belytchko, Superconvergent patch recovery with equilibrium and conjoint interpolant enhancement, *Int. J. Num. Meth. Engng.* **37** (1994), 1517–1536.

4. J. Aalto, Built-in field equations for recovery procedures, *Proceedings of the Second International Conference on Computational Structural Technology (CST 94): Advances in post and preprocessing for finite element technology*, (Edited by M. Papadrakakis and B.H.V. Topping), Civil-Comp Press, pp. 125–135; also, *Computers and Structures* (accepted 16th November 1995).

5. J. Aalto and M. Perälä, Two robust patch recovery methods with built-in field equations and boundary conditions, *Finite Element Methods: Superconvergence, Post-Processing and A Posteriori Estimates*, University of Jyväskylä, Jyväskylä, Finland, July 1-4, 1996, 18p. (In press).

6. J. Aalto and M. Åman, Polynomial representations with built-in field equations for patch recovery procedures, *Proceedings of the Third International Conference on Computational Structural Technology (CST 96): Advances in finite element technology*, (Edited by B.H.V. Topping), Civil-Comp Press, pp. 109-125.

7. J. Aalto and M. Perälä, Built-in field equations for patch recovery procedures using weighted residuals, *Proceedings of the Third International Conference on Computational Structural Technology (CST 96): Advances in finite element technology*, (Edited by B.H.V. Topping), Civil-Comp Press, pp. 135-150.

8. O.C. Zienkiewicz and J.Z. Zhu, A simple error estimator and adaptive procedure for practical engineering analysis, *Int. J. Num. Meth. Engng.* **24** (1987), 337–357.

AN ALTERNATIVE FE-SOLUTION STRATEGY FOR ELASTOSTATIC PROBLEMS

J. TOIVOLA and J. MÄKINEN
Applied Mechanics
Tampere University of Technology
P.O. Box 589, FIN-33101 Tampere

ABSTRACT

Conventional finite element solution method for 3D elastostatic problems starts from Navier's equation and uses Galerkin method for building of discretized equilibrium equations. This procedure leads to same computational code as minimization of total potential energy or use of principle of virtual work. In this paper, the problem is reformulated using Papkovich-Neuber-solution, in which the Navier's equation is substituted by Poisson equations. By considering two computationally most demanding phases of numerical solution, namely numerical integration of elemental stiffness matrices and triangulation of global stiffness matrix, it is shown that the proposed procedure is potentionally 5 to 25 times faster than conventional method. It is also shown that the amount of memory needed to store the global stiffness matrix is reduced by almost 90%. Because the standard Galerkin method leads to problems at boundary, solution procedure is formulated using least squares method.

INTRODUCTION

Computational speed is a desired property of numerical solution of partial differential equations. In elastostatics, the tendency of creating FE-model from geometrical model increases the size of FE-model considerably compared to situation, where FE-model is created from scratch. Also the use of shape optimization procedures leads to excessive number of analyses, and thus the analysis speed tends to be a limiting factor in the practical use of these methods.

When using conventional FE-procedures to solve governing equation of displacements, the matrix product

$$[N]^T[E][N] \tag{1}$$

needs to be evaluated at each integration point of an element, where matrices $[N]$ and $[E]$ in 3D-situation are of the form

$$[N]=\begin{bmatrix}[B_1] & [B_2] & \cdots & [B_{n_e}]\end{bmatrix}$$

$$[B_i]=\begin{bmatrix} N_{i,x} & 0 & 0 \\ 0 & N_{i,y} & 0 \\ 0 & 0 & N_{i,z} \\ N_{i,y} & N_{i,x} & 0 \\ 0 & N_{i,z} & N_{i,y} \\ N_{i,z} & 0 & N_{i,x} \end{bmatrix} \quad [E]=\overline{E}\begin{bmatrix} 1-\nu & \nu & \nu & 0 & 0 & 0 \\ \nu & 1-\nu & \nu & 0 & 0 & 0 \\ \nu & \nu & 1-\nu & 0 & 0 & 0 \\ 0 & 0 & 0 & \overline{\nu} & 0 & 0 \\ 0 & 0 & 0 & 0 & \overline{\nu} & 0 \\ 0 & 0 & 0 & 0 & 0 & \overline{\nu} \end{bmatrix}$$

$$\overline{E}=\frac{E}{(1+\nu)(1-2\nu)}, \qquad \overline{\nu}=\tfrac{1}{2}(1-2\nu) \tag{2}$$

Taking the special structure of these matrices in to account, it can be shown that matrix product (1) needs approximately $13n_e^2$ flops, where n_e is the number of nodes in element, and 1 flop is either addition/subtraction or multiplication/division [22].

When solving one Poisson equation using conventional FEM, the matrix product

$$[N]^T[N] \tag{3}$$

needs to be evaluted at each integration point, where matrix $[N]$ has a form

$$[N]=\begin{bmatrix} N_{1,x} & N_{2,x} & \cdots & N_{n_e,x} \\ N_{1,y} & N_{2,y} & \cdots & N_{n_e,y} \\ N_{1,z} & N_{2,z} & \cdots & N_{n_e,z} \end{bmatrix} \tag{4}$$

This matrix product needs approximately $\frac{5}{2}n_e^2$ flops, and is thus over five times faster to evaluate than corresponding matrix product in elastostatic problems.

The triangulation of global stiffness matrix in elastostatics needs approximately $N\cdot B^2$ flops [22], where N is the number of global DOFs and B is the bandwidth. If the same element mesh is used to solve one Poisson equation, both the bandwidth and number of global DOFs drops to one third, and hence in this situation the triangulation needs $\frac{1}{27}N\cdot B^2$ flops. Hence, triangulation of global stiffness matrix for one Poisson equation with the same element mesh as in elastostatic problem is approximately 27 times faster.

Using the same notation as in previous paragraph, the amount of memory needed to store the global stiffness matrix in elastostatic problem is proportional to $N\cdot B$. To solve one Poisson equation, this is reduced to $\frac{1}{9}N\cdot B$. Thus, the amount of memory needed drops almost 90 %.

It is clear that governing equation of elastostatic problem, namely Navier equation, cannot be substituted by one Poisson equation, and hence the above conclusions needs to reconsidered. Some aspects of this matter is slightly addressed later when the problem is stated more clearly. The results above are very approximative, the evaluation of Jacobian

of geometric mapping and evaluation of derivatives of shape functions are not considered. Also, different formulations should be compared based on some well defined error-norm, not based on same element mesh. Finally, only numerical experiments show, in what proportions the numerical integration of elemental stiffness matrix and triangulation of global stiffness matrix appears in total computation time.

The displacement field $\mathbf{u}$ of linear, materially isotropic and homogeneous, elastostatic problem is governed by Navier's equation [21, 23]

$$(\lambda+\mu)\nabla(\nabla\cdot\mathbf{u})+\mu\nabla^2\mathbf{u}+\mathbf{f}=\mathbf{0} \tag{5}$$

where λ and μ are Lame´s contants and $\mathbf{f}$ is the body force density. The solution to Navier equation is known to exist and is unique when global rigid body displacement is not possible [21]. In this paper, the region occupied by elastic material is denoted as D and boundary of D is denoted as Γ.

Using the definition of vector Laplacian

$$\nabla^2\mathbf{u}=\nabla(\nabla\cdot\mathbf{u})-\nabla\times(\nabla\times\mathbf{u}) \tag{6}$$

it is possible to write Navier's equation in the forms

$$(\lambda+2\mu)\nabla(\nabla\cdot\mathbf{u})-\mu\nabla\times(\nabla\times\mathbf{u})+\mathbf{f}=\mathbf{0} \tag{7a}$$

$$(\lambda+2\mu)\nabla^2\mathbf{u}+(\lambda+\mu)\nabla\times(\nabla\times\mathbf{u})+\mathbf{f}=\mathbf{0} \tag{7b}$$

From equations (5) and (7b) it is seen, that Navier's equation is almost vector Poisson equation, the only difference being either the term $\nabla(\nabla\cdot\mathbf{u})$ or $\nabla\times(\nabla\times\mathbf{u})$.

There are several possibilities to substitute Navier equation (5) by several Poisson equations (or vector Poisson equation). Among these, the displacement field $\mathbf{u}$ can be decomposed to irrotational and solenoidal components [24, 25], which are governed by vector Poisson equations, the displacement field can be represented by scalar and vector potentials [24, 25], which in turn are governed by Poisson and vector Poisson equations, or new representations of the displacement field can be defined in a suitable manner to reach vector Poisson equation for unknown vector field, as in Betti's method [20]. All of these methods seems to be inconvinient considering numerical solution. For this reason, the Papkovich-Neuber-formulation is considered in this paper.

GALERKIN VECTOR AND PAPKOVICH-NEUBER FORMULATION

It is natural to examine representations in which the displacement field is built up from second derivatives of a potential function. The most general such a form is found to be

$$\mathbf{u}=\left[c\nabla^2-\nabla(\nabla\cdot)\right]\mathbf{F} \tag{8}$$

where c is an arbitrary constant which will be assigned so as to simplify the representation. Substituting representation (8) to Navier equation (5) and realizing that

$$\nabla\left[\nabla\cdot\left(\nabla^2\mathbf{F}\right)\right]=\nabla\{\nabla\cdot[\nabla(\nabla\cdot\mathbf{F})]\}=\nabla^2[\nabla(\nabla\cdot\mathbf{F})] \tag{9}$$

and setting

$$c=\frac{\lambda+2\mu}{\lambda+\mu}=2(1-\nu) \tag{10}$$

yields

$$2(1-\nu)\mu\nabla^4\mathbf{F}=-\mathbf{f} \tag{11}$$

The vector $\mathbf{F}$ is known as Galerkin vector. Its use in numerical solution is inconvenient because of biharmonic operator ∇^4 and because it is not unique [21, 23].

Papkovich [1, 2] and Neuber [3] independly defined an auxliary vector

$$\psi=-\tfrac{1}{2}\nabla^2\mathbf{F} \tag{12}$$

which is governed by

$$\nabla^2\psi=\frac{1}{4(1-\nu)\mu}\mathbf{f} \tag{13}$$

as can be seen from equations (11) and (12). When it is noted that

$$\nabla^2(\mathbf{r}\cdot\psi)=2\nabla\cdot\psi+\frac{1}{4(1-\nu)\mu}\mathbf{r}\cdot\mathbf{f} \tag{14}$$

and from definition (12)

$$\nabla\cdot\psi=-\tfrac{1}{2}\nabla^2(\nabla\cdot\mathbf{F}) \tag{15}$$

it is clear that

$$\nabla^2(\mathbf{r}\cdot\psi+\nabla\cdot\mathbf{F})=\frac{1}{4(1-\nu)\mu}\mathbf{r}\cdot\mathbf{f} \tag{16}$$

Now, an auxliary function θ is defined by

$$\nabla^2\theta=-\frac{1}{4(1-\nu)\mu}\mathbf{r}\cdot\mathbf{f} \tag{17}$$

so that finally

$$\nabla^2(\mathbf{r}\cdot\boldsymbol{\psi}+\nabla\cdot\mathbf{F}+\theta)=0 \tag{18}$$

This shows that

$$-\nabla\cdot\mathbf{F}=\mathbf{r}\cdot\boldsymbol{\psi}+\phi'+\theta \tag{19}$$

where scalar field ϕ' is harmonic, i.e. $\nabla^2\phi'=0$. Combining ϕ' and θ by $\phi=\phi'+\theta$ leads to Poisson equation

$$\nabla^2\phi=-\frac{1}{4(1-\nu)\mu}\mathbf{r}\cdot\mathbf{f} \tag{20}$$

for scalar field ϕ.

Substituting $\nabla^2\mathbf{F}$ from (12) and $-\nabla\cdot\mathbf{F}$ from (19) to Galerkin vector representation (8) leads to displacement formula

$$\mathbf{u}=-4(1-\nu)\boldsymbol{\psi}+\nabla(\mathbf{r}\cdot\boldsymbol{\psi}+\phi) \tag{21}$$

and governing equations for unknown functions are (13) and (20). In this paper, the unknown functions $\boldsymbol{\psi}$ and ϕ are called Papkovich-Neuber-potentials.

COMPLETENESS, EXISTENCE AND UNIQUENESS OF PAPKOVICH-NEUBER-POTENTIALS

There are four unknown functions, namely ϕ and three components of $\boldsymbol{\psi}$. Because naturally there arise only three boundary conditions at each point of the boundary, there have to be an additional equation constraining the four unknowns or possibly one additional boundary condition. Looking for fourth equation, $\boldsymbol{\psi}$ is expressed as

$$\boldsymbol{\psi}=-\tfrac{1}{2}\nabla^2\mathbf{F}=-\tfrac{1}{2}\nabla(\nabla\cdot\mathbf{F})+\tfrac{1}{2}\nabla\times(\nabla\times\mathbf{F}) \tag{22}$$

Now, $\nabla\times\big(\nabla(\nabla\cdot\mathbf{F})\big)=\mathbf{0}$ and $\nabla\cdot\big(\nabla\times(\nabla\times\mathbf{F})\big)=0$. Hence, equation (22) expresses $\boldsymbol{\psi}$ decomposed to irrotational and solenoidal parts

$$\boldsymbol{\psi}=\boldsymbol{\psi}_{irr}+\boldsymbol{\psi}_{sol} \tag{23}$$

where

$$\boldsymbol{\psi}_{irr}=-\tfrac{1}{2}\nabla(\nabla\cdot\mathbf{F}) \qquad \nabla\times\boldsymbol{\psi}_{irr}=\mathbf{0} \tag{24a}$$

$$\boldsymbol{\psi}_{sol}=\tfrac{1}{2}\nabla\times(\nabla\times\mathbf{F}) \qquad \nabla\cdot\boldsymbol{\psi}_{sol}=0 \tag{24b}$$

Taking the gradient of (19) with the aid of (24a) results

$$2\psi_{irr} = \nabla(\mathbf{r}\cdot\psi + \phi) \tag{25}$$

which is the required equation constraining the unknows. To show that this equation indeed remove the redundancy, (25) is substituted to (21)

$$\mathbf{u} = -4(1-\nu)\psi + 2\psi_{irr} \tag{26}$$

where scalar field ϕ is eliminated. Thus, at the boundary the Papkovich-Neuber vector ψ have to be decomposed to irrotational and solenoidal components in order to represent the boundary conditions, which is inconvinient in numerical solution. This can be avoided, if scalar field ϕ is substituted by scalar field

$$\alpha = \mathbf{r}\cdot\psi + \phi \tag{27}$$

Now, from (25) with the aid of (24b)

$$\nabla^2\alpha = 2\nabla\cdot\psi \tag{28}$$

Equation (28) couples the unknown fields α and ψ and thus bandwidth of global stiffness matrix is increased. Formulation with scalar field (27) is due to Freiberger [21].

The completeness of Papkovich-Neuber-formulations (21) and (26) has been proved by Naghdi and Hsu [7] for regular, bounded and multiply connected region of space. For other proofs of completeness and related matters, intrested reader is referred to [4-6, 8-14]. It is clear that the excistence of Papkovich-Neuber-potentials follows from the completeness of representations (21) and (26).

Historically, because Papkovich in his formulation includes ϕ in [1] and excludes it in [2] and Neuber [3] claims that ϕ or any component of ψ could be set to zero without violating the completeness, many attemps has been made to show that this is possible. The results are as follows: ϕ can be set to zero

a) if region D occupied by elastic material is star-shaped with respect to one of it's points and 4ν is not an integer [(5), 6, 10, 12, 13, 14]. If 4ν is an integer, ϕ can be limited to the set of solid harmonics of degree $4-4\nu$ [10, 13].

b) if region D is bounded and r-convex [14].

c) if region D is unbounded and a radial line cuts boundary transversely in at most one point [14].

d) if region D is bounded, periphractic domain between two surfaces, which are star-shaped with respect to same point [14].

Any rectangular component of ψ can be set to zero if region occupied by elastic material is convex [6]. Additionally, if region is convex in e.g. z-direction, ψ_z could be set to zero [6, 14]. Although one element is always star-shaped with respect to one of it's points or usually convex, these results cannot be used for general region and hence in numerical formulation of the problem.

Considering the numerical solution, the most tempting formulation using Papkovich-Neuber-potentials is

$$\mathbf{u} = -4(1-\nu)\boldsymbol{\psi} + \nabla(\mathbf{r}\cdot\boldsymbol{\psi} + \phi) \tag{29}$$

where

$$\nabla^2\boldsymbol{\psi} = \frac{1}{4(1-\nu)\mu}\mathbf{f} \tag{30a}$$

$$\nabla^2\phi = -\frac{1}{4(1-\nu)\mu}\mathbf{r}\cdot\mathbf{f} \tag{30b}$$

Now, an attempt is made to add one boundary condition to the problem so that representation (29) is also unique. First, the null displacement field $\mathbf{u}_0 \equiv \mathbf{0}$ is sought using (29). For this field

$$\mathbf{u}_0 = -4(1-\nu)\boldsymbol{\psi}_0 + \nabla(\mathbf{r}\cdot\boldsymbol{\psi}_0 + \phi_0) \equiv \mathbf{0} \tag{31}$$

where $\boldsymbol{\psi}_0$ and ϕ_0 are harmonic, i.e. $\nabla^2\boldsymbol{\psi}_0 = \mathbf{0}$ and $\nabla^2\phi_0 = 0$. Taking the divergence and rotor of equation (31) leads to

$$\nabla\cdot\boldsymbol{\psi}_0 = 0 \tag{32a}$$

$$\nabla\times\boldsymbol{\psi}_0 = \mathbf{0} \tag{32b}$$

Equation (32b) guarantees the existence of scalar potential Φ_0

$$\boldsymbol{\psi}_0 = \nabla\Phi_0 \tag{33}$$

and equation (32a) leads now to

$$\nabla^2\Phi_0 = 0 \tag{34}$$

Null displacement field (31) is now represented as

$$\mathbf{u}_0 = \nabla(\mathbf{r}\cdot\nabla\Phi_0 - 4(1-\nu)\Phi_0 + \phi_0) \equiv \mathbf{0} \tag{35}$$

This shows that

$$\phi_0 = 4(1-\nu)\Phi_0 - \mathbf{r}\cdot\nabla\Phi_0 + c_0 \tag{36}$$

where c_0 is constant. According to equation (36), because scalar potential Φ_0 exists also Papkovich-Neuber-potential ϕ_0 exists. To sum up, the general null displacement field can be represented by

$$\psi_0 = \nabla\Phi_0 \tag{37a}$$
$$\phi_0 = 4(1-\nu)\Phi_0 - \mathbf{r}\cdot\nabla\Phi_0 + c_0 \tag{37b}$$
$$\nabla^2\Phi_0 = 0 \tag{37c}$$

Next, it is assumed, that some complete representation

$$\mathbf{u} = -4(1-\nu)\psi_1 + \nabla(\mathbf{r}\cdot\psi_1 + \phi_1) \tag{38}$$

can be found for a problem at hand. It is possible to add null displacement field to (38), i.e. Papkovich-Neuber-potentials

$$\psi = \psi_1 + \nabla\Phi_0 \tag{39a}$$
$$\phi = \phi_1 + 4(1-\nu)\Phi_0 - \mathbf{r}\cdot\nabla\Phi_0 + c_0 \tag{39b}$$

represents the same displacement field as (38). Defining boundary conditions for scalar potential Φ_0 fixes ψ_0 and ϕ_0, and representation (39) is still complete. Now, the question arises, it is possible to define such boudary conditions for Φ_0 so as to render representation (39) also unique?

Indeed, this is possible, as Stippes [10] has shown. The most convinient method in numerical solution is to force $\mathbf{r}\cdot\psi + \phi$ to be zero at boundary. It is only needed to show that this is possible by defining suitable boundary conditions for Φ_0. From (39)

$$\mathbf{r}\cdot\psi + \phi = \mathbf{r}\cdot\psi_1 + \phi_1 + 4(1-\nu)\Phi_0 + c_0 = 0 \text{ on } \Gamma \tag{40}$$

and defining $c_0 = 0$ and

$$4(1-\nu)\Phi_0 = -\mathbf{r}\cdot\psi_1 - \phi_1 \text{ on } \Gamma \tag{41}$$

the scalar potential Φ_0 is fixed. Thus, representation (29) or (39) with $\mathbf{r}\cdot\psi + \phi = 0$ at boundary is complete.

It is simple mater to show that representation (29) or (39) with $\mathbf{r}\cdot\psi + \phi = 0$ at boundary is unique, if the null displacement field $\mathbf{u}_0 \equiv \mathbf{0}$ can be produced if and only if $\psi_0 = \mathbf{0}$ and $\phi_0 = 0$, where Papkovich-Neuber-potentials ψ_0 and ϕ_0 fulfills the boundary condition $\mathbf{r}\cdot\psi_0 + \phi_0 = 0$ at boundary Γ. According to (32a), $\nabla\cdot\psi_0 = 0$ on D. But (14) shows that

$$\nabla^2(\mathbf{r}\cdot\psi_0 + \phi_0) = 2\nabla\cdot\psi_0 = 0 \text{ on D} \tag{42}$$

Because $\mathbf{r}\cdot\boldsymbol{\psi}_0+\phi_0=0$ also at boundary, it must zero everywhere. Then (31) leads to $-4(1-\nu)\boldsymbol{\psi}_0=0$ on D, and hence $\boldsymbol{\psi}_0\equiv\mathbf{0}$, $\phi_0\equiv 0$.

EXPRESSIONS FOR DISPLACEMENTS, STRAINS AND STRESSES IN CARTESIAN ORTHOGONAL COORDINATE SYSTEM

In order to represent the boundary conditions arising from elastostatics, displacements and stresses needs to be expressed using Papkovich-Neuber-potentials. In Cartesian orthogonal coordinate system, the displacement components are

$$\begin{aligned}
u_x &= -(3-4\nu)\psi_x + x\psi_{x,x} + y\psi_{y,x} + z\psi_{z,x} + \phi_{,x}\\
u_y &= -(3-4\nu)\psi_y + x\psi_{x,y} + y\psi_{y,y} + z\psi_{z,y} + \phi_{,y}\\
u_z &= -(3-4\nu)\psi_z + x\psi_{x,z} + y\psi_{y,z} + z\psi_{z,z} + \phi_{,z}
\end{aligned} \tag{43}$$

The strains are

$$\begin{aligned}
\varepsilon_x &= -2(1-2\nu)\psi_{x,x} + x\psi_{x,xx} + y\psi_{y,xx} + z\psi_{z,xx} + \phi_{,xx}\\
\varepsilon_y &= -2(1-2\nu)\psi_{y,y} + x\psi_{x,yy} + y\psi_{y,yy} + z\psi_{z,yy} + \phi_{,yy}\\
\varepsilon_z &= -2(1-2\nu)\psi_{z,z} + x\psi_{x,zz} + y\psi_{y,zz} + z\psi_{z,zz} + \phi_{,zz}\\
\gamma_{xy} &= 2\left[-(1-2\nu)\left(\psi_{x,y}+\psi_{y,x}\right) + x\psi_{x,xy} + y\psi_{y,xy} + z\psi_{z,xy} + \phi_{,xy}\right]\\
\gamma_{yz} &= 2\left[-(1-2\nu)\left(\psi_{y,z}+\psi_{z,y}\right) + x\psi_{x,yz} + y\psi_{y,yz} + z\psi_{z,yz} + \phi_{,yz}\right]\\
\gamma_{xz} &= 2\left[-(1-2\nu)\left(\psi_{x,z}+\psi_{z,x}\right) + x\psi_{x,xz} + y\psi_{y,xz} + z\psi_{z,xz} + \phi_{,xz}\right]
\end{aligned} \tag{44}$$

and the stresses are

$$\begin{aligned}
\sigma_x &= \frac{E}{1+\nu}\left[-2\left(\nu\nabla\cdot\boldsymbol{\psi} + (1-2\nu)\psi_{x,x}\right) + x\psi_{x,xx} + y\psi_{y,xx} + z\psi_{z,xx} + \phi_{,xx}\right]\\
\sigma_y &= \frac{E}{1+\nu}\left[-2\left(\nu\nabla\cdot\boldsymbol{\psi} + (1-2\nu)\psi_{y,y}\right) + x\psi_{x,yy} + y\psi_{y,yy} + z\psi_{z,yy} + \phi_{,yy}\right]\\
\sigma_z &= \frac{E}{1+\nu}\left[-2\left(\nu\nabla\cdot\boldsymbol{\psi} + (1-2\nu)\psi_{z,z}\right) + x\psi_{x,zz} + y\psi_{y,zz} + z\psi_{z,zz} + \phi_{,zz}\right]\\
\tau_{xy} &= \frac{E}{1+\nu}\left[-(1-2\nu)\left(\psi_{x,y}+\psi_{y,x}\right) + x\psi_{x,xy} + y\psi_{y,xy} + z\psi_{z,xy} + \phi_{,xy}\right]\\
\tau_{yz} &= \frac{E}{1+\nu}\left[-(1-2\nu)\left(\psi_{y,z}+\psi_{z,y}\right) + x\psi_{x,yz} + y\psi_{y,yz} + z\psi_{z,yz} + \phi_{,yz}\right]\\
\tau_{xz} &= \frac{E}{1+\nu}\left[-(1-2\nu)\left(\psi_{x,z}+\psi_{z,x}\right) + x\psi_{x,xz} + y\psi_{y,xz} + z\psi_{z,xz} + \phi_{,xz}\right]
\end{aligned} \tag{45}$$

In the expressions for normal stresses, the governing equations (13) and (20) has been used.

NUMERICAL SOLUTION WITH FEM

As is shown above, the displacement field of linear, materially isotropic and homogeneous elastostatic problem can be represented as

$$\mathbf{u} = -4(1-\nu)\boldsymbol{\psi} + \nabla(\mathbf{r}\cdot\boldsymbol{\psi} + \phi) \tag{46}$$

where Papkovich-Neuber-potentials are governed by

$$\nabla^2\boldsymbol{\psi} = \frac{1}{4(1-\nu)\mu}\mathbf{f} \qquad \text{on D} \tag{47a}$$

$$\nabla^2\phi = -\frac{1}{4(1-\nu)\mu}\mathbf{r}\cdot\mathbf{f} \qquad \text{on D} \tag{47b}$$

One boundary condition is

$$\mathbf{r}\cdot\boldsymbol{\psi} + \phi = 0 \quad \text{on } \Gamma \tag{48}$$

Other boundary conditions arises from elastostatics. Two types, which arises most often, are

$$\text{Type I:} \qquad \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_I \tag{49a}$$

$$\text{Type II:} \qquad \mathbf{t} = \bar{\mathbf{t}} \text{ on } \Gamma_{II} \tag{49b}$$

where $\Gamma = \Gamma_I \cup \Gamma_{II}$, $\Gamma_I \cap \Gamma_{II} = \varnothing$, overbar denotes prescriped function and $\mathbf{t}$ is traction vector.

Some important remarks concerning the problem are as follows:

1. In order to uncouple the components of $\boldsymbol{\psi}$ in (47a), the rectangular cartesian components should be used.
2. It is impossible to choose such a approximation for unknown fields, that some (essential) boundary conditions are satisfied *a priori*.
3. In order to get advantages of the formulation, same shape/trial functions should be used for every unknown field inside the domain D.
4. At the boundary Γ, shape/trial functions can be of different type for unknown fields in order to represent boundary conditions efficiently.

When the problem is solved using standard finite element procedure with Galerkin method, each governing equation is multiplied by test function and integrated over region under consideration. This procedure will lead to equations

$$\int_\Omega \begin{bmatrix} \nabla^2\psi_x v_x \\ \nabla^2\psi_y v_y \\ \nabla^2\psi_z v_z \\ \nabla^2\phi\, w \end{bmatrix} d\Omega = \frac{1}{4(1-\nu)\mu}\int_\Omega \begin{bmatrix} f_x v_x \\ f_y v_y \\ f_z v_z \\ -\mathbf{r}\cdot\mathbf{f} w \end{bmatrix} d\Omega \tag{50}$$

In order to get symmetric stiffness matrix, Green's formula is used

$$\int_\Omega \begin{bmatrix} \nabla\psi_x \cdot \nabla v_x \\ \nabla\psi_y \cdot \nabla v_y \\ \nabla\psi_z \cdot \nabla v_z \\ \nabla\phi \cdot \nabla w \end{bmatrix} d\Omega = \int_\Gamma \begin{bmatrix} v_x \mathbf{n} \cdot \nabla\psi_x \\ v_y \mathbf{n} \cdot \nabla\psi_y \\ v_z \mathbf{n} \cdot \nabla\psi_z \\ w\mathbf{n} \cdot \nabla\phi \end{bmatrix} d\Gamma + \frac{1}{4(1-\nu)\mu} \int_\Omega \begin{bmatrix} -f_x v_x \\ -f_y v_y \\ -f_z v_z \\ \mathbf{r} \cdot \mathbf{f} w \end{bmatrix} d\Omega \qquad (51)$$

Comparing this to expressions of displacements and stresses, it is seen that boundary conditions do not appear in the first integral on right hand side (there are second derivatives in stresses but not in (51), not to mention boundary condition (48)). Obviously, there are no means by which the second derivatives of unknown fields can be incorporated to boundary integrals. For this reason, the Galerkin method is not suitable for numerical solution of this problem.

Because partial integration by the use of Green's formula do not produce suitable boundary integrals, the partial integration should be avoided. One possible formulation is the least squares method, in which the approximations of unknown fields are written as

$$\tilde{\psi}_x = \left[N_x\right]\left\{a_x\right\},\ \tilde{\psi}_y = \left[N_y\right]\left\{a_y\right\},\ \tilde{\psi}_z = \left[N_z\right]\left\{a_z\right\},\ \tilde{\phi} = \left[N_\phi\right]\left\{a_\phi\right\} \qquad (52)$$

where $\left[N_x\right]$, $\left[N_y\right]$, $\left[N_z\right]$ and $\left[N_\phi\right]$ are row matrices (vectors) of global shape functions and $\left\{a_x\right\}$, $\left\{a_y\right\}$, $\left\{a_z\right\}$ and $\left\{a_\phi\right\}$ are column vectors of unknown nodal values. Approximations (52) are substituted to (47) and because these approximations are not exact solutions, residuals R_x, R_y, R_z and R_ϕ are generated. These residuals are functions of unknown nodal values. Next, a non-negative function of unknown nodal values is formed as

$$\begin{aligned} I\left(\left\{a_x\right\},\left\{a_y\right\},\left\{a_z\right\},\left\{a_\phi\right\}\right) = {} & \alpha_x \int_\Omega R_x^2 d\Omega + \alpha_y \int_\Omega R_y^2 d\Omega + \alpha_z \int_\Omega R_z^2 d\Omega + \alpha_\phi \int_\Omega R_\phi^2 d\Omega + \\ & + \alpha_I \int_{\Gamma_I} (\mathbf{u} - \bar{\mathbf{u}}) \cdot (\mathbf{u} - \bar{\mathbf{u}}) d\Gamma + \alpha_{II} \int_{\Gamma_{II}} (\mathbf{t} - \bar{\mathbf{t}}) \cdot (\mathbf{t} - \bar{\mathbf{t}}) d\Gamma + \\ & + \alpha_\Gamma \int_\Gamma (\mathbf{r} \cdot \psi + \phi)^2 d\Gamma \end{aligned} \qquad (53)$$

where the α_i's are suitable positive weighting coefficients. Obviously, it is rational to choose $\alpha_x = \alpha_y = \alpha_z = 1$. Minimization of function (53) with respect to unknown nodal values will yield equations, from which the unknown nodal values can be solved.

It is important to note, that in order to generate nonzero residuals, at least quadratic shape functions should be used, because governing equations (47) includes second derivatives of unknown fields. Also, because formulation by least squares method is considerable different from standard Galerkin method, comparison by simple flop count could be totally misleading.

If nodal values at boundary for each unknown field is collected to one column vector $\{a_B\}$ and all unknown nodal values are collected to one column vector

$$\{a\} = \left[\{a_{xD}\}^T \quad \{a_{yD}\}^T \quad \{a_{zD}\}^T \quad \{a_{\phi D}\}^T \quad \{a_B\}^T \right]^T \tag{54}$$

where subscript D denotes unknown nodal values inside domain D, and the same shape functions are used for each unknown field inside the domain D, the procedure above will lead to symmetric positive definite stiffness matrix

$$[K] = \begin{bmatrix} [K_D] & [0] & [0] & [0] & [K_{xB}]^T \\ [0] & [K_D] & [0] & [0] & [K_{yB}]^T \\ [0] & [0] & [K_D] & [0] & [K_{zB}]^T \\ [0] & [0] & [0] & \alpha_\phi [K_D] & [K_{\phi B}]^T \\ [K_{xB}] & [K_{yB}] & [K_{zB}] & [K_{\phi B}] & [K_B] \end{bmatrix} \tag{55}$$

In triangulation process, the submatrix $[K_D]$ needs only be triangulated once. The number of rows in this matrix is equal to number of inner nodes and its bandwidth is less than bandwidth in solution of one Poisson equation in the same domain using standard Galerkin procedure. Thus, if the proportion of number of inner nodes to number of boundary nodes is large, the procedure should be efficient when concerning speed of solution.

However, there are still plenty of problems. Because boundary integrals in (53) contains derivatives of unknown fields, all nodal values in elements adjacent to boundary falls in to vector $\{a_B\}$. This will signicantly decrease the efficiency of the procedure. Also, the values of weighting coefficients should be determined in some suitable manner. Because these same problems arises in least squares Trefftz finite element method, and because Trefftz method offers superior accuracy, problem should be attacked by this method [15-19].

CONCLUSIONS

An alternative solution strategy for elastostatic problems was proposed. Using Papkovich-Neuber-solution, Navier's equation was substituted by vector Poisson and Poisson eaquation. Based on literature, the completeness of resulting representation was concluded. Using published results, the representation was also forced unique. It was shown that standard Galerkin and least squares finite element methods leads to some problems, and hence it was concluded that some other numerical procedure should be used. Authors consideres the Trefftz finite element method the most appropriate.

REFERENCES

1. Папковиц, П. Ф., *Выражение общего интеграла основых уравнений теории упругости через гармонические функции*, Известия Академии Наук СССР, Отделсние математических и естественных паук 10, 1932, 1425-1435 (Papkovich, P.F., *Expression of the general integral of the basic equations of elasticity through harmonic functions*, Izv. Akad. Nauk SSSR, Phys.-Math. Ser **10**, 1932, pp. 1425-1435)
2. Papkovich, P.F., *Solution générale des équations différentielles fondamentales d'élasticité, exprimée par trois fonctions harmoniques*, Comptes Rendus Acad. Sci. **195**, Paris, 1932, pp. 513-515
3. Neuber, H., *Ein neuer Ansatz zur Lösung räumlicher Probleme der Elastizitätstheorie. Der Hohlkegel unter Einzellast als Beispiel.*, Ztschr. f. angew. Math. Mech. (ZAMM) **14**, 1934, pp. 203-212
4. Mindlin, R.D., *Note on the Galerkin and Papkovitch Stress Functions*, Bull. Amer. Math. Soc., Vol **42**, 1936, pp. 373-376
5. Слободянский, М. Г., *Общие формы решений уравнений упругости для односвязных и многосвязных областей, выраженные через гармонические функции*, Прикладная математика и механика (Институт механики Академии наук Союза ССР), Там 18, 1954, 55-74 (Slobodyansky, M.G., *General forms of solutions of equations of elasticity for singly and multiply connected domains expressed through harmonic functions*, Prikl. Mat. Mekh. Akad. Nauk SSSR, Vol. **18**, 1954, pp. 55-74)
6. Eubanks, R.A., Sternberg, E., *On the Completeness of the Boussinesq-Papkovich Stress Functions*, J. Rational Mech. Anal, Vol. **5**, No. 5, 1956, pp. 735-746
7. Naghdi, P.M., Hsu, C.S., *On a Representation of Displacements in Linear Elasticity in Terms of Three Stress Functions*, Journal of Mathematics and Mechanics, Vol. **10**, No. 2, 1961, pp. 233-245
8. Sternberg, E., Gurtin, M.E., *On the completeness of certain stress functions in the linear theory of elasticity*, Proc. 4th U.S. Nat. Cong. Appl. Mech., 1962, pp. 793-797
9 Gurtin, M.E., *On Helmholtz's Theorem and the Completeness of the Papkovich-Neuber Stress Functions for Infinite Domains*, Archive for Rational Mechanics and Analysis, Vol. **9**, 1962, pp. 225-233
10. Stippes, M., *Completeness of the Papkovich potentials*, Quarterly of Applied Mathematics, Vol. **26**, No. 4, 1969, pp. 477-483
11. Pecknold, D.A.W., *On the role of the Stokes-Helmholtz decomposition in the derivation of displacement potentials in classical elasticity*, Journal of Elasticity, Vol. **1**, No. 2, 1971, pp. 171-174
12. Bentham, J.P., *Note on the Boussinesq-Papkovich stress-functions*, Journal of Elasticity, Vol. **9**, No. 2, 1979, pp. 201-206
13. Tran Cong, T., Steven, G.P., *On the representation of elastic displacement field in terms of three harmonic functions*, Journal of Elasticity, Vol. **9**, No. 3, 1979, pp. 325-333
14. Millar, R.F., *On the completeness of the Papkovich potentials*, Quarterly of Applied Mathematics, Vol. **41**, No. 4, 1984, pp. 385-393
15. Jirousek, J., Teodorescu, P., *Large finite element method for the solution of problems in the theory of elasticity*, Computers & Structures, Vol. **15**, No. 5, 1982, pp. 575-587

16. Jirousek, J., Wroblewski, A., *Least squares T-elements: Equivalent FE and BE forms of a substructure-oriented boundary solution approach*, Communications in Numerical Methods in Engineering, Vol. **10**, 1994, pp. 21-32
17. Jirousek, J., Stojek, M., *Numerical assessment of a new T-element approach*, Computers & Structures, Vol. **57**, No. 3, 1995, pp. 367-378
18. Kita, E., Kamiya, N., *Trefftz method: an overview*, Adv. in Engineering Software, Vol. **24**, 1995, pp. 3-12
19. Jirousek, J., Wroblewski, A., *T-elements: a finite element approach with advantages of boundary solution methods*, Adv. in Engineering Software, Vol. **24**, 1995, pp. 71-88
20. Sneddon, I.N., Berry, D.S., *The Classical Theory of Elasticity*, In: S. Flügge (ed.), Encyclopedia of Physics, Vol VI: Elasticity and Plasticity, Springer-Verlag, Berlin, 1958
21. Gurtin, M.E., *The Linear Theory of Elasticity*, In: S. Flügge (chief ed.), Encyclopedia of Physics, Vol VIa/2: Mechanics of Solids II, (ed. by C. Truesdell), Springer-Verlag, Berlin, 1972
22. Golub, G.H, Van Loan, C.F., *Matrix Computations*, 2nd ed., The John Hopkins University Press, Baltimore and London, 1989
23. Barber, J.R., *Elasticity*, Kluwer Academic Publishers, Dordrecht/Boston/London, 1992
24. Poruchikov, V.B., *Methods of the Classical Theory of Elastodynamics*, Springer-Verlag, Berlin, 1993
25. Brekhovskikh, L.M., Goncharov, V., *Mechanics of Continua and Wave Dynamics*, 2nd ed., Springer-Verlag, Berlin, 1994

FEM ANALYSIS OF A TRAVELING PAPER WEB AND SURROUNDING AIR

JARI LAUKKANEN and ANTTI PRAMILA
University of Oulu
Department of Mechanical Engineering
Laboratory of Engineering Mechanics
BOX 444 Linnanmaa, 90571 Oulu, FINLAND

ABSTRACT

The increase of running speed in paper machines and printing presses has increased the attention received by the vibration and stability problems of the traveling paper web. The present paper shows a new numerical model where the finite width of the paper web and the coupling of the vibration of the web and surrounding air are taken into account. A novel modification of *Lanczos* algorithm has been developed in order to improve the accuracy of the eigenvalues obtained. Numerical results have been compared with available analytical and experimental results in some special cases. The agreement has been good. Moreover, parametric studies concerning the effect of some design parameters of the system have been done.

INTRODUCTION

With the increase in speed of paper machines and printing presses, the problems associated with the vibration and stability of paper web have received much attention. The achievable operating speed is limited by the instability of the web, by excessive vibration amplitudes or

by the threshold of the formation of wrinkles. Therefore the understanding of the effects of the various design parameters to the dynamic behaviour of the system has become more and more important.

The present vibration problem belongs to a broader class of problems called vibration and stability of axially moving material. Typical such systems having engineering importance are fluid conveying pipes, traveling strings, belt drives and band saws. Recent reviews of the research on these fields can be found from references [1] and [2]. The axially moving material problems involve the salient feature of having three inertial terms instead of one in "usual" vibration problems. The two additional terms are due to the convective acceleration caused by the continuous traveling of the material in its own plane.

Because of the dimensions of the practical problem under consideration (width of the web larger than the roll distance) one dimensional models are not adequate. Previously there have been only a few models where the axially moving material has been described as a two dimensional plate or membrane medium, e.g. [3] and [4]. The effect of the surrounding fluid has, however, been neglected in the abovementioned studies. The coupling of the vibratory motion of a submerged solid body to the surrounding fluid decreases generally the natural frequencies and the lighter is the vibrating solid the more pronounced is this effect. Experimental [5] and [6] analytical results obtained earlier have shown that with dimensions typical in a pilot paper mill the first natural frequency will be overestimated by up to 400% if the coupling is neglected. Preliminary results obtained by FEM and based on the ideal fluid assumption [7] showed that also the geometry of the surrounding fluid field has a considerable effect on the dynamic behaviour of the paper web.

The present study is continuation to the study described in [7]. Now, the compressibility of the air is taken into account. A special purpose FEM program system has been developed for the analysis [8]. It contains a modified version of the *Lanczos* algorithm with convergence control and improved computational accuracy of eigenvectors. A separate post processing program has been developed for the visualization of the computational results.

The numerical results agree with the previously obtained analytical and experimental results available. Parametric studies done on the effect of the geometry of the fluid domain have, for example, revealed that the increase of the diameter of the rolls decreases the eigenfrequencies of the web.

EQUATIONS OF MOTION

The system under consideration consists of a membrane (width b , length a) traveling between two roll-supports with constant velocity v in the positive x-direction originally in the x,y-plane subject to in-plane forces $T_{xx}(x,y)$, $T_{xy}(x,y)$ and $T_{yy}(x,y)$ per unit length (assumed to be independent of time) and a transverse force $q(x,y,t)$ per unit area. The transverse motion of the sheet is formulated using spatial coordinates. Thus, the kinetic energy of the portion of the membrane between the supports is

$$T = \frac{1}{2}\int_{-\frac{b}{2}}^{\frac{b}{2}}\int_{0}^{a}\rho_p\left[\left(\frac{\partial w}{\partial t} + v\,\frac{\partial w}{\partial x}\right)^2 + v^2\right]dx\,dy \quad , \tag{1}$$

where ρ_p denotes the mass per unit area of the sheet. The potential energy of the membrane between supports is

$$V = \frac{1}{2}\int_{-\frac{b}{2}}^{\frac{b}{2}}\int_{0}^{a}\left(T_{xx}\frac{\partial^2 w}{\partial x^2} + 2T_{xy}\frac{\partial^2 w}{\partial x\,\partial y} + T_{yy}\frac{\partial^2 w}{\partial y^2}\right)dx\,dy \quad . \tag{2}$$

The equation of motion in the undamped case can be obtained using *Hamilton's* principle in the same way as in a similar problem - dynamics of fluid conveying pipes. The *Hamilton's* principle takes the familiar form

$$\delta\int_{t_1}^{t_2}(T - V)\,dt = 0 \tag{3}$$

provided that the transverse displacement is prescribed as zero both at the inlet boundary and at the outlet boundary, as is the case in the present problem.

By substituting expressions (1) and (2) for T and V in equation (3) we obtain in a standard manner the equation of motion

$$\rho_p \frac{\partial^2 w}{\partial t^2} + 2\rho_p v \frac{\partial^2 w}{\partial x \partial t} + \rho_p v^2 \frac{\partial^2 w}{\partial x^2} - T_{xx} \frac{\partial^2 w}{\partial x^2} - 2T_{xy} \frac{\partial^2 w}{\partial x \partial y} - T_{yy} \frac{\partial^2 w}{\partial y^2} = q \quad . \tag{4}$$

The present equation of motion differs from the usual membrane equation in the two additional inertia terms (second and third) due to the in-plane velocity v.

For the fluid around the traveling membrane the following assumptions are made: 1) velocities are small enough for convective effects to be omitted 2) the pressure-density behaviour is locally linear and varies by a small amount only 3) the viscous effects can be neglected. Based on these assumptions the governing equation in fluid domain is

$$\nabla^2 p - \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} = 0 \quad , \tag{5}$$

where $c = \sqrt{B/\rho}$ is the speed of the sound and B is the bulk modulus. This equation is the well known *Helmholtz* equation. If the fluid does not separate from the membrane, the pressure gradient and the velocity of the membrane along the outward normal must satisfy equation

$$\frac{\partial p}{\partial n} = -\rho \frac{\partial v_n}{\partial t} = -\rho \dot{v}_n = -\rho \ddot{u}_n \quad . \tag{6}$$

The normal is pointing from the fluid to the membrane. If the fluid is in contact with the rigid wall the boundary condition becomes $\partial p/\partial n = 0$. If the fluid is assumed to extend to infinity (e.g. in comparison with some available analytical results), the motion is assumed to vanish at infinity and therefore the dynamic pressure must vanish there, i.e. $p = 0$. In a numerical representation an infinite boundary has always to be truncated at some sufficiently large distance. At such a boundary the truncation boundary condition

$$\frac{\partial p}{\partial n} = -\frac{1}{c} \frac{\partial p}{\partial t} \tag{7}$$

must be introduced.

When expressions (1) and (2) are substituted into equation (3) and w is replaced by FEM trial $w(x,y,t) = \boldsymbol{N}\boldsymbol{u}(t)$ and correspondingly in equation (5) p is replaced by FEM trial $p(x,y,z,t) = \check{\boldsymbol{N}}\boldsymbol{p}(t)$ and *Galerkin's* method is used for it, we obtain finally the equations of motion

$$\begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \rho\boldsymbol{S} & \boldsymbol{Q} \end{bmatrix} \begin{Bmatrix} \ddot{\boldsymbol{u}} \\ \ddot{\boldsymbol{p}} \end{Bmatrix} + \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} \begin{Bmatrix} \dot{\boldsymbol{u}} \\ \dot{\boldsymbol{p}} \end{Bmatrix} + \begin{bmatrix} \boldsymbol{K} & -\boldsymbol{S}^T \\ \boldsymbol{0} & \boldsymbol{H} \end{bmatrix} \begin{Bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{Bmatrix} = \begin{Bmatrix} f_s \\ f_f \end{Bmatrix} , \tag{8}$$

where $\boldsymbol{K}$, $\boldsymbol{M}$ and $\boldsymbol{H}$, $\boldsymbol{Q}$ are the stiffness and mass matrices of the membrane and the fluid respectively. $\boldsymbol{G}$ is the skew-symmetric gyroscopic inertia matrix of the membrane and $\boldsymbol{D}$ is the radiation damping matrix of the fluid. f_s and f_f are the force vectors of structure and fluid respectively. The coupling between the surrounding fluid and membrane part occurs via the matrix $\boldsymbol{S}$.

Substituting $(\boldsymbol{u}\ \boldsymbol{p})^T = \boldsymbol{z}e^{\lambda t}$ into equation (8) and neglecting the right-hand side force vector yields

$$\left(\lambda^2 \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \rho\boldsymbol{S} & \boldsymbol{Q} \end{bmatrix} + \lambda \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} + \begin{bmatrix} \boldsymbol{K} & -\boldsymbol{S}^T \\ \boldsymbol{0} & \boldsymbol{H} \end{bmatrix} \right) \boldsymbol{z} = \boldsymbol{0} , \tag{9}$$

which is an unsymmetric eigenvalue problem with complex eigenvalues and eigenvectors. The eigenvalues are of type $\lambda = \sigma + i\,\omega$, where σ and ω are real numbers and i is the imaginary unit. Thus, two types of instability are possible: $\lambda = 0$ indicates divergence and $\sigma > 0$ indicates flutter.

RESULTS

The eigenvalue problem (9) was solved by computer program [8] based on *Lanczos* algorithm. A membrane travelling between two roll-supports was studied using FEM model of 110 bilinear elements for membrane and 6644 8-node brick elements for fluid. The length of the membrane is $a = 2.4$ m , length/width ratio $a/b = 5.1$. In figure 1 the four lowest non-dimensional eigenfrequencies $F = f_i\ 2a\sqrt{\rho_p/T_x}$ as function of the non-dimensional velocity

$V = v_i \sqrt{\rho_p / T_x}$ are presented and the importance of surrounding fluid for the eigenfrequencies of the light membranes is clearly seen. The experimental results in the figure are from reference [6].

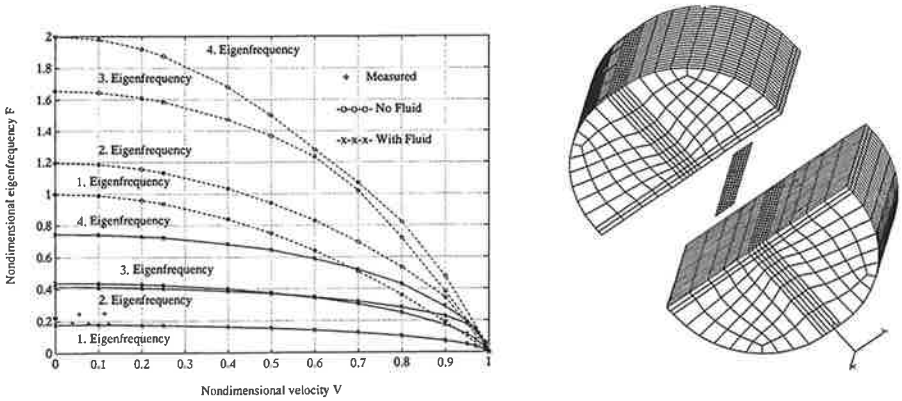


Figure 1. Non-dimensional eigenfrequency as function of non-dimensional velocity and the model of the membrane and surrounding air [8]. The membrane and the fluid are separated for clarity. Boundary condition $p = 0$ on the cylindrical fluid-boundary. Density of air $\rho = 1.3$ kg/m^3 and speed of the sound $c = 340$ m/s. The weight per unit area of the membrane is $\rho_p = 35.5\ 10^{-3}$ kg/m^2 and the tensions $T_x = 600$ N/m, $T_y = 10$ N/m, $T_{xy} = 0$. The height of the fluid area is $5.25b$ and the width $12.38b$.

The effects of coupling of two membrane parts are studied using L-shape membrane traveling through three roll-supports whose diameter is neglected. The width of the membrane is 0.47 m and the length of the horizontal and vertical parts are 2.4 m and 2.6 m, respectively. The membrane is modelled using 200 bilinear elements and fluid area contains 13641 8-node brick elements. The coupling effect is best seen in eigenmodes and in figure 2 it can be seen when V = 0.5 ... 0.9 [8].

In practice the diameters of the supporting rolls can't be neglected because they change the shape of fluid domain and affect the fluid flow. The effect of the diameter of roll-supports is studied by using the model in figure 3. The membrane is modeled with 480 bilinear elements and the fluid with 10560 8-node brick elements. The width of the membrane, which is

traveling with constant velocity through two two-roll nips, is 8 m and the length 2.4 m. The membrane is assumed to travel between two rigid walls. From the results in figure 3 it is clearly seen the decreasing of eigenfrequencies when the roll diameter is increasing [8].

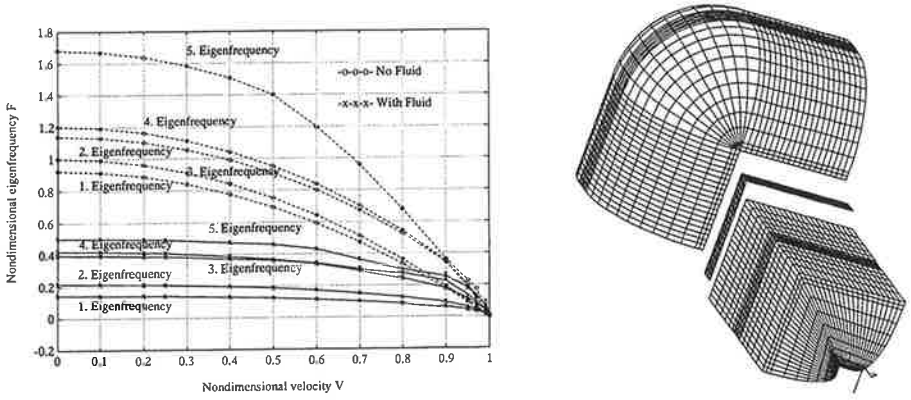


Figure 2. Non-dimensional eigenfrequency as function of non-dimensional velocity ($F = f_i 2a\sqrt{\rho_p/T_x}$, a = 2.4 m) and the model of the membrane and surrounding air. The membrane and the fluid are separated for clarity. Boundary condition $p = 0$ on the cylindrical fluid-boundary. Density of air $\rho = 1.3$ kg/m^3 and speed of the sound $c = 340$ m/s. The weight per unit area of the membrane is $\rho_p = 35.5\ 10^{-3}$ kg/m^2 and the tensions $T_x = 600$ N/m, $T_y = 10$ N/m, $T_{xy} = 0$ length/width ration $a/b = 5.1$ for horizontal section and $a/b = 5.53$ for vertical section. The height of the fluid area is $5.25b$ and the width $12.38b$.

The system (8) is a coupled second-order differential equation. Various solution schemes for coupled problems have been suggested by Park and Felippa [9], Paul [10] and Felippa and Geers [11]. The difficulties with field elimination methods are that order of resulting differential equation is higher, sparseness of matrices are lost and special algorithms are required due to new initial conditions. The method of simultaneous solutions also poses some computational difficulties because the resulting equations (8) are unsymmetric. Attempts to make them symmetric leads to loss in bandedness of the resulting equations. The method of partitioning overcomes the above mentioned limitations. Here the structure or the fluid field may be integrated by implicit, explicit or mixed time integration scheme on two different

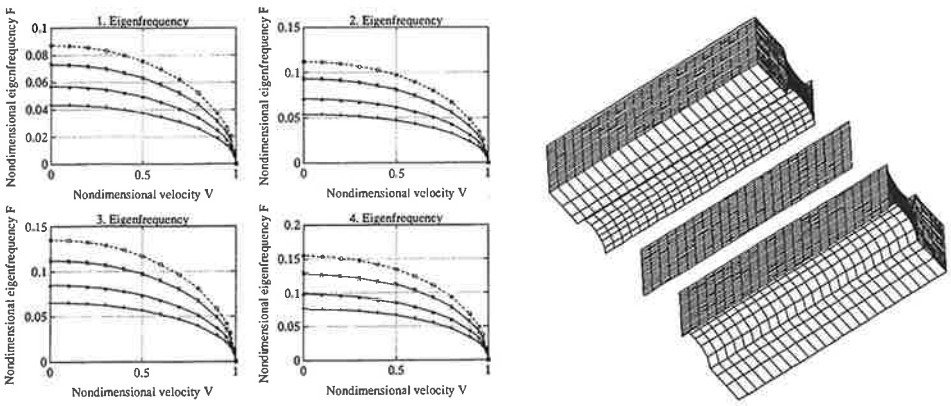


Figure 3. Non-dimensional eigenfrequency as function of non-dimensional velocity and different diameters of supporting rolls. -o-o- $d/a = 0.00$, -x-x- $d/a = 0.21$, -*-*- $d/a = 0.42$, -+-+- $d/a = 0.625$. The model of the membrane and surrounding air. The membrane and the fluid are separated for clarity. Boundary condition $dp/dn = 0$ on short faces. Density of air $\rho = 1.3$ kg/m^3 and speed of the sound $c = 340$ m/s. The weight per unit area of the membrane is $\rho_p = 35.5\ 10^{-3}$ kg/m^2 and the tensions $T_x = 600$ N/m, $T_y = 10$ N/m, $T_{xy} = 0$. The height of the fluid area is $0.75b$ and the width b.

meshes in a staggered fashion and interaction effects can also be accounted. In practice there is two solution sequences: 1) first structure then fluid, i.e. predicted pressure is applied to the structure and the corrected response after solution of the structure equation is transferred to the fluid to take into account the interaction effect. 2) is just opposite, first fluid then structure. Coupled problems with various mesh partitioning schemes along with the predictor-multi-corrector algorithm are not easily amenable to stability analysis. Some general notations are available in references [9]-[11], where it is concluded that stability depends on integrator, mesh partition, predictor formula and computational path. However, based on information available, stable algorithms can be selected and optimum mesh partitioning is possible. Based on our experience of solving equation (8) solution sequence 1) is recommended because it gives the best convergence behaviour. In the present work the Newmark's implicit integration scheme in predictor-multi-corrector form with parameters $\beta = 0.25$ and $\gamma = 0.5$ is used. A

tolerance of 1.0E-06 was used as convergence criteria on the ratio of norm of incremental field variables (pressure or displacement) with norm of total field variable.

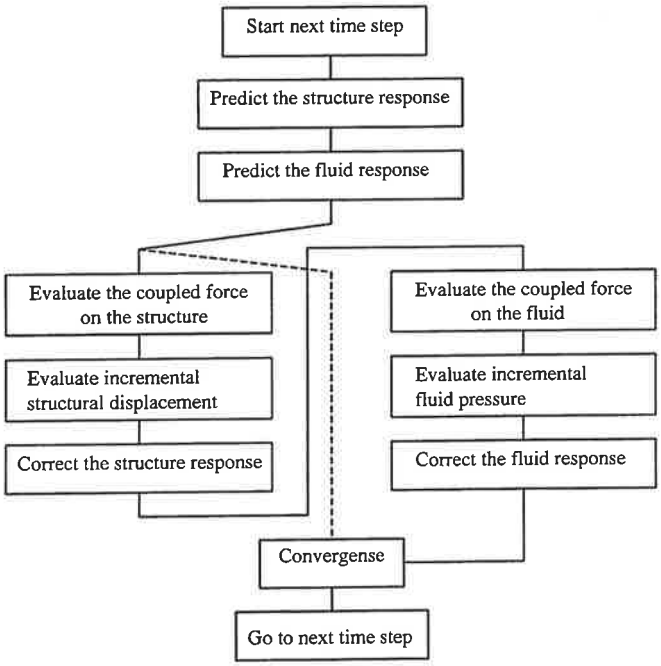


Figure 4. Flow diagram for fluid-structure interaction problem.

The effects of dynamic loading to a membrane traveling between two roll-supports was studied using FEM model in figure 1. The dynamic load applied at every node at line $x = a/7.3$ is shown in figure 5. The displacement responses with and without surrounding fluid are shown in figures 6. The response is plotted based on mid-span displacements ($x = a/2, y = b/2$). From figures 6 it is clearly seen the effect of velocity to the response.

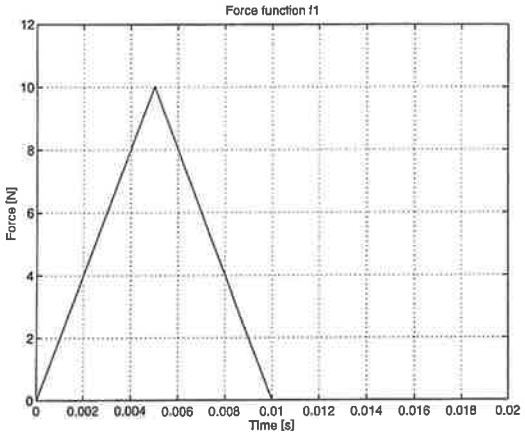


Figure 5. Time dependent loading **f1** used in calculation.

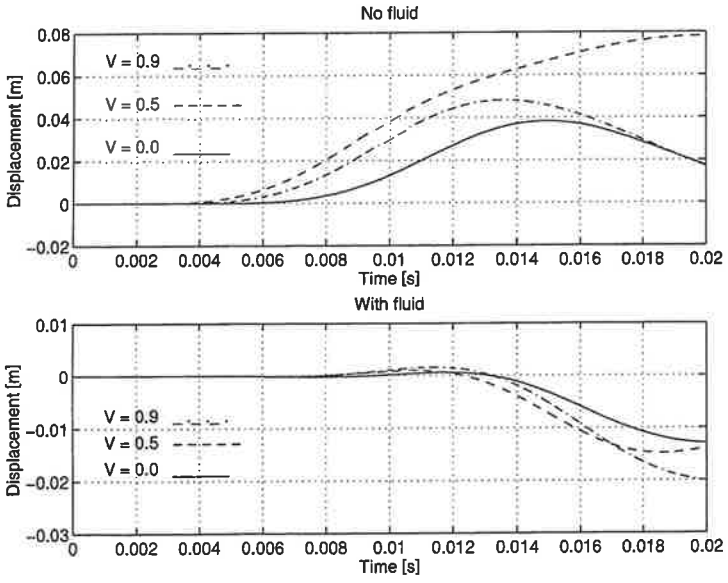


Figure 6. The mid-span ($x = a/2, y = b/2$) dispacement response due to load **f1** at line $x = a/7.3$ with and without fluid, $\Delta t = 0.01$ ms.

REFERENCES

1. Paidoussis, M.P., Li, G.X., *Pipes conveying fluid: a model dynamical problem*, Journal of Fluids and Structures, (1993) 7, 137-204.
2. Wickert, J.A., Mote, Jr., C.D., *Current research on the vibration and stability of axially-moving materials*, Shock and Vibration Digest, May 1988, 3-13.
3. Ulsoy, A.G., Mote, Jr., C.D., *Band saw vibration and stability*, Shock and Vibration Digest, 10, 3-5, (1978).
4. Lengoc, L., McCullion, H., *Wide bandsaw blades under cutting conditions*, Journal of Sound and Vibration, 186(1), 125-142 Part I, 143-162 Part II, 163-179 Part III, (1995).
5. Pramila, A., *Sheet flutter and the interaction between sheet and air*, TAPPI Journal, 69, 79-74, (1986).
6. Pramila, A., *Natural frequencies of a submerged axially moving band*, Journal of Sound and Vibration, 113(1), 198-203, (1987).
7. Niemi, J., Parmila, A., *FEM-analysis of transverse vibrations of an axially moving membrane immersed in ideal fluid*, International Journal for Numerical Methods in Engineering, Vol 24, 2301-2313, (1987).
8. Laukkanen, J., *Numerical vibration analysis of axially moving membrane and surrounding fluid*, (Licentiate thesis, in Finnish), University of Oulu, Department of Mechanical Engineering, (1993).
9. Park, K.C., Felippa, C.A., *Partitioned analysis of coupled systems*, in Computational Methods for Transient Analysis (Eds. T. Belytschko and T.J.R. Hughes), North-Holland, Amsterdam, Ch. 4 (1983)
10. Paul, D.K., *Single and coupled multifield problems*, PhD Thesis, University College of Swansea (1982).
11. Felippa, C.A., Geers, T.L., *Partitioned analysis for coupled mechanical systems*, Eng. Comput., 5, 123-133 (1988)

A FINITE DIFFERENCE SHOOTING METHOD FOR GENERATING AXISYMMETRIC ELEMENTS

PENTTI TUOMINEN
Laboratory of Structural Engineering
University of Oulu
Kasarmintie 4, 90100 Oulu, FINLAND

1. PURPOSE OF THE METHOD AND A BRIEF DESCRIPTION OF IT

The paper deals with the case of the finite difference method (FDM) in connection with the finite element method (FEM). The purpose of the method is to generate macroelements for FEM. If possible elements should be some way natural with a known accuracy.

The method is a direct method using numerical solutions of differential equations. It is a variation of FDM and is formulated as a pure transfer matrix method. The basis of the differential equations are the expressions of normal forces and couples.

Seven different element types have been programmed. These elements are linear and orthotropic. Their displacements are small and loadings axisymmetric. As an example are presented equations and the flow of calculations for open spherical shells.

2. NOTATIONS

In the present method the meridional line of an axisymmetric structure is divided into n equal segments which determine grid points from 1 to $n + 1$. The grid point 1 is situated at the final boundary. The grid point sare indicated with subindeces. 'Index' $i + ½$ or a corresponding fraction is used for the mean point between grid points i and $i + 1$. The mesh lenght between two neighbouring grid points is Δs. Superindices s and t indicate

meridional and circumferential directions of the element. In Figure 1 there is presented the geometry of a spherical shell to light the notations used. The stress resultants and couples are further shown in Figure 2 as positive.

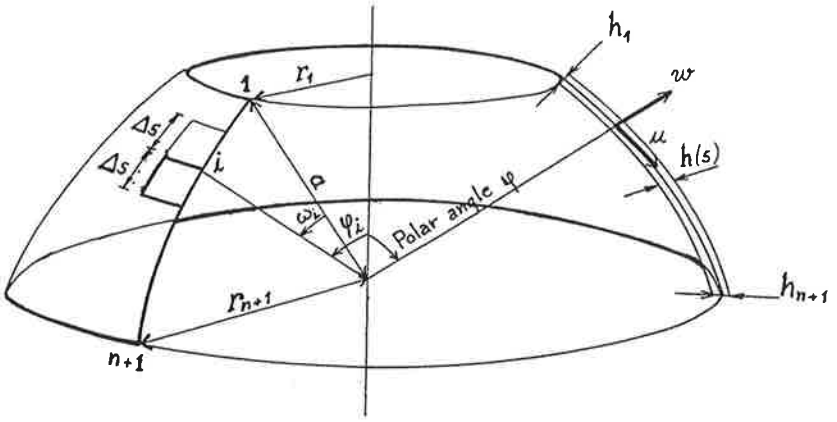


FIGURE 1: Dimensions and displacements of a spherical shell. Quantities are shown as positive.

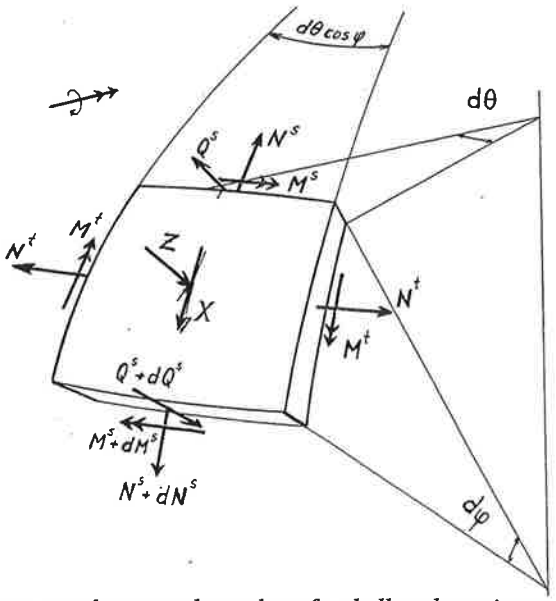


FIGURE 2: Stress resultants and couples of a shell under axisymmetric loading.

3. DIFFERENTIAL EQUATIONS

The usual system of six first order equations is replaced with one first order and one second order differential equation combined with numerical integration scheme for handling the right hand sides of the equations. The two basic equations are obtained by rewriting the expressions of the meridional normal force and couple in an inverted order. Equations for a spherical shell are [1, pp. 433-434]

$$\frac{du}{ds}+\mu\cos\phi\ \frac{u}{r}+(1+\mu)\ \frac{w}{a}\ =\ \frac{N^s}{C^s} \tag{1}$$

$$\frac{d^2w}{ds^2}+\frac{\mu\cos\phi}{r}\frac{dw}{ds}+(1+\mu)\frac{w}{a^2}=\frac{M^s}{B^s}+\frac{1}{a}\frac{N^s}{C^s}, \tag{2}$$

where μ is Poisson's ratio. B^s and C^s are the meridional bending and stretching stiffnesses of the shell. Other termes of equations are presented in Figures 1 and 2. The first version to solve these equation using the shooting method at hand was presented in [3, pp. 407-416]. For a cylindrical shell or for a beam structure Equations 1 and 2 get their simplest forms. The left hand side of the former equation is then $du/ds+\mu w/a^2$ and that of the latter one $d^2w/ds^2+\mu w/a^2$. The normal force N^s stays also away from the right hand side of Equation 2. The meaning of these equations is obvious.

4. HANDLING OF THE LEFT HAND SIDE OF DIFFERENTIAL EQUATIONS

Discretizing the derivatives of the displacements u and w is made with the finite differences

$$\left(\frac{du}{ds}\right)_{i+\frac{1}{2}}=\frac{-u_i+u_{i+1}}{\Delta s}+0(\Delta s^2), \tag{3}$$

$$\left(\frac{dw}{ds}\right)_i=\frac{-w_{i-1}+w_{i+1}}{2\Delta s}+0(\Delta s^2)\ , \tag{4}$$

$$\left(\frac{d^2w}{ds^2}\right)_i=\frac{w_{i-1}-2w_i+w_{i+1}}{\Delta s^2}+0(\Delta s^2)\ , \tag{5}$$

and

$$\psi_i = \frac{u_i}{a} - \left(\frac{dw}{ds}\right)_i = \frac{u_i}{a} - \frac{-w_{i-1} + w_{i+1}}{2\Delta s} + 0(\Delta s^2) \quad . \tag{6}$$

Equation 6 gives the rotation of the meridian at grid point i and $0(\Delta s^n)$ is a term which goes to zero as Δs^n. After some manipulations three recursive equations are obtained

$$d_j u_{i+1} = \left[1 - \frac{\mu_j \Delta s \cos\phi_j}{2r_j}\right] u_i - \left(\frac{\Delta s}{2a} + \frac{\mu_j \Delta s \sin\phi_j}{2r_j}\right)(w_i + w_{i+1}) + \Delta s \frac{N_j^2}{C_j^s} + 0(\Delta s^3) \quad , \quad \left(j = i + \frac{1}{2}\right) \tag{7}$$

$$w_{i+1} = \left[1 - \frac{\mu_i \Delta s \cos\phi_i}{2r_i}\right] \frac{\Delta s}{a} u_i + \left[1 - \frac{\Delta s^2}{2a^2}(1 - \mu_i)\right] w_i - \Delta s \left(1 - \frac{\mu_i \Delta s \cos\phi_i}{2r_i}\right) \psi_i + \frac{\Delta s^2}{2}\left(\frac{M_i^s}{B_i^s} + \frac{1}{a}\frac{N_i^s}{C_i^s}\right) + 0(\Delta s^3) \tag{8}$$

$$\psi_{i+1} = \frac{u_{i+1}}{a} - \frac{-w_i + w_{i+1}}{d_{i+1}\Delta s} + \frac{(1 + \mu_{i+1})\Delta s^2}{d_{i+1} 2a^2} w_{i+1} - \frac{\Delta s}{2d_{i+1}}\left(\frac{M_{i+1}^s}{B_{i+1}^s} + \frac{1}{a}\frac{N_{i+1}^s}{C_{i+1}^s}\right) + 0(\Delta s^3) \quad , \tag{9}$$

In these equations the notation d_i is $d_i = 1 + \frac{\mu_i \Delta s \cos\phi_i}{2r_i}$.

Equations determine the displacement vector

$$v_i = \begin{Bmatrix} u_{i+1} \\ w_{i+1} \\ \psi_{i+1} \end{Bmatrix} \quad ,$$

if the normal force and the bending moment of Equations 7, 8 and 9 can be calculated at grid points $j = i + \frac{1}{2}$ and i+1. This problem is considered in the next section.

5. HANDLING THE RIGHT HAND SIDES OF DIFFERENTIAL EQUATIONS

Equations of this section are based on consideration the equilibrium of a meridional strip between grid points i and i+1 on a spherical shell. Considerations of loadings are omitted for shorting this paper. The strip is shown in Figure 3.

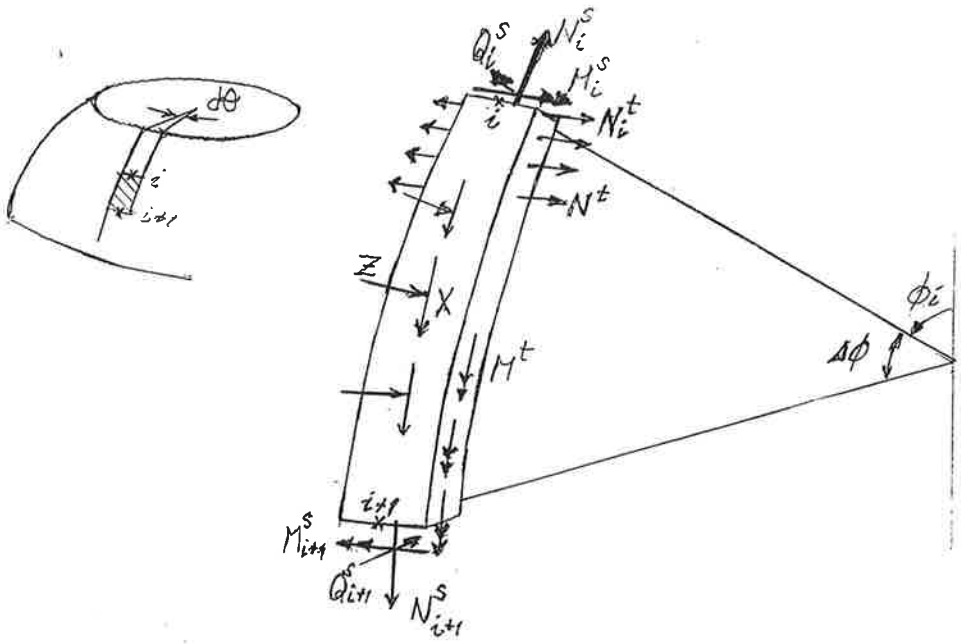


FIGURE 3: A meridional strip of a spherical shell and forces acting on it.

The consideration of equilibrium in the direction of N_{i+1}^s will lead to the expression

$$N_{i+1}^s = \frac{r_i}{r_{i+1}}\left(N_i^s \cos\Delta\phi + Q_i^s \sin\Delta\phi\right) + \frac{\cos\phi_{i+1}}{r_{i+1}} \int_{S_i}^{S_{i+1}} N' ds$$
$$- \frac{1}{r_{i+1}} \int_{S_i}^{S_{i+1}} \left[Z \sin(\phi_{i+1} - \phi) + X \cos(\phi_{i+1} - \phi)\right] r ds \qquad (11)$$

for the meridional normal force.

The quadrature of the circumferential normal force N^t is approximated using the the trapezoidal rule as

$$\int_{S_i}^{S_{i+1}} N^t ds \approx \frac{\Delta s}{2}\left(N_i^t + N_{i+1}^t\right). \tag{12}$$

The substitution

$$N_{i+1}^t = N_k^t = \mu_k N_k^S + \frac{(1+\mu_k \upsilon_k)C_k^t}{r^k}(u_k \cos\phi_k + w_k \sin\phi_k) \tag{13}$$

into 12 and further into 11 with some rearranging give the recursive formula

$$\begin{aligned}
&\left(1-\frac{\mu_{i+1}\Delta s\cos\phi_{i+1}}{2r_{i+1}}\right)N_{i+1}^s = \frac{r_i}{r_{i+1}}\left(N_i^s\cos\Delta\phi + Q_i^s\sin\Delta\phi\right)\\
&+\frac{\Delta s\cos\phi_{i+1}}{2r_{i+1}}N_i^t + \frac{\Delta s\cos\phi_{i+1}}{2r_{i+1}}\frac{(1-\mu_{i+1}\upsilon_{i+1})C_{i+1}^t}{r_{i+1}}(u_{i+1}\cos\phi_{i+1} + w_{i+1}\sin\phi_{i+1})\\
&-\frac{1}{r_{i+1}}\int_{s_i}^{s_{i+1}}\left[Z\sin(\phi_{i+1}-\phi) + X\cos(\phi_{i+1}-\phi)\right]r ds
\end{aligned} \tag{14}$$

for calculating the normal force N_{i+1}^s. After that the circumferential normal force N_{i+1}^t can be computed according to Equation 13.

At the mean point $j = i+1/2$ the formula for N^s has a shorter form

$$\begin{aligned}
N_j^s = &\frac{r_i}{r_j}\left(N_i^s\cos\frac{\Delta\phi}{2} + Q_i^s\sin\frac{\Delta\phi}{2}\right) + \frac{\Delta s\cos\phi_j}{2r_j}N_i^t\\
&-\frac{1}{r_j}\int_{s_i}^{s_j}\left[Z\sin(\phi_i-\phi) + X\cos(\phi_j-\phi)\right]r ds\,.
\end{aligned} \tag{15}$$

where $j = i + 1/2$.

The consideration of the equilibrium in the direction of the shearing force Q_{i+1}^s will give the equation

$$Q_{i+1}^{s} = \frac{r_i}{r_{i+1}}\left(-N_i^s \sin\Delta\phi + Q_i^s \cos\Delta\phi\right) - \frac{\Delta s}{2a}\left(N_i' + N_{i+1}'\right) - \frac{1}{r_{i+1}}\int_{s_i}^{s_{i+1}}\left[Z\cos(\phi_{i+1}-\phi) - X\sin(\phi_{i+1}-\phi)\right]r ds\,, \tag{16}$$

where the circumferential normal force N_{i+1}' was calculated using Equation 13.

The moment equilibrium of the strip about the circumferential axis through the grid point $i+1$ produces the expression

$$M_{i+1}^{s} = \frac{r_i}{r_{i+1}}\left[M_i^s + N_i^s a(1-\cos\Delta\phi)\right] + Q_i^s a\sin\Delta\phi - \frac{1}{r_{i+1}}\int_{s_i}^{s_{i+1}} N' a(\cos\phi_i - \cos\phi) ds + \frac{1}{r_{i+1}}\int_{s_i}^{s_{i+1}} M'\cos\phi ds + \frac{1}{r_{i+1}}\int_{s_i}^{s_{i+1}}\left[Z\sin(\phi_{i+1}-\phi)\right] + X\left[1-\cos(\phi_{i+1}-\phi)\right] a r ds \tag{17}$$

for the meridional couple.

In this equation the integrals are again aprroximated using the trapezoidal rule. The two first quadratures are now

$$\int_{s_i}^{s_{i+1}} N'(z_{i+1}-z) ds = \Delta s N_i' \Delta z \tag{18}$$

and

$$\int_{s_i}^{s_{i+1}} M'\cos\phi ds \approx \frac{\Delta s}{2}\left(M_i'\cos\phi_i + M_{i+1}'\cos\phi_{i+1}\right). \tag{19}$$

The expression 18 can be calculated as N_i' is known. In the approximation 19 the couple M_{i+1}' is unknown. For it can be written Equation

$$M_{i+1}' = \mu_{i+1} M_{i+1}' - \frac{(1-\mu_{i+1}\nu_{i+1})B_{i+1}'}{r_{i+1}}\psi_{i+1}\,. \tag{20}$$

The substitution of ψ_{i+1} (from Equation 9) into 20 and this further into Equation 19 will produce with rearranging the final formula

$$(1-b_{i+1})M^s_{i+1} = \frac{r_i}{r_{i+1}}\left[M^i_s + N^s_i a(1-\cos\Delta\phi) + Q^s_i a\sin\Delta\phi\right] - \frac{\Delta s}{2r_{i+1}} N^t_i \Delta z$$
$$+\frac{\Delta s}{2r_{i+1}} M^t_i \cos\phi_i - \frac{\Delta s}{2r_{i+1}} \frac{(1-\mu_{i+1}\nu_{i+1})B^t_{i+1}\cos\phi_{i+1}}{r_{i+1}}\left[\frac{u_i}{a} + \frac{w_i - w_{i+1}}{d_{i+1}\Delta s}\right. \quad (21)$$
$$\left. +\frac{(1+\mu_{i+1})\Delta s}{2d_{i+1}a}\cdot\frac{w_{i+1}}{a} - \frac{\Delta s}{2d_{i+1}}\cdot\frac{1}{a}\frac{N^s_{i+1}}{C^s_{i+1}}\right] + \text{influence of loading}$$

where d_{i+1} is given in Equation 9 and the shortened notation b_{i+1} is

$$b_{i+1} = \frac{\Delta s}{2d_{i+1}r_{i+1}}\left(\mu_{i+1} + \frac{\cos\phi_{i+1}}{2}\frac{\Delta s}{r_{i+1}}\frac{B^t_{i+1}}{B^s_{i+1}}\right). \quad (22)$$

6. FLOW OF THE CALCULATIONS

The equations of the two former sections are used to generate the transfer matrix between displacement and force vectors at grid points *i* and *i+1*. The displacement vector is given in Equation 9 and the force vector at point *i* is

$$f_i = \begin{Bmatrix} N^s_i \\ Q^s_s \\ M^s_i \end{Bmatrix} \quad (23)$$

The transfer matrix is a 6x6 - matrix. In Calculations matrix FIR (6,6) (first boundary) includes the dependence of vectors $\mathbf{v_i}$ and $\mathbf{f_i}$ on themselves and is thus a unit matrix **I**. The matrix SEC (6,6) (second boundary) includes the dependence of vectors $\mathbf{v_{i+1}}$ and $\mathbf{f_{i+1}}$ on vectors $\mathbf{v_i}$ and $\mathbf{f_i}$. It is calculated in a loop of program (Figure 4). Both matrices can have two additinal rows for the circumferential normal force and couple at points *i* and *i+1*.

The use of equations of former sections is given in the next flow chart shown in Figure 4.

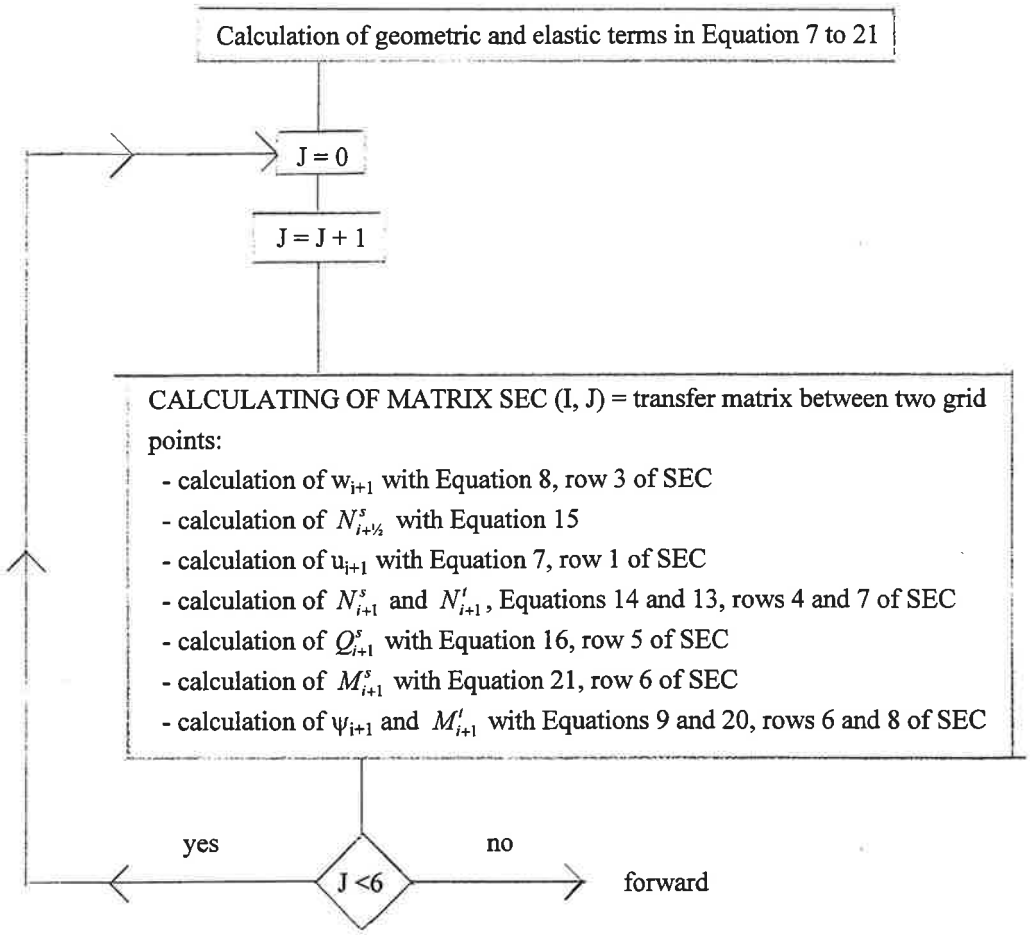


FIGURE 4: Flow of calculations for determining the transfer matrix SEC between grid points i and i+1.

The transfer matrix between two grid points was calculated. The transfer matrix of the whole structure is obtained by repeating n times calculations combined with a matrix multiplication. Results contain a discretization error.

To improve results the number of meridional segments is incresed and calculations are carried out several times (≤5). In increasing the number of meridional segments Bulirsch'

queue (n = 1, 2, 3, 4 and 6) [2, p. 109], is used as multipliers to limit the round-off error. After that an extrapolation toward zero length of Δs is used.

Due to extrapolation the transfer matrix (and loading vector) should be quite accurate. The stiffmess matrix and equivalent modal loads can now be calculated with good accuracy. The extrapolation is discussed in the next section.

7. EXTRAPOLATION

To decrease discretization error in calculated results Richardson type extrapolation [2, pp. 108-110] is carried out toward zero grid length. The basis of any extrapolation is some kind presumption about the nature of the error. Neville's algorithm [2, pp. 109-110] contains the form

$$e = \Sigma a_i \Delta s^i \quad , \quad i = 1,\ 2,\ 3,... \tag{23}$$

for the discretization error. In using Romberg's method [2, pp. 146-148] the error assumption is similar but includes only even powers of Δs.

In the phases of extrapolation terms of power series 23 are eliminated in turn beginning with the lowest power. When a person considers results of the first phase of Neville's algorithm, he can find the error to be approximately proportional to the second power of the grid length Δs. It is a very natural choice to use Romberg's method in the next phases of the extrapolation. In this study this simple idea was experimented. The powers in the series 23 are now i = 1, 2, 4, 6 and so on.

It is rather surprising that the modification of Neville's algorithm works giving after a four-phase extrapolation results with a rate of convergence eight for equations considered in this study. An exception seems to be found. It is discussed very briefly a little later. A condition for the convergence rate is the use of the trapezoidal rule in the numerical quadratures carried out. Integral 19 will lead to Equation 21. The corresponding equation for a ring plate [4, p. 36] was written after a discussion with associate professor Juha Paavola. The modified Neville's algorithm has given good results when the thickness of the structure has been linearly or parabolically varying. The rate of convergence has decreased when the thickness varys piecewise linearly.

Then the exception; When considering a spherical cap it is needed a routine of its own type for determining the transfer matrix between the polar point (number 0) and the point (number 1) with a distance Δs away from the pole. The matrix multiplication with the transfer matrices described in Sections 3 to 6, produces now matrices the convergence of which at least partly seems not to obey the presumption 23 in its modified form with powers $i = 1, 2, 4, 6, ...$

8. NUMERICAL EXAMPLES

In this section three examples are presented to verify the convergence rate of the method. In these there are used two error indicators. The first one

$$e = \max \frac{|k_{ij} - k_{ji}|}{\sqrt{k_{ii} k_{jj}}} , \tag{24}$$

is a measure for the asymmetry of stiffness matrix [5, p. 34]. In the expression k_{ij} is the element on row i and column j of stiffness matrix **K**.

Error parameter e is compared with the parameter

$$g = \max \frac{|k_{ij} - a_{ij}|}{\sqrt{a_{ii} a_{jj}}} \times sign(k_{ij} - a_{ij}) \tag{25}$$

for verifying its reliability as a measure of calculation error. In Equation 25 a_{ij} is the element corresponding k_{ij} in a stiffness matrix **A** calculated analytically or with a high accuracy that is with two to four correct digits more than matrix **K**.

Example 1 considers an isotropic ring plate with the inner radius 2.0 m and the outer radius 5.0 m. Thickness of the plate varys linearly and is 0.4 m at the inner and 0.20 m at the outer boundary. Young's modulus of and Poisson's ratio of the structure are 10 000 MN/m^2 and 1/3. The structure was calculated as a conical shell with a half apex angle of 90 degrees using five calculation loops. Data and error parameters are given in Table 1.

TABLE 1. Ring plate example. In columns are given number of meridional segments in the first and fifth loops and error parameters e and g. The bottom line contains proportions of values on two former lines.

Number of segments *n...6n*	**Error parameters**	
	e	***g***
18...108	$0.4452 \cdot 10^{-12}$	-
6...36	$0.1905 \cdot 10^{-8}$	$-0.9801 \cdot 10^{-8}$
8...48	$0.1987 \cdot 10^{-9}$	$-0.1075 \cdot 10^{-8}$
$(8/6)^8=9.99$	9.59	9.12

The second example is a slightly orthotropic cylindrical shell with the radius a = 10.0 m and lenght l = 2.0 m. Young's moduli are E^s = 16 000 MN/m^s and E^t = *10 000 MN/m^2*. Poisson's ratios are $\nu = 0.30$ and $\mu = 0.1875$. The thickness of the structure is constant h = 0.10 m. Results are given in Table 2.

TABLE 2. An orthotropic cylindrical shell. Content of the table is similar with the one in Example 1.

n...6n	***e***	***g***
32...142	$0.3352 \cdot 10^{-12}$	-
6...36	$0.3151 \cdot 10^{-6}$	$-0.3146 \cdot 10^{-8}$
12...72	$0.1188 \cdot 10^{-8}$	$-0.1186 \cdot 10^{-8}$
$(12/6)^8=256$	265.2	265.2

In two first examples results compared also with the analytical ones computed with ESAS-program of prof. Hannu Outinen [6] . Coincidence with them is excellent.

The third example structure is an isotropic spherical shell the radius of which is 10 m and thickness n= 0.10 m. Angles of boundaries are $\phi_1 = 30°$ and $\phi_{n+1} = 40°$. Elastic constants are E = 10 000 MN/m^2 and $\nu = 0.200$. Results are in Table 3.

TABLE 3. Results for an isotropic spherical shell. Content of the table is similar with the one in Example 1.

n...6n	***e***	***g***
30...180	$0.1237 \cdot 10^{-11}$	-
10...60	$0.2129 \cdot 10^{-8}$	$-0.2254 \cdot 10^{-8}$
6...36	$0.1273 \cdot 10^{-6}$	$-0.1364 \cdot 10^{-6}$
$(10/6)^8=59.54$	59.8	60.0

10. CONCLUSIONS

Finite difference approximations combined with numerical integration using the trapezoidal rule and with an extrapolation give a good rate of convergence in solutions of differential equations of this study.

The asymmetry of the stiffness matrix seems to be a realiable measure of the calculation error in generating shell elements. A more thorough consideration will show that for the stretching state of a ring plate the asymmetry fails totally as a measure of error.

As a numerical method the shooting method presented would be well applied in computing composite structures.

For cylindrical shells with constant thicknesses h one can found contours of error parameters e and g (Eq. 24 and 25). The calculation error is proportional to the term $l/\sqrt{ah}$, where l is the length and a the radius of the cylinder. If Geckeler's approximation is regarded valid in calculating shell structures, also the length of the meridional line of other shell types is limited, when a certain error level is tried to catch in computing. In one example carried out with a long cylindrical shell this limitation could be avoided by shooting from opposite boundaries of the shell and coupling the results at the line of symmetry. The method may be experimented with success also with asymmetric structures.

11. REFERENCES

1. K. Girkmann, *Flächentragwerke*, Sechste Amflage, Wien Springer-Verlag 1963.
2. M. Mäkelä, O. Nevanlinna and J. Virkkunen, *Numeerinen matematiikka,* toinen painos, Gaudeamus, Mänttä 1984.
3. P. Tuominen, *Generation of a Spherical Shell Element Using a Shooting Method,* Domes from Antiquity to the Present, Proceedings of IASS-MSU Symposium, Mimar Sinan University Istanbul, Turkey 1983 pp. 407-416.
4. P. Tuominen, *Generation of an Axisymmetric Ring Plate Element Using a Shooting Method*, Acta Polytechnica Scandinavia, Civil Engineering and Building Construction Series No. 100, Helsinki 1995.
5. H. Outinen, *Paksuudeltaan lineaarisesti muuttuvan ortotrooppisen kartiokuoren rotaatiosymmetrinen statiikka (Axisymmetric statics of orthotropic conical shell with tapered thickness),* Tampere University of Technology, Applied Mechanics, Report 38, Tampere 1987. 30 pp.
6. H. Outinen, *ESAS-Manuaali (ESAS- manual)*, Tampere University of Technology, Applied Mechanics, Report 43, Tampere 1987. 50 pp.

INFINITE ELEMENT STRIPS AND h–CONVERGENCE

J. AALTO and K. KUULA
University of Oulu
Department of Civil Engineering
PL 191
90101 Oulu
FINLAND

ABSTRACT

Infinite strips of mapped elements are proposed as a generalization of mapped infinite elements. Combined with standard isoparametric elements infinite element strips can be used to solve boundary value problems with unbounded domains effectively. Typical diffusion and plane elasticity problems are considered as numerical examples. Experimental convergence studies of the error in energy show that it is possible to achieve better rate of h-convergence, if infinite element strips instead of conventional infinite elements are used.

1. INTRODUCTION

Unbounded domains cause difficulties in standard finite element analysis. One possibility to overcome these difficulties is to cover the domain of the problem using standard finite elements and special infinite elements. Infinite elements connect the outer boundaries of the standard grid to the infinite boundaries of the domain. Mapped infinite elements first proposed by Zienkiewicz et. al. [1] are attractive and easy to implement.

The most simple combination of standard parametric and mapped infinite elements is obtained, if equal degree of interpolation of the basic unknown functions (in natural coordinates) is used in both element types. There is, however, a shortcoming in this attractive and widely used combination: The possibility to increase the accuracy of the analysis by making the grid denser is limited. The reason for this is simply that the number of degrees of freedom does not change and the approximation does not improve in the longitudinal direction of the infinite elements.

The paper presents a simple way to avoid this defect. The idea is just to use a mapped strip of elements (whose last element is infinite) instead of a single mapped infinite element between the finite element grid and the infinite boundary of the domain. Such strips of elements are called here *"infinite element strips"*.

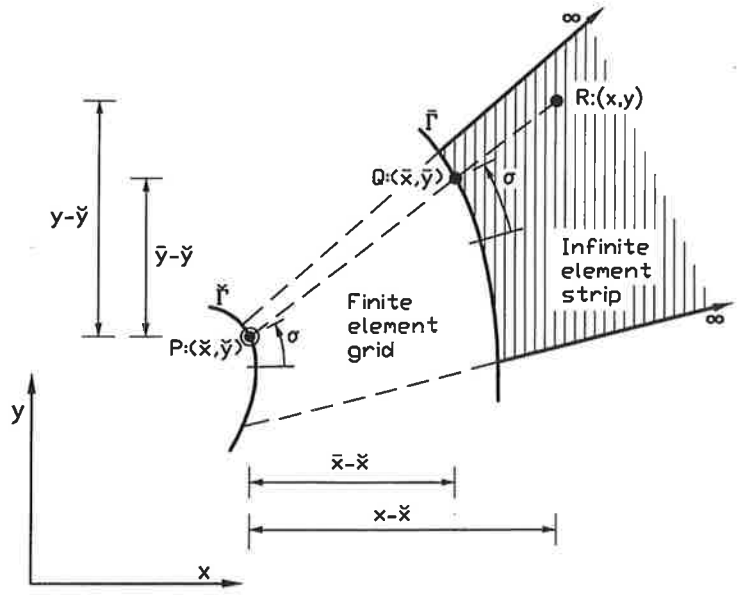


Fig. 1: Geometrical mapping of an infinite element strip

2. GEOMETRICAL MAPPING OF AN INFINITE ELEMENT STRIP

The geometry of an infinite element strip can be defined using two curves: so called pole line $\breve{\Gamma}$ and so called interface line $\bar{\Gamma}$. The latter connects the infinite element strip to the standard finite element grid. Pole line $\breve{\Gamma}$ and interface line $\bar{\Gamma}$ are expressed as

$$\begin{aligned} x &= \breve{x}(\sigma), \\ y &= \breve{y}(\sigma) \end{aligned} \tag{1}$$

and

$$\begin{aligned} x &= \bar{x}(\sigma), \\ y &= \bar{y}(\sigma), \end{aligned} \tag{2}$$

respectively, where σ is a parametric coordinate.

Coordinates x and y of a point R in the infinite element strip are defined by scaling the coordinate differences $\bar{x}-\breve{x}$ and $\bar{y}-\breve{y}$ (see Fig. 1) with a multiplier c, which may vary from 0 to ∞. Thus we have

$$\begin{aligned} x - \breve{x} &= c(\bar{x} - \breve{x}), \\ y - \breve{y} &= c(\bar{y} - \breve{y}). \end{aligned} \tag{3}$$

Choosing $c = 1/(1-\rho)$, where ρ is another parametric coordinate, we get

$$\begin{aligned} x(\rho,\sigma) &= \frac{1}{1-\rho}[\bar{x}(\sigma) - \rho\breve{x}(\sigma)], \\ y(\rho,\sigma) &= \frac{1}{1-\rho}[\bar{y}(\sigma) - \rho\breve{y}(\sigma)]. \end{aligned} \tag{4}$$

It is easy to see that $x(0,\sigma) = \bar{x}(\sigma)$, $y(0,\sigma) = \bar{y}(\sigma)$ and $x(1,\sigma) = y(1,\sigma) = \infty$. Thus if $\rho = 0$, we are on the interface line $\bar{\Gamma}$, and if $\rho = 1$, we are on the infinite boundary of the domain.

A natural way of expressing the coordinates $\breve{x}(\sigma)$, $\breve{y}(\sigma)$ and $\bar{x}(\sigma)$, $\bar{y}(\sigma)$ of the pole line $\breve{\Gamma}$ and the interface line $\bar{\Gamma}$, respectively, is to regard these lines as parametric line elements and use Lagrange interpolation. Thus we have

$$\begin{aligned} \breve{x}(\sigma) &= \sum_{i=1}^{m} L_i^m(\sigma)\breve{x}_i, \\ \breve{y}(\sigma) &= \sum_{i=1}^{m} L_i^m(\sigma)\breve{y}_i \end{aligned} \tag{5}$$

and

$$\begin{aligned} \bar{x}(\sigma) &= \sum_{i=1}^{m} L_i^m(\sigma)\bar{x}_i, \\ \bar{y}(\sigma) &= \sum_{i=1}^{m} L_i^m(\sigma)\bar{y}_i, \end{aligned} \tag{6}$$

where

$$L_i^m(\sigma) = \prod_{\substack{j=1 \\ j\neq i}}^{m} \frac{\sigma - \sigma_j}{\sigma_i - \sigma_j}, \tag{7}$$

m is the number of nodes of the line element and $\breve{x}_i$, $\breve{y}_i$ and $\bar{x}_i$, $\bar{y}_i$ are the correspoding nodal coordinates. Combining equations (4), (5) and (6) finally gives

$$\begin{aligned} x(\rho,\sigma) &= \sum_{i=1}^{m} \frac{1}{1-\rho}(\bar{x}_i - \rho\breve{x}_i)L_i^m(\sigma), \\ y(\rho,\sigma) &= \sum_{i=1}^{m} \frac{1}{1-\rho}(\bar{y}_i - \rho\breve{y}_i)L_i^m(\sigma). \end{aligned} \tag{8}$$

Equations (8) express the geometrical mapping of an infinite element strip.

3. MAPPED ELEMENTS IN AN INFINITE ELEMENT STRIP

The relations between the natural coordinates ξ and η of element e and the strip coordinates ρ and σ are assumed to be linear and of form

$$\begin{aligned} \rho &= \frac{\rho_1^e + \rho_2^e}{2} + \frac{\rho_2^e - \rho_1^e}{2}\xi, \\ \sigma &= \eta, \end{aligned} \tag{9}$$

(see Figs. 2b and c). With the help of equations (8) and (9) one thus obtains

$$\begin{aligned} x(\xi,\eta) &= \sum_{i=1}^{m} \frac{2\bar{x}_i - [\rho_1^e + \rho_2^e + (\rho_2^e - \rho_1^e)\xi]\breve{x}_i}{2 - \rho_1^e - \rho_2^e - (\rho_2^e - \rho_1^e)\xi} L_i^m(\eta), \\ y(\xi,\eta) &= \sum_{i=1}^{m} \frac{2\bar{y}_i - [\rho_1^e + \rho_2^e + (\rho_2^e - \rho_1^e)\xi]\breve{y}_i}{2 - \rho_1^e - \rho_2^e - (\rho_2^e - \rho_1^e)\xi} L_i^m(\eta). \end{aligned} \tag{10}$$

Equations (10) express the geometrical mapping of an element in an infinite element strip.

4. PROPERTIES OF MAPPED ELEMENTS

4.1 Properties along a radial line

The decaying solution $\mathbf{u}(x,y)$ of a boundary value problem, which should be approximated using the elements of an infinite element strip, can be expressed as an infinite series of form

$$\mathbf{u}(r,\theta) = \mathbf{u}_0(\theta) + \frac{\mathbf{u}_1(\theta)}{r} + \frac{\mathbf{u}_2(\theta)}{r^2} + \frac{\mathbf{u}_3(\theta)}{r^3} + \cdots, \tag{11}$$

where r and θ are polar coordinates so that r is measured from the pole P and $\mathbf{u}_i(\theta)$ are smooth functions of θ. Along a radial line $\theta = \theta_0$ it can be expressed as

$$\mathbf{u}(r,\theta_0) = \boldsymbol{\alpha}_0 + \frac{\boldsymbol{\alpha}_1}{r} + \frac{\boldsymbol{\alpha}_2}{r^2} + \frac{\boldsymbol{\alpha}_3}{r^3} + \cdots, \tag{12}$$

where $\boldsymbol{\alpha}_i = \mathbf{u}_i(\theta_0)$. A possibility to get a priori information of the quality of the elements of our infinite element strip is to study, how well they can approximate the function (12) along a radial line. This is done in the following.

The finite element approximation of function $\mathbf{u}(x,y)$ within element e of an infinite element strip is a polynomial (of the natural coordinates ξ and η) and of form

$$\hat{\mathbf{u}}(\xi,\eta) = \mathbf{a}_0 + \mathbf{a}_1\xi + \mathbf{a}_2\eta + \mathbf{a}_3\xi^2 + \mathbf{a}_4\xi\eta + \mathbf{a}_5\eta^2 + \cdots. \tag{13}$$

The number of terms in this polynomial depends on the element type. Using the linear relations (9) it can be written as

$$\hat{\mathbf{u}}(\rho,\sigma) = \mathbf{b}_0 + \mathbf{b}_1\rho + \mathbf{b}_2\sigma + \mathbf{b}_3\rho^2 + \mathbf{b}_4\rho\sigma + \mathbf{b}_5\sigma^2 + \cdots, \tag{14}$$

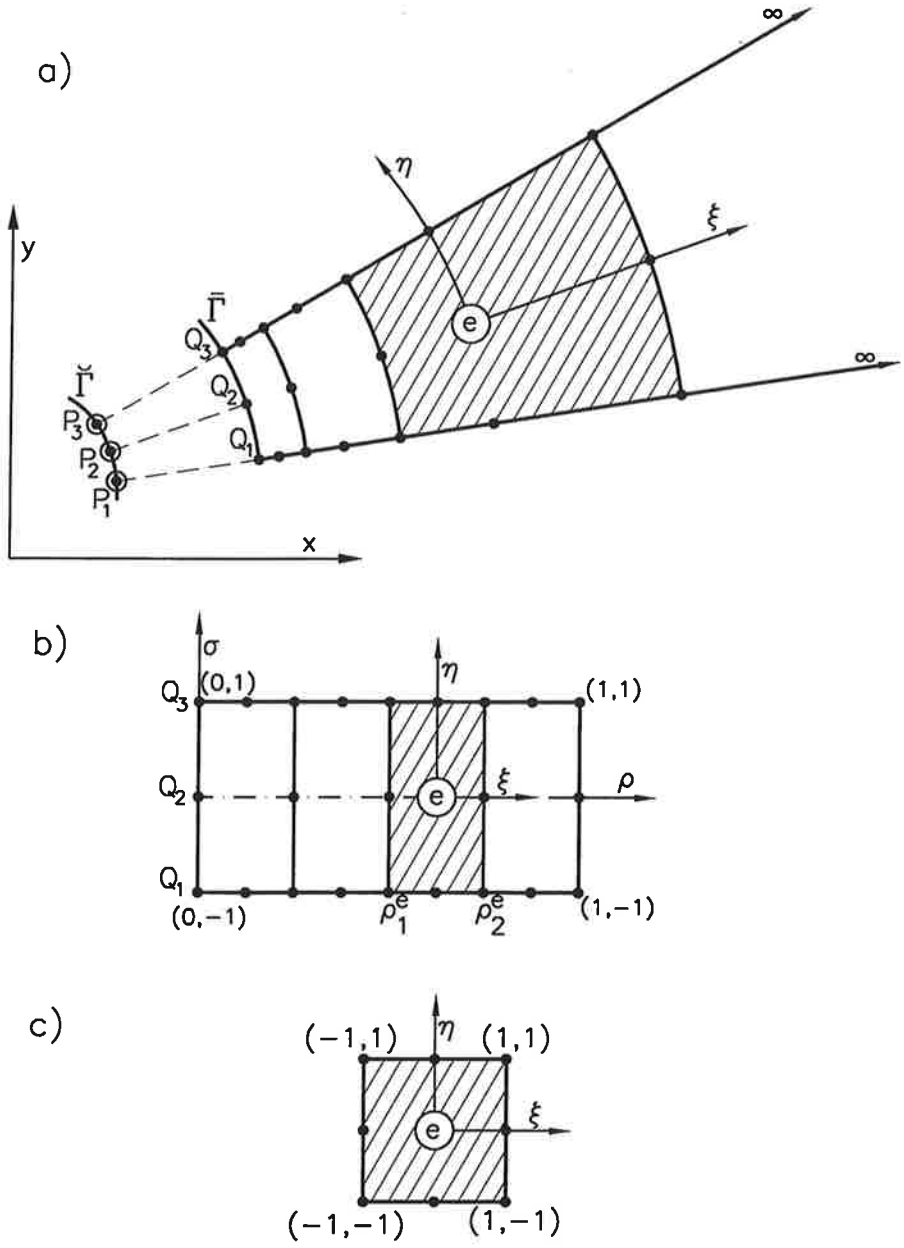


Fig. 2: A quadratic Serendip element in an infinite element strip:
(a) physical, (b) strip- and (c) natural coordinates

where

$$\begin{aligned}
\mathbf{b}_0 &= \mathbf{a}_0 - \frac{\rho_1^e + \rho_2^e}{\rho_2^e - \rho_1^e}\mathbf{a}_1 + (\frac{\rho_1^e + \rho_2^e}{\rho_2^e - \rho_1^e})^2\mathbf{a}_3, \\
\mathbf{b}_1 &= \frac{2}{\rho_2^e - \rho_1^e}\mathbf{a}_1 - \frac{4(\rho_1^e + \rho_2^e)}{(\rho_2^e - \rho_1^e)^2}\mathbf{a}_3, \\
\mathbf{b}_2 &= \mathbf{a}_2 - \frac{\rho_1^e + \rho_2^e}{\rho_2^e - \rho_1^e}\mathbf{a}_4, \\
\mathbf{b}_3 &= \frac{4}{(\rho_2^e - \rho_1^e)^2}\mathbf{a}_3, \\
\mathbf{b}_4 &= \frac{2}{\rho_2^e - \rho_1^e}\mathbf{a}_4, \\
\mathbf{b}_5 &= \mathbf{a}_5 \\
&, \cdots.
\end{aligned} \tag{15}$$

The distances of points R:(x, y) and Q:$(\bar{x}, \bar{y})$ from the pole P:$(\check{x}, \check{y})$ (which is considered to be fixed here) are

$$r = \sqrt{(x - \check{x})^2 + (y - \check{y})^2} \tag{16}$$

and

$$\bar{r} = \sqrt{(\bar{x} - \check{x})^2 + (\bar{y} - \check{y})^2}, \tag{17}$$

respectively. With the help of equations (4), (16) and (17) one easily gets the result

$$r = \frac{1}{1 - \rho}\bar{r}, \tag{18}$$

or

$$\rho = 1 - \frac{\bar{r}(\sigma)}{r}. \tag{19}$$

Substitution of this into equation (14) results to

$$\hat{\mathbf{u}}(r, \sigma) = \mathbf{b}_0 + \mathbf{b}_1 + \mathbf{b}_3 + (\mathbf{b}_2 + \mathbf{b}_4)\sigma + \mathbf{b}_5\sigma^2 - (\mathbf{b}_1 + \mathbf{b}_3 + \mathbf{b}_4\sigma)\bar{r}(\sigma)\frac{1}{r} + \mathbf{b}_3\bar{r}^2(\sigma)\frac{1}{r^2} + \cdots. \tag{19}$$

Consider now the approximation $\hat{\mathbf{u}}$ along a radial line. Let us denote the corresponding values of the coordinate σ and angle θ as σ_0 and θ_0, respectively. They are related by equation

$$\tan\theta_0 = \frac{\bar{y}(\sigma_0) - \check{y}}{\bar{x}(\sigma_0) - \check{x}}. \tag{20}$$

We now get

$$\hat{\mathbf{u}}(r, \sigma_0) = \mathbf{c}_0 + \frac{\mathbf{c}_1}{r} + \frac{\mathbf{c}_2}{r^2} + \cdots, \tag{21}$$

where

$$\begin{aligned}
\mathbf{c}_0 &= \mathbf{b}_0 + \mathbf{b}_1 + \mathbf{b}_3 + (\mathbf{b}_2 + \mathbf{b}_4)\sigma_0 + \mathbf{b}_5\sigma_0^2, \\
\mathbf{c}_1 &= -(\mathbf{b}_1 + \mathbf{b}_3 + \mathbf{b}_4\sigma_0)\bar{r}(\sigma_0), \\
\mathbf{c}_2 &= \mathbf{b}_3\bar{r}^2(\sigma_0), \\
&\cdots .
\end{aligned} \tag{22}$$

Comparison of equations (22) and (12) show now, that along a radial line the finite element approximation of a mapped element in an infinite element strip is able to reproduce certain first terms of the corresponding infinite series form of the analytical solution. If the element is bilinear these terms are $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1/r$ and if the element is quadratic (Serendip or Lagrange) these terms are $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1/r$ and $\boldsymbol{\alpha}_2/r^2$ etc.

4.2 Infinite element strip with one element

Consider an infinite element strip with one element. Now we have $\rho_1^e = 0$ and $\rho_2^e = 1$ and equations (9) and (10) give

$$\begin{aligned}
x(\xi,\eta) &= \sum_{i=1}^{m} \frac{2\bar{x}_i - (1+\xi)\breve{x}_i}{1-\xi} L_i^m(\eta), \\
y(\xi,\eta) &= \sum_{i=1}^{m} \frac{2\bar{y}_i - (1+\xi)\breve{y}_i}{1-\xi} L_i^m(\eta).
\end{aligned} \tag{23}$$

If the element is a quadratic Lagrange quadrilateral of Fig. 3, we have $m = 3$

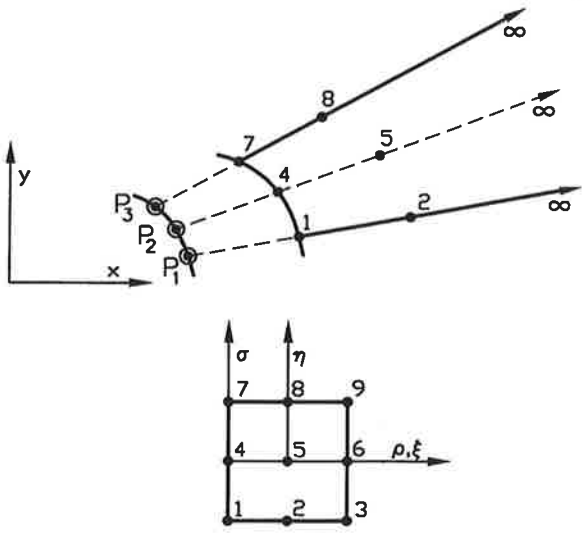


Fig. 3: An infinite element strip with one quadratic Lagrange quadrilateral

$$
\begin{aligned}
L_1^3(\eta) &= -\frac{1}{2}\eta(1-\eta), \\
L_2^3(\eta) &= 1-\eta^2, \\
L_3^3(\eta) &= \frac{1}{2}\eta(1+\eta),
\end{aligned}
\tag{24}
$$

and

$$
\begin{aligned}
\bar{x}_1 &= x_1, \quad \bar{y}_1 = y_1, \\
\bar{x}_2 &= x_4, \quad \bar{y}_2 = y_4, \\
\bar{x}_3 &= x_7, \quad \bar{y}_3 = y_7.
\end{aligned}
\tag{25}
$$

We get for the x-coordinates of the nodes at $\xi = 0$

$$
\begin{aligned}
x_2 &= x(0,-1) = 2\bar{x}_1 - \breve{x}_1 = 2x_1 - \breve{x}_1, \\
x_5 &= x(0,0) = 2\bar{x}_2 - \breve{x}_2 = 2x_4 - \breve{x}_2, \\
x_8 &= x(0,+1) = 2\bar{x}_3 - \breve{x}_3 = 2x_7 - \breve{x}_3.
\end{aligned}
\tag{26}
$$

Equations (26) can be solved for the x-coordinates $\breve{x}_i$ of the poles. The result is

$$
\begin{aligned}
\breve{x}_1 &= 2x_1 - x_2, \\
\breve{x}_2 &= 2x_4 - x_5, \\
\breve{x}_3 &= 2x_7 - x_8.
\end{aligned}
\tag{27}
$$

Similar equations result for the y-coordinates of the poles. Substitution of these into equations (23) gives

$$
\begin{aligned}
x(\xi,\eta) &= [(2x_1 - x_2)N_0(\xi) + x_2 N_2(\xi)]L_1^3(\eta) \\
&\quad + [(2x_4 - x_5)N_0(\xi) + x_5 N_2(\xi)]L_2^3(\eta) \\
&\quad + [(2x_7 - x_8)N_0(\xi) + x_8 N_2(\xi)]L_3^3(\eta), \\
y(\xi,\eta) &= [(2y_1 - y_2)N_0(\xi) + y_2 N_2(\xi)]L_1^3(\eta) \\
&\quad + [(2y_4 - y_5)N_0(\xi) + y_5 N_2(\xi)]L_2^3(\eta) \\
&\quad + [(2y_7 - y_8)N_0(\xi) + y_8 N_2(\xi)]L_3^3(\eta),
\end{aligned}
\tag{28}
$$

where

$$
\begin{aligned}
N_0(\xi) &= -\frac{\xi}{1-\xi}, \\
N_2(\xi) &= \frac{1}{1-\xi}.
\end{aligned}
\tag{29}
$$

The geometrical mapping of equations (28) is identical to that of the mapped infinite element of reference [1].

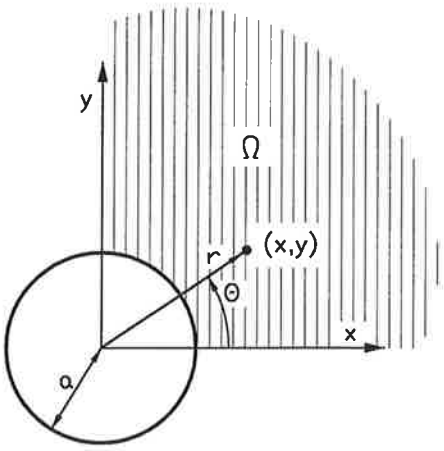


Fig. 4: Notation for both the flow problem and the elasticity problem

5. NUMERICAL EXAMPLES

5.1 Example problems

Potential flow around a circular cylinder. This problem is referred as "flow problem" in the following. Consider irrotational flow of an ideal fluid around a circular cylinder of radius a. The velocities far ($r \to \infty$) from the cylinder (see Fig. 4) are $v_x = v_\infty$ and $v_y = 0$. Expressed in terms of the velocity potential $\phi(x, y)$, which is related to the velocities $v_x(x, y)$ and $v_y(x, y)$ by equations

$$v_x = -\frac{\partial \phi}{\partial x}, \qquad v_y = -\frac{\partial \phi}{\partial y}, \tag{30}$$

this problem is governed by the Laplace equation. The Neumann boundary condition on the surface of the cylinder is

$$v_{\mathrm{n}} \equiv -\frac{\partial \phi}{\partial \mathrm{n}} = 0. \tag{31}$$

Because of symmetry we can restrict our consideration to the first quadrant of the x, y-plane and assume the Dirichlet and Neumann boundary conditions $v_y(x, 0) = 0$ and $\phi(0, y) = 0$ on the x- and y-axes, respectively.

The potential ϕ will be infinite on the infinite boundaries of the domain and the corresponding nodal values should be infinite too. Thus ϕ as such cannot be used as an unknown in the numerical analysis. We will take the potential difference

$$\phi^* = \phi - \phi_0, \tag{32}$$

where ϕ_0 is the potential corresponding to uniform flow with velocities $v_x = v_\infty$ and $v_y = 0$, as a new unknown of the problem. It is easy to see, that the potential ϕ_0 is

$$\phi_0 = -v_\infty x. \tag{33}$$

Far from the cylinder ($r \to \infty$) the influence of the cylinder vanishes and $\phi = \phi_0$. Thus on infinte boundaries of the domain the potental difference ϕ^* is finite and has the value $\phi^* = 0$.

The modified boundary value problem, which is expressed in terms of the potential difference $\phi^*(x, y)$, is governed by the Laplace equation and the symmetry boundary conditions, but the boundary condition (31) on the surface of the cylinder is changed. With the help of equations (32) and (33) it can be written as

$$v_{\mathrm{n}}^* \equiv -\frac{\partial \phi^*}{\partial \mathrm{n}} = -\mathrm{n}_x v_\infty, \tag{34}$$

where the superscript $*$ refers to the modified boundary value problem and $\mathrm{n}_x \equiv \cos\theta$ is the x-component of the unit normal of the surface of the cylinder. The equation $\phi^* = 0$ on the infinite boundaries of the domain can further be taken as an additional Dirichlet boundary condition to the modified boundary value problem.

Relative energy norm of the error of the finite element solution $\hat{\phi}^*$ is

$$\eta_E = \frac{\|\phi^* - \hat{\phi}^*\|_E}{\|\phi^*\|_E}, \tag{35}$$

where

$$\|\phi\|_E = \int_{\hat{\Omega}} (v_x^2 + v_y^2) d\Omega \tag{36}$$

and $\hat{\Omega}$ is the mesh of finite elements and infinite element strips. The relative norm η_E is used as an error measure in the experimental convergence studies of the flow problem.

The analytical solution of the flow problem (see Fig. 4), which is used in the error analysis, is

$$\phi = -v_\infty (r + \frac{a^2}{r}) \cos\theta. \tag{37}$$

Stretching of an infinite plate with a circular hole. This problem is referred as "elasticity problem" in the following. Consider an infinite elastic plate with a circular hole of radius a. The state of stress far ($r \to \infty$) from the hole (see Fig. 4) is assumed to be $\sigma_x = \sigma_\infty$, $\sigma_y = 0$ and $\tau_{xy} = 0$. Expressed in terms of the displacements $u(x, y)$ and $v(x, y)$ this problem is governed by the Navier equations of plane elasticity. The Neumann boundary conditions on the surface of the hole are

$$t_x \equiv n_x \sigma_x + n_y \tau_{xy} = 0, \quad t_y \equiv n_x \tau_{xy} + n_y \sigma_y = 0, \tag{38}$$

Because of symmetry we can restrict our consideration to the first quadrant of the x, y-plane and assume the mixed boundary conditions $t_x(x,0) = v(x,0) = 0$ and $u(0,y) = t_y(0,y) = 0$ on the x- and y-axes, respectively.

The displacements u and v will be infinite on the infinite boundaries of the domain and the corresponding nodal values should be infinite too. Thus u and v as such cannot be used as unknowns in the numerical analysis. We will take the displacement differences

$$u^* = u - u_0, \quad v^* = v - v_0, \tag{39}$$

where u_0 and v_0 are the displacements corresponding to uniform state of stress $\sigma_x = \sigma_\infty$, $\sigma_y = 0$ and $\tau_{xy} = 0$, as new unknowns of the problem. It is easy to see that the displacements u_0 and v_0 are

$$u_0 = \frac{\sigma_\infty}{E}x, \quad v_0 = -\frac{\nu\sigma_\infty}{E}y. \tag{40}$$

Far form the hole ($r \to \infty$) the influence of the hole vanishes and $u = u_0$ and $v = v_0$. Thus on infinte boundaries of the domain the displacement differences u^* and v^* are finite and have the values $u^* = 0$ and $v^* = 0$.

The modified boundary value problem, which is expressed in terms of the displacement differences $u^*(x,y)$ and $v^*(x,y)$, is governed by the Navier equations and the symmetry boundary conditions, but the boundary conditions on the surface of the hole are changed. They can easily be written as

$$t_x^* \equiv n_x\sigma_x^* + n_y\tau_{xy}^* = -n_x\sigma_\infty, \quad t_y^* \equiv n_x\tau_{xy}^* + n_y\sigma_y^* = 0, \tag{41}$$

where the superscript $*$ refers to the quantitites of the modified boundary value problem. The equations $u^* = 0$ and $v^* = 0$ on the infinite boundaries of the domain can further be taken as additional Dirichlet boundary conditions to the modified boundary value problem.

Relative energy norm of the error of the finite element solution $\hat{\mathbf{u}}^* = [\hat{u}^*, \hat{v}^*]^T$ is

$$\eta_E = \frac{\|\mathbf{u}^* - \hat{\mathbf{u}}^*\|_E}{\|\mathbf{u}^*\|_E}, \tag{42}$$

where

$$\|\mathbf{u}\|_E = \int_{\hat{\Omega}} \epsilon^T \mathbf{D} \epsilon d\Omega, \tag{43}$$

$$\epsilon = \begin{Bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{Bmatrix}, \quad \mathbf{D} = \frac{E}{1-\nu^2}\begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{bmatrix} \tag{44}$$

and $\hat{\Omega}$ is the mesh of finite elements and infinite element strips. The relative norm η_E is used as an error measure in the experimental convergence studies of the elasticity problem.

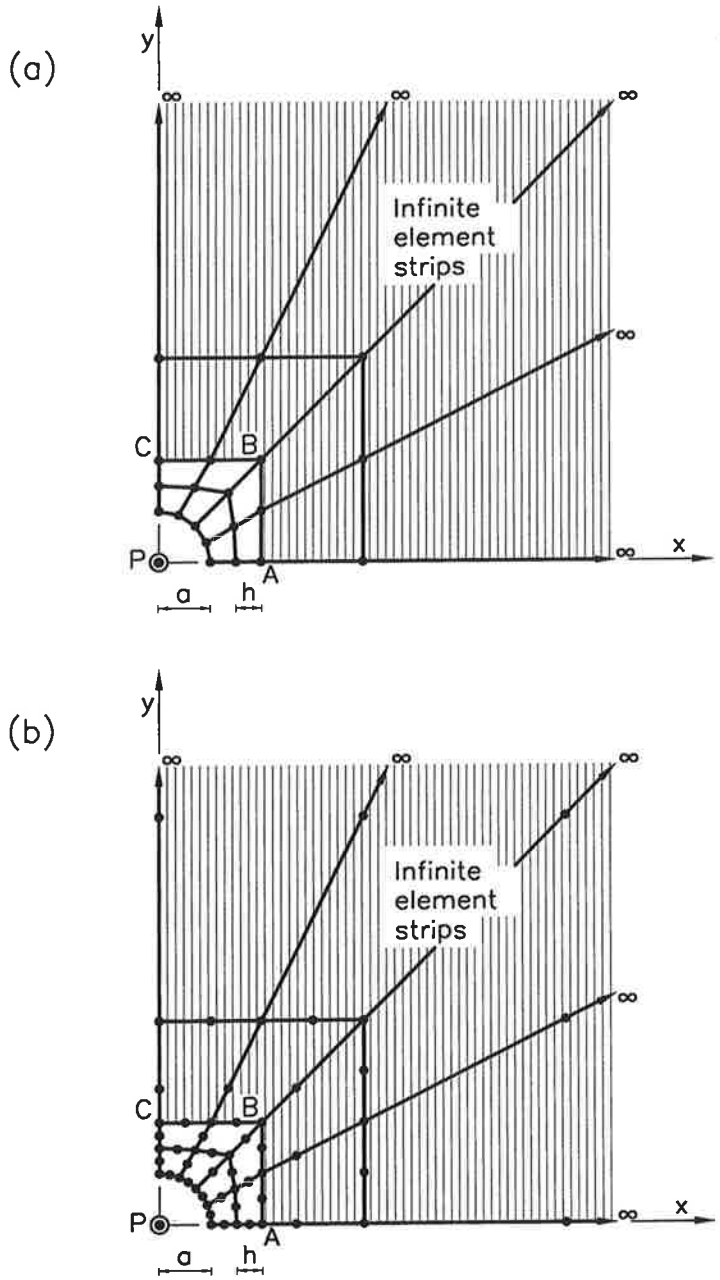


Fig. 5: Typical grids ($h/a = 0.5$) with infinite element strips:
(a) bilinear elements (b) quadratic Serendip elements

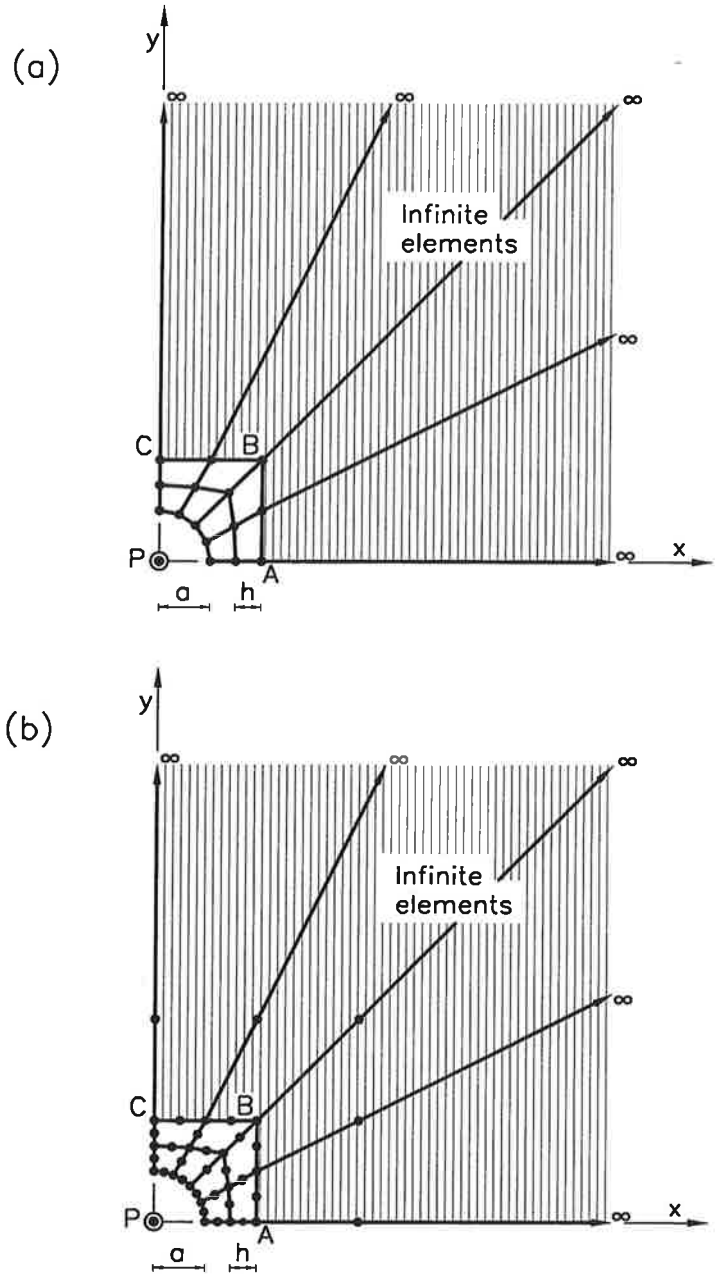


Fig. 6: Typical grids ($h/a = 0.5$) with mapped infinite elements of reference [1]:
(a) bilinear elements (b) quadratic Serendip elements

The analytical solution of the elasticity problem (see Fig. 4), which is used in the error analysis, is

$$u = \frac{\sigma_\infty a}{E}\left\{\frac{r}{a}\cos\theta + \frac{a}{r}\left[2\cos\theta + \frac{1+\nu}{2}\cos 3\theta\right] - \frac{1+\nu}{2}\frac{a^3}{r^3}\cos 3\theta\right\},$$
$$v = \frac{\sigma_\infty a}{E}\left\{-\nu\frac{r}{a}\sin\theta + \frac{a}{r}\left[-(1-\nu)\sin\theta + \frac{1+\nu}{2}\sin 3\theta\right] - \frac{1+\nu}{2}\frac{a^3}{r^3}\sin 3\theta\right\}. \tag{45}$$

5.2 Grids

Both example problems (the flow problem and the elasticity problem) were analyzed using identical grids. The pole P of all the infinite element strips is located at the centre of the circle (cylinder/hole). The interface line $\bar{\Gamma}$ consists of lines AB and BC in Fig. 5. Five uniformly refined grids with $h/a = 1, 0.5, 0.25, 0.125, 0.0625$ were constructed for experimental convergence studies. Typical grids ($h/a = 0.5$) of bilinear and quadratic Serendip elements, which are composed of standard finite elements and infinite element strips are shown in Fig. 5. Similar grids ($h/a = 0.5$), which are composed of standard finite elements and mapped infinite elements of reference [1] are shown in Fig. 6.

5.3 Experimental convergence study

Fig. 7 presents results of experimental convergence study of the flow problem using bilinear and quadratic Serendip elements. The analysis was first performed using grids with infinite element strips (see Fig. 5). The experimental rates of convergence seem to approach the values 1 and 2 for the bilinear and quadratic elements, respectively. These values 1 and 2 are the ideal values, which should be approached using standard finite elements in bounded domains (with no singularities). The analysis was repeated using grids with mapped infinite elements of reference [1] (see Fig. 6). Practically identical results were obtained.

This result might raise the question: Do infinite element strips bring any improvement compared to mapped infinite elements? The coincidence of the results can, however, be shown to be caused by the simplicity of the anlytical solution (37). Along a radial line $\theta = \theta_0$ the analytical potential difference ϕ^* is of form

$$\phi^*(r, \theta_0) = \frac{\alpha_1}{r}, \tag{46}$$

where

$$\alpha_1 = -v_\infty a^2 \cos\theta_0. \tag{47}$$

and it does not contain higher degree terms of $1/r$. Based on the reasoning in section 4.2, a bilinear element (and also a quadratic Serendip element) in an infinite element strip is able to reproduce this result. Thus one element in an infinite element strip (a mapped infinite element) is needed and additional elements do not improve the result.

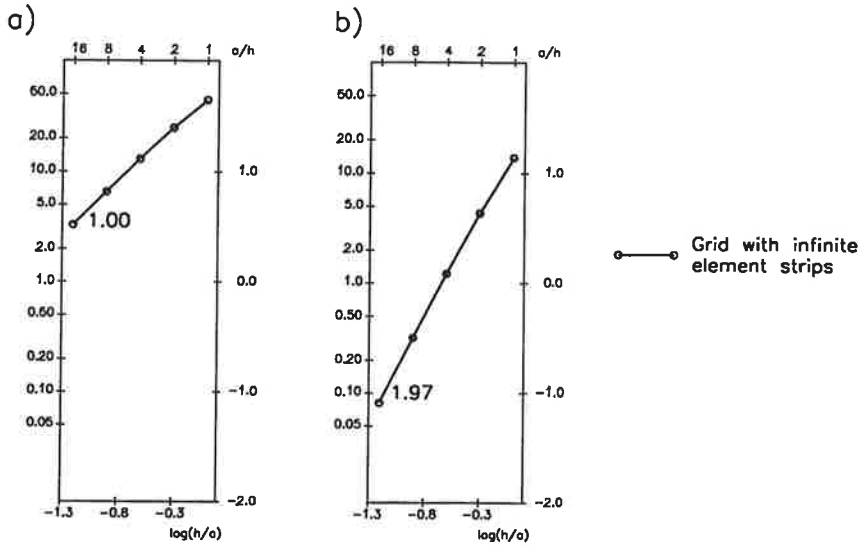


Fig. 7: Experimental convergence study of the flow problem:
(a) bilinear elements (b) quadratic Serendip elements

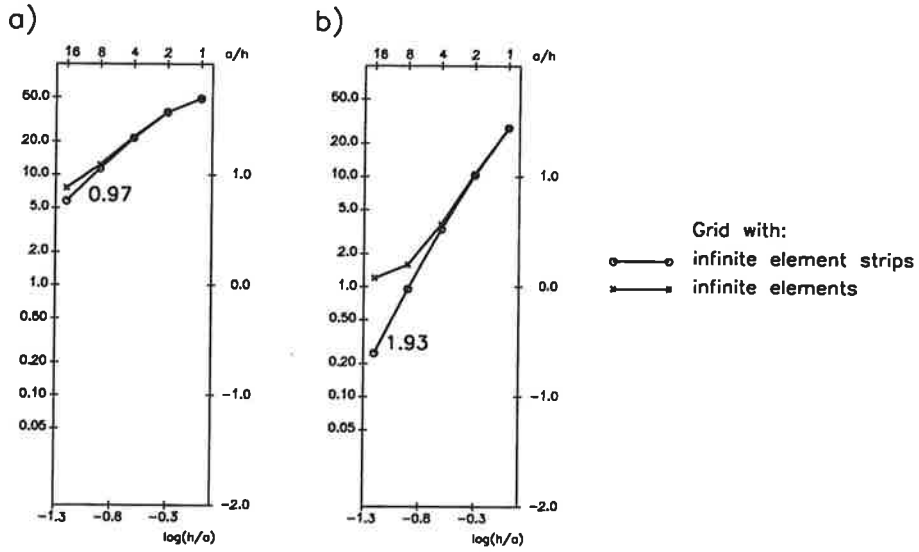


Fig. 8: Experimental convergence study of the elasticity problem:
(a) bilinear elements (b) quadratic Serendip elements

Fig. 8 presents results of experimental convergence study of the elasticity problem using bilinear and quadratic Serendip elements. The analysis was first performed using grids with infinite element strips (see Fig. 5). The experimental rates of convergence seem to approach the ideal values 1 and 2 for the bilinear and quadratic elements, respectively. The analysis was repeated using grids with mapped infinite elements of reference [1] (see Fig. 6). These results are not any more identical and they do not seem to converge.

Infinite element strips did bring a considerable improvement compared to mapped infinite elements. The analytical solution (45) was complicated enough to reveal this. Along a radial line $\theta = \theta_0$ the analytical displacement difference $\mathbf{u}^* = [u^*, v^*]^T$ is of form

$$\mathbf{u}^*(r, \theta_0) = \frac{\boldsymbol{\alpha}_1}{r} + \frac{\boldsymbol{\alpha}_3}{r^3}, \tag{48}$$

where

$$\boldsymbol{\alpha}_1 = \frac{\sigma_\infty a^2}{E} \left\{ \begin{array}{c} 2\cos\theta_0 + \frac{1+\nu}{2}\cos 3\theta_0 \\ -(1-\nu)\sin\theta_0 + \frac{1+\nu}{2}\sin 3\theta_0 \end{array} \right\} \tag{49}$$

and

$$\boldsymbol{\alpha}_3 = -\frac{\sigma_\infty (1+\nu) a^4}{2E} \left\{ \begin{array}{c} \cos 3\theta_0 \\ \sin 3\theta_0 \end{array} \right\}. \tag{50}$$

Based on the reasoning in section 4.2 a quadratic element (and also a bilinear element) in an infinite element strip is not able to reproduce this result. Thus the result is worst with one element in an infinite element srip (a mapped infinite element) and it improves with increasing number of elements.

6. CONCLUSIONS

The paper presented a systematic procedure of constructing mapped strip of elements, which can be used instead of mapped infinite elements between the finite element grid and the infinite boundary of the domain. Based on experimental convergence studies and some theoretical reasoning the paper also showed, that usage of such *"infinite element strips"* improves the accuracy of the analysis of unbounded domains compared to mapped infinite elements. Only two-dimensional problems were considered but generalization to three-dimensions is straightforward.

REFERENCES

1. O.C. Zienkiewicz, C. Emson and P. Bettess, A novel boundary infinite element, *Int. J. Num. Meth. Engng.* **19**, 393–404, (1983).

Optimal Dynamic Absorber for a Rotating Rayleigh Beam

MARKO JORKAMA
Valmet Winders, Wärtsilänkatu 100, FIN-04400 Järvenpää, Finland
AND
RAIMO VON HERTZEN
Laboratory of Theoretical and Applied Mechanics, Helsinki University of Technology, 02150 Espoo, Finland

ABSTRACT

A uniform rotating Rayleigh beam, carrying translational springs and dynamic vibration absorbers at its ends, is used as a basic model for a flexible rotor with dynamic vibration absorbers attached to the bearing houses. A closed form analytic solution for an arbitrary periodic load is derived. The necessity of taking into account the rotational coupling in the determination of the optimal tuning parameters is demonstrated by studying a numerical example of a paper machine roll. The values of the optimal absorber damping and spring constant are calculated as a function of the rotational speed of the roll. Finally, the effectiveness of the absorber is analyzed as a function of the absorber size.

1. INTRODUCTION

The *dynamic vibration absorber* or simply *dynamic absorber* was invented in the beginning of the 19th century by Frahm [1], and since then, it has proven to be an indispensable device to reduce the undesirable vibration in many applications such as gas turbines and engines, ship rolling, helicopters, electrical transmission lines etc. This discrete dynamic absorber was first analyzed by Ormondroyd and Den Hartog in 1928 [2], and the optimum damping was later derived by Brock [3]. Their studies covered a main system consisting of a mass and spring and a dynamic absorber with a mass, spring and viscous damper. For this system it was possible to obtain analytical expressions for the optimum tuning and damping of the absorber. Later in 1981 Thompson [4] extended the study to a viscously and hysteretically damped main system. He presented a numerical method for the determination of the optimal tuning and damping.

Young [5] was the first to consider the application of dynamic absorbers to beams in 1952. Snowdon [6] considered the optimization of the discrete absorber on beams with various boundary conditions when structural damping was present. Jacquot [7] used an approximate method in which the analogy established between a beam and a SDOF system allows the use of the optimum absorber parameters for the latter to determine the ones for the beam. The main system damping was not included in his theory so that the analytical resuts of Den Hartog and Brock could be applied. H. N Özgüven and B. Candir [8] extended Jacquot's treatment by considering a hysteretically damped beam with two dynamic absorbers for suppressing the first two resonances of the beam. A further extension was made by D. N Manikanahally and M. J. Crocker [9] who included mounted rigid masses in their beam model. The previous works, however, do not account for the rotational motion of the beam. This is an important factor, escpecially in high speed machinery.

In this paper a general closed form solution for a rotating uniform Rayleigh beam, with dynamic vibration absorbers, is presented. An example of a rotating paper machine roll with translational springs, dampers and dynamic absorbers at its ends is studied. The present theory can be utilized in suppressing e.g. nip induced vibrations in paper machinery.

2. THEORY

The equations of motion of a uniform rotating Rayleigh beam (see Fig.1) can be written in the inertial coordinates as

$$\rho A\ddot{u} - \rho I\ddot{u}'' + 2\rho I\Omega\dot{v}'' + C_i(\dot{u} - \Omega v) + EIu'''' = f_u, \tag{1}$$

$$\rho A\ddot{v} - \rho I\ddot{v}'' - 2\rho I\Omega\dot{u}'' + C_i(\dot{v} + \Omega u) + EIv'''' = f_v, \tag{2}$$

where $u(Z,t)$ and $v(Z,t)$ are the horizontal and vertical displacement fields, respectively, and $f_u(Z,t)$ and $f_v(Z,t)$ are the components of an arbitrary time periodic load. The density, cross-sectional area, moment of area, modulus of elasticity and rotational speed for the beam are ρ, A, I, E and Ω, respectively. The internal damping of the beam, proportional to the vibration velocity of the beam relative to the rotating coordinate system, is described by the linear viscous damping coefficient C_i and the length of the beam is L. The spring constants and viscous damping coefficients at the ends of the beam in the horizontal and vertical directions are K, C, $\hat{K}$ and $\hat{C}$, respectively.

When equations (1) and (2) are Fourier transformed with respect to time, a pair of ordinary differential equations for Z is obtained:

$$\hat{u}'''' + \frac{\rho}{E}\omega^2\hat{u}'' + i2\frac{\rho}{E}\Omega\omega\hat{v}'' - \frac{\rho A}{EI}\omega^2\hat{u} + i\frac{C_i}{EI}\omega\hat{u} - \frac{C_i}{EI}\Omega\hat{v} = \frac{1}{EI}\hat{f}_u, \tag{3}$$

$$\hat{v}'''' + \frac{\rho}{E}\omega^2\hat{v}'' - i2\frac{\rho}{E}\Omega\omega\hat{u}'' - \frac{\rho A}{EI}\omega^2\hat{v} + i\frac{C_i}{EI}\omega\hat{v} + \frac{C_i}{EI}\Omega\hat{u} = \frac{1}{EI}\hat{f}_v. \tag{4}$$

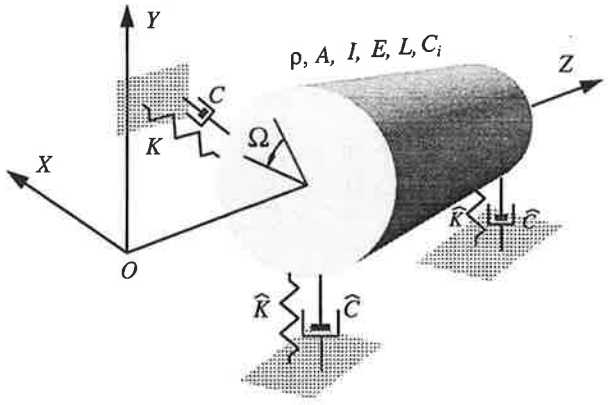


Figure 1. Spinning Rayleigh beam resting on springs and viscous dampers.

The general solution of equations (3) and (4) can be shown to be

$$\hat{u}(Z) = \sum_{n=1}^{4} \{[A_n^- + g_n^-(Z)]\Phi_n^-(Z) + [A_n^+ + g_n^+(Z)]\Phi_n^+(Z)\}, \tag{5}$$

$$\hat{v}(Z) = i\sum_{n=1}^{4} \{[A_n^- + g_n^-(Z)]\Phi_n^-(Z) - [A_n^+ + g_n^+(Z)]\Phi_n^+(Z)\}, \tag{6}$$

where the complete set of basis functions are given by

$$\Phi_1^{\pm}(Z) = \sin \nu^{\pm} Z,\ \Phi_2^{\pm}(Z) = \cos \nu^{\pm} Z,\ \Phi_3^{\pm}(Z) = \sinh \kappa^{\pm} Z,\ \Phi_4^{\pm}(Z) = \cosh \kappa^{\pm} Z, \tag{7}$$

the g-functions accounting for the load by

$$g_1^{\pm}(Z) = -\int_0^Z \frac{\hat{f}_u(z) \pm i\hat{f}_v(z)}{2EI\nu^{\pm}(\nu^{\pm 2} + \kappa^{\pm 2})} \cos \nu^{\pm} z\, dz, \tag{8}$$

$$g_2^{\pm}(Z) = \int_0^Z \frac{\hat{f}_u(z) \pm i\hat{f}_v(z)}{2EI\nu^{\pm}(\nu^{\pm 2} + \kappa^{\pm 2})} \sin \nu^{\pm} z\, dz, \tag{9}$$

$$g_3^{\pm}(Z) = \int_0^Z \frac{\hat{f}_u(z) \pm i\hat{f}_v(z)}{2EI\kappa^{\pm}(\nu^{\pm 2} + \kappa^{\pm 2})} \cosh \kappa^{\pm} z\, dz, \tag{10}$$

$$g_4^{\pm}(Z) = -\int_0^Z \frac{\hat{f}_u(z) \pm i\hat{f}_v(z)}{2EI\kappa^{\pm}(\nu^{\pm 2} + \kappa^{\pm 2})} \sinh \kappa^{\pm} z\, dz, \tag{11}$$

and

$$\nu^{\pm} = \sqrt{\frac{1}{2}\left\{\frac{\rho}{E}\omega(\omega \pm 2\Omega) + \sqrt{[\frac{\rho}{E}\omega(\omega \pm 2\Omega)]^2 + 4\frac{\rho A}{EI}\omega^2 - i4\frac{C_i}{EI}(\omega \pm \Omega)}\right\}}, \tag{12}$$

$$\kappa^{\pm} = \sqrt{\frac{1}{2}\left\{-\frac{\rho}{E}\omega(\omega \pm 2\Omega) + \sqrt{[\frac{\rho}{E}\omega(\omega \pm 2\Omega)]^2 + 4\frac{\rho A}{EI}\omega^2 - i4\frac{C_i}{EI}(\omega \pm \Omega)}\right\}}. \tag{13}$$

The boundary conditions of the problem are determined by the springs and dampers at the beam ends (see Fig.1) and two identical dynamic vibration absorbers which here are supposed to execute vertical motion at the beam ends as well. The Fourier transformed boundary conditions (14)-(19) and equations of motion for the absorbers (20) and (21) can be shown to be

$$EI\hat{u}'''(0) + \rho I\omega^2\hat{u}'(0) + i2\rho I\Omega\omega\hat{v}'(0) + (K + iC\omega)\hat{u}(0) = 0, \tag{14}$$

$$EI\hat{v}'''(0) + \rho I\omega^2\hat{v}'(0) - i2\rho I\Omega\omega\hat{u}'(0) + (\bar{K} + i\bar{C}\omega)\hat{v}(0) = (k_a + i\omega c_a)\hat{V}_0, \tag{15}$$

$$\hat{u}''(0) = 0,\ \hat{v}''(0) = 0, \tag{16}$$

$$EI\hat{u}'''(L) + \rho I\omega^2\hat{u}'(L) + i2\rho I\Omega\omega\hat{v}'(L) - (K + iC\omega)\hat{u}(L) = 0, \tag{17}$$

$$-EI\hat{v}'''(L) - \rho I\omega^2\hat{v}'(L) + i2\rho I\Omega\omega\hat{u}'(L) + (\bar{K} + i\bar{C}\omega)\hat{v}(L) = (k_a + i\omega c_a)\hat{V}_L, \tag{18}$$

$$\hat{u}''(L) = 0,\ \hat{v}''(L) = 0, \tag{19}$$

$$(k_a + i\omega c_a - m_a\omega^2)\hat{V}_0 = (k_a + i\omega c_a)\hat{v}(0), \tag{20}$$

$$(k_a + i\omega c_a - m_a\omega^2)\hat{V}_L = (k_a + i\omega c_a)\hat{v}(L), \tag{21}$$

where

$$\bar{K} = \hat{K} + k_a, \tag{22}$$

$$\bar{C} = \hat{C} + c_a. \tag{23}$$

Above m_a, c_a, k_a, $\hat{V}_0$ and $\hat{V}_L$ are the mass, viscous damping constant, spring constant and Fourier transforms of the displacements of the absorbers at locations 0 and L,

respectively. The coeffiecients $A_n^{\pm}$ $(n = 1, ..., 4)$, $\hat{V}_0$ and $\hat{V}_L$ can be solved by substituting the expressions (5) and (6) into the equations (14)-(21).

Finally, if T is the period of the loading functions f_u and f_v in equations (1) and (2), then the complete solution of the problem is

$$u(Z,t) = \sum_{n=-\infty}^{n=\infty} \hat{u}_n(Z) e^{in\omega t}, \tag{24}$$

$$v(Z,t) = \sum_{n=-\infty}^{n=\infty} \hat{v}_n(Z) e^{in\omega t}, \tag{25}$$

where

$$\omega = \frac{2\pi}{T} \tag{26}$$

and the functions $\hat{u}_n(Z)$ and $\hat{v}_n(Z)$ are obtained from equations (5)-(21) with $n\omega$ in place of ω.

3. APPLICATION TO A PAPER MACHINE ROLL

In order to illustrate the dependence of the optimal parameters of the dynamic absorber on the rotational speed of the beam Ω, the present theory is applied to a paper machine roll. We specify a uniform vertical load $\hat{f}_v = 1$, $\hat{f}_u = 0$ on the roll and calculate the vertical response $\hat{v}_1(L/2, \omega)$, i.e, the frequency response at the center of the roll. The optimization criterion used, for a fixed absorber mass m_a, to find the optimal values of c_a and k_a is

$$\min_{c_a, k_a} \{\max_{\omega} \hat{v}_1(L/2, \omega)\}. \tag{27}$$

The parameter values used in the calculations are shown in Table 1.

TABLE 1

Structural parameters used in the example

Parameter	Notation	Value
Modulus of elasticity	E	$2.106 \cdot 10^{11}$ N/m
Cross sectional area	A	0.1257 m^2
Roll density	ρ	7830 kg/m^3
Area moment of inertia	I	$1.010 \cdot 10^{-2}$ m^4
Horizontal bearing stiffness	K	$5.5 \cdot 10^8$ N/m
Vertical bearing stiffness	$\hat{K}$	$6.0 \cdot 10^8$ N/m
Horizontal and vertical bearing viscous damping coeffs.	C, $\hat{C}$	$3.94 \cdot 10^4$ Ns/m
Internal damping coefficient for the roll tube	C_i	807.65 Ns/m^2
Absorber mass	m_a	300 kg

The frequency response function in the neighbourhood of the lowest resonances is plotted in Fig.2 for three different Ω values. The absorber parameters $c_a = 2253\,\mathrm{Ns/m}$ and $k_a = 7.0086 \cdot 10^6\,\mathrm{N/m}$ used, determined by the condition (27), are optimal for $\Omega = 130\,\mathrm{rad/s}$. Note that the familiar condition of equal height of the peaks at resonance for the optimality also seems to apply in this case. However, due to the coupling between the horizontal and vertical displacements u and v, an additional third peak appears between the two conventional peaks which correspond to the motion of the beam and absorber mass in phase and antiphase. The third peak is due to the lowest resonance in the horizontal direction at $\omega = 150\,\mathrm{rad/s}$. Actually, there is an *antiresonance* in the vertical direction at this value of ω, because the resonance in the horizontal direction brings about an energy transfer from the vertical to the horizontal vibrations. As a result the third peak appears at $\omega = 151\,\mathrm{rad/s}$. On the other hand, if the lowest horizontal resonance frequency falls outside the region between the leftmost peak and the resonance frequency of the bare roll without absorbers, only two peaks will appear. Also, for decreasing Ω the coupling gets weaker and the midmost peak vanishes (case $\Omega = 0$). It should be noted that the values of c_a and k_a are not optimal for $\Omega = 0$ and 80 rad/s, which clearly demonstrates that the gyroscopic and dissipational coupling bears a considerable effect on the optimal parameters values of the dynamic absorber.

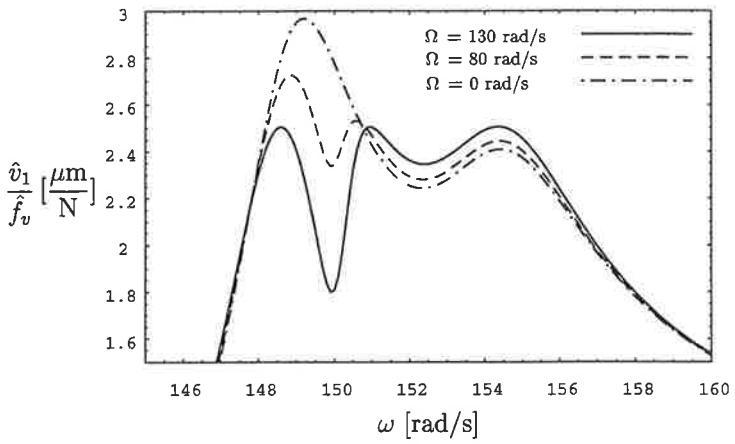


Figure 2. Frequency response function in the vertical direction for three roll rotational speeds and $\frac{K}{\tilde{K}} = 92\,\%$.

The calculated optimal absorber parameters c_a and k_a as a function of the rotational speed for three different bearing stiffness ratios $\frac{K}{\tilde{K}}$ are shown in Fig.3. It can been seen that for $\frac{K}{\tilde{K}}$ close to unity the optimal tuning depends relatively strongly on Ω. This relates to the considerations above. When the lowest horizontal and vertical resonances are close to each other, the horizontal motion will interfere with the vertical motion thereby affecting the optimal tuning. Note also that percentually c_a seems to be more sensitive to the coupling.

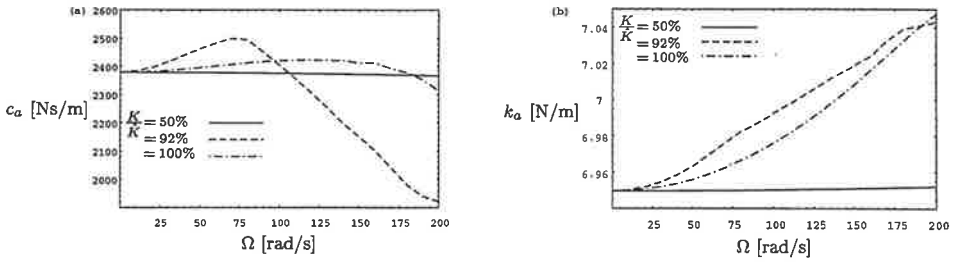


Figure 3. Optimal absorber parameters (a) c_a and (b) k_a as a function of the rotational speed for three bearing stiffness ratios.

From a practical point of view the effectiveness of the absorber is of considerable interest. In order to study this, the amplitude reduction factor η will be defined as the ratio of the maximum values of the frequency response function in the neighbourhood of the lowest resonance for the optimally tuned roll and the roll without absorbers. The function η as a function of the absorber size is plotted in Fig.4 for $\Omega = 80\,\text{rad/s}$. It can be seen that η falls steeply near the origin indicating that even with small absorber masses a considerable vibration attenuation can be achieved. When the absorber size increases further, the η-function levels out and only a minor improvement is obtained.

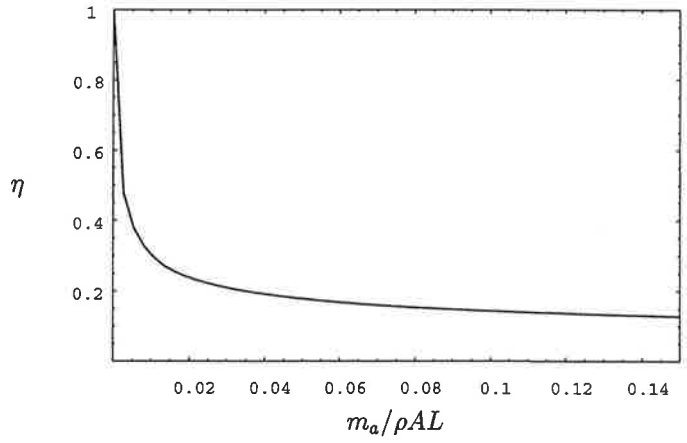


Figure 4. Reduction of the maximum of the frequency response function relative to that of the bare roll as a function of the absorber size for $\Omega = 80\,\text{rad/s}$.

4. CONCLUSIONS

A closed form analytic solution for a rotating Rayleigh beam driven by an arbitrary periodic load and with dynamic vibration absorbers at its ends is presented. The boundary conditions consist of horizontal and vertical translational springs and viscous dampers at the ends of the beam. However, the procedure is general and can be used

for any linear boundary conditions and additional dynamic absorbers as well. The extension to an arbitrary nonperiodic load is also straigthforward – just replace the Fourier sums by the Fourier integrals.

As an application a numerical example of a paper machine roll is studied. The significant effect of the gyroscopic and dissipational coupling on the optimal dynamic absorber tuning is demonstrated and the optimal absorber parameters are presented as a function of the rotational speed of the roll. It was found that, for certain ratios of the horizontal and vertical bearing stiffnesses, the frequency response function at the optimal tuning displays three peaks of equal heights instead of the conventional two ones. Finally, the effectiveness of the dynamic absorber is studied as a function of the absorber size. The conclusion is that even for very small absorber masses a considerable vibration attenuation can be achieved.

References

[1] H. Frahm. Device for Damping Vibrations of Bodies, US Patent No. 989 958, October 1909.

[2] J. Ormondroyd and J.P. Den Hartog. The Theory of the Dynamic Vibration Absorber. *Transactions of the American Society of Mechanical Engineers*, 16:109–117, 1928.

[3] J. E. Brock. A note on the damped vibration absorber. *Transactions of the American Society of Mechanical Engineers*, 68:A284, 1946.

[4] A. G. Thompson. Optimum Tuning and Damping of a Dynamic Vibration Absorber Applied to a Force Excited and Damped Primary System. *Journal of Sound and Vibration*, 77:403–415, 1981.

[5] D. Young. Theory of dynamic vibration absorbers for beams. In *Proceedings of the First U.S. National Congress of Applied Mechanics*, pages 91–96, 1952.

[6] J. C. Snowdon. Vibration of a cantilever to which dynamic absorbers are attached. *Journal of the Acoustical Society of America*, 39:878–886, 1966.

[7] R. G. Jacquot. Optimal dynamic vibration absorbers for general beam systems. *Journal of Sound and Vibration*, 60:535–542, 1978.

[8] H. N. Özgüven and B. Candir. Suppressing the first and second resonances of beams by dynamic vibration absorbers. *Journal of Sound and Vibration*, 111:377–390, 1985.

[9] D. N. Manikanahally and M. J. Crocker. Vibration absorbers for hysteretically damped mass-loaded beams. *Journal of Vibration and Acoustics*, 113:116–122, 1991.

DESIGN OF THE DECELERATION DYNAMICS AND TRIBOLOGY OF A HIGH SPEED ROTOR

H. MARTIKKA* and M. KUOSA†

*Faculty of Construction Design, Department of Mechanical Engineering

† Department of Energy Technology

Lappeenranta University of Technology, P.O.Box 20,

FIN-53851 Lappeenranta, FINLAND

ABSTRACT

The deceleration dynamics and tribology of a high speed rotor system are studied. The rotor is vertical and in normal operation it is supported by magnetic bearings. In emergency braking situations its should be stopped fast with minimal radial and axial damage to a ring sliding bearing. A list of feasible tribological material pairs are tested by simulation for selecting the best ones for field tests. Wear rates and wear volumes under various assumptions are estimated and also power consumptions and temperature rises. Dynamic behaviour of the rotor bearing system is studied using analytical modelling and simulation solution and also using a multibody 3D dynamics program Working Model. The simulations are used to design the experimental testings. There is a need for more quantitive material models and interdisciplinary design when designing tribological advanced high speed products. Present tools give reasonable results but integrated design tools are needed.

1. INTRODUCTION

Customers using high speed machinery need to decelerate them in emergencies. They are satisfied with such a mechanical braking system which endures cost-effectively sufficiently many decelerations. The aim of this study is to serve this goal. It consists of five subgoals:

The first goal is to study the deceleration dynamics of the rotor bearing system using analytical modelling and simulation. The aim is to use this model to optimize the designs. The second goal is to simulate the deceleration dynamics using a 3D multibody dynamics simulation program and CAD modelling of the geometry. The third goal is to study the tribology of the system for selecting the optimal wear resistant material pair. The fourth goal is to calculate the transient temperatures of the critical bearing for use in material selection. The final fifth goal is to utilize the previous results in designing the testing program and make recommendations.

2. DECELERATION DYNAMICS OF THE ROTOR USING ANALYTICAL DYNAMICS AND SIMULATION SOLUTION

The Jeffcott flexible-rotor model used as a starting model according to Childs [1]. The equations of motion for the system shown in Fig.1 are

$$\begin{aligned} m\ddot{R}_X + k_r R_X &= f_X + ma_X\dot{\phi}^2 + ma_Y\ddot{\phi} \\ m\ddot{R}_Y + k_r R_Y &= f_Y + ma_Y\dot{\phi}^2 - ma_X\ddot{\phi} \\ J_z\ddot{\phi} &= T_z + ma_Y\ddot{R}_X - ma_X\ddot{R}_Y \end{aligned} \tag{1}$$

where m is the mass of the rotor, mass of the shaft is not considered, k_r is the shaft stiffness coefficient, J_z is the moment of inertia of the disk about its Z-axis. The components of the external force vector are $\mathbf{f} = \{f_X \; f_Y\}$ and the component of the external moment vector along Z-axis is T_z. The transverse motion vector of the rotor is $\mathbf{R} = \{R_X \; R_Y\}$. The vector $\mathbf{a}$ is the imbalance vector. The mass is located at vector

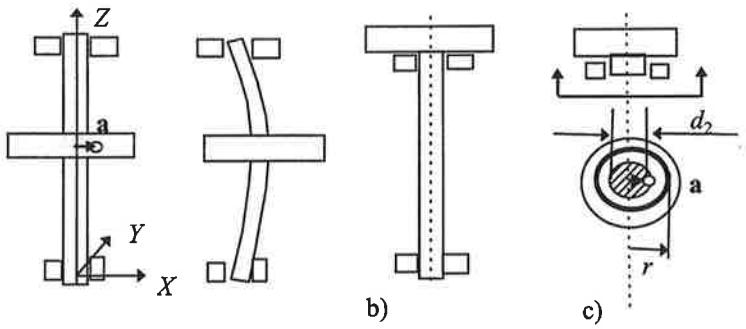
$$\begin{aligned} &\mathbf{S} = \mathbf{R} + \mathbf{A} = R_X + iR_Y + (a_X + ia_Y) \\ &\mathbf{S} = \mathbf{R} + \mathbf{a}e^{i\phi} \qquad \mathbf{a} = a_x + ia_y = ae^{i\gamma} \\ &\ddot{\mathbf{S}} = \ddot{\mathbf{R}} + \ddot{\mathbf{A}} \\ &\ddot{\mathbf{S}} = \ddot{\mathbf{R}} + \mathbf{a}e^{i\phi}\left(-\dot{\phi}^2 + i\ddot{\phi}\right) = \ddot{\mathbf{R}} + (a_X + ia_Y)\left(-\dot{\phi}^2 + i\ddot{\phi}\right) = \ddot{R} - \begin{bmatrix} a_X\dot{\phi}^2 + a_Y\ddot{\phi} \\ a_Y\dot{\phi}^2 - a_X\ddot{\phi} \end{bmatrix} \end{aligned} \tag{2}$$

The principle of virtual work may be used to express the equations

$$\begin{aligned} \left(X_{act} - m\ddot{S}_X\right)\delta x = 0 &= \left(f_X - kR_X - m\left\{\ddot{R}_X - a_X\dot{\phi}^2 - a_Y\ddot{\phi}\right\}\right)\delta x = 0 \\ \left(Y_{act} - m\ddot{S}_Y\right)\delta y = 0 &= \left(f_Y - kR_Y - m\left\{\ddot{R}_Y - a_Y\dot{\phi}^2 + a_X\ddot{\phi}\right\}\right)\delta y = 0 \end{aligned} \tag{3}$$

The torque balance is

$$\left(T_Z\hat{\mathbf{k}} - J\ddot{\phi}\hat{\mathbf{k}} - \mathbf{a}\times m\ddot{\mathbf{R}}\right)\bullet\hat{\mathbf{k}}\delta\phi = 0$$
$$\mathbf{a}\times m\ddot{\mathbf{R}} = \left(a_X + ia_Y\right)\times m\left(\ddot{R}_X + i\ddot{R}_Y\right) = m\left(a_X\ddot{R}_Y - a_Y\ddot{R}_X\right) \tag{4}$$



a) b) c)

FIGURE 1: Rotor shaft models. a) The Jeffcott flexible-rotor model where the disk is at the middle of the shaft, b) present model and c) approximate model

The dynamical equations of motion were solved as follows. First the highest derivatives must be solved as functions of the lower ones. Substituting the angular accelaration into *X Y* equations gives

$$A\,\ddot{R}_X + B\,\ddot{R}_Y = C + D\dot{\phi}^2$$
$$A'\,\ddot{R}_X + B'\,\ddot{R}_Y = C' + D'\dot{\phi}^2 \tag{5}$$

The highest derivative components can now be solved as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} \ddot{R}_X \\ \ddot{R}_Y \end{bmatrix} = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$
$$\ddot{R}_X = \left(k_1a_{22} - k_2a_{12}\right)/D \quad \ddot{R}_Y = \left(k_2a_{11} - k_1a_{21}\right)/D \tag{6}$$
$$D = a_{11}a_{22} - a_{12}a_{21}$$

$$\left[m - \frac{ma_Y}{J_z}ma_Y\right]\ddot{R}_X + \left[\frac{ma_Y}{J_z}ma_X\right]\ddot{R}_Y = f_X + ma_X\dot{\phi}^2 + \frac{ma_Y}{J_z}T_Z$$
$$\left[\frac{ma_X}{J_z}ma_Y\right]\ddot{R}_X + \left[m + \frac{ma_X}{J_z}ma_X\right]\ddot{R}_Y = f_Y + ma_Y\dot{\phi}^2 - \frac{ma_X}{J_z}T_Z \tag{7}$$

The state and derivative variables are defined as follows

STATE RX RXt RY RYt q qt B

DER dRX dRXt dRY dRYt dq dqt dB

For the state variable solution the following definitions are made

$$\begin{aligned} dR_X &= R_{Xt} \quad , \qquad dR_Y = R_{Yt} \\ dR_{Xt} &= R_{Xtt}(\phi_t) \quad , \quad R_{Yt} = R_{Ytt}(\phi_t) \\ d\phi &= \phi_t \quad , \qquad d\phi_t = \frac{1}{J_z} M = \frac{1}{J_z}\left(T_Z + ma_Y R_{Xtt} - ma_X R_{Ytt}\right) \\ d\beta &= \beta_t \end{aligned} \tag{8}$$

The following assumptions are made :

a The rotor moves in global X and Y directions and not in vertical Z direction

b Radial clearance is 0.3 mm, and unbalance vector is $a = 1 \cdot 10^{-6}$ m

c The movement of the total mass m is considered

The components of the unbalance vector in the global XYZ and the local rotor fixed xyz coordinates are

$$\begin{aligned} a_X &= a_x \cos\phi - a_y \sin\phi \\ a_Y &= a_x \sin\phi + a_y \cos\phi \\ a_x &= a\cos\gamma \quad = a \quad , \quad a_y = a\sin\gamma = 0 \end{aligned} \tag{9}$$

The total decelerating torque is sum of three torques

$$T_Z = T_{air} + T_{Zax} + T_{Zrad} = -k\dot{\phi}^2 - \mu Z_{ax} mgr - \mu Z_{rad} N r_2 \tag{10}$$

Here T_{air} is the decelerating torque due to air flow resistance at the rotor. A simple model is

$$T_{air} = k\dot{\phi}^2 \ , \quad k = 6.112 \cdot 10^{-6} \tag{11a}$$

A more accurate empirical model is as follows

$$\begin{aligned} T_{air} &= -sign(\dot{\phi})(if \quad \dot{\phi} > 1355 \quad then \quad K_1 \quad else \quad K_f) \\ K_f &= if \quad \dot{\phi} < 1355 \quad and \quad \dot{\phi} \geq 773 \quad then \quad K_2 \quad else \quad K_3 \\ K_i &= a_i + b_i\dot{\phi} + c_i\dot{\phi}^2 + d_i\dot{\phi}^3 + e_i\dot{\phi}^4 \quad , \quad i = 1...3 \end{aligned} \tag{11b}$$

The sign functions for the axial and radial friction surfaces are

$$\begin{aligned} Z_{ax} &= sign(V_{ax}) \quad , \quad V_{ax} = r\dot{\phi} \\ Z_{rad} &= sign(V_{rad}) \quad , \quad V_{rad} = r_2\dot{\beta} - r_2\dot{\phi} \end{aligned} \tag{12}$$

here r is friction radius of the upper surface between the bearing ring, r_1 is radius in the inner surface of the bearing ring, r_2 is the radius of the shaft through the bearing ring. Now here F is normal radial force at a contact and Δr is radial clearance (0.3 mm). F causes

$$p = R_X - \Delta r \tag{13}$$

where p is change of diameter d_1 due to compression by the normal force F by which the shaft compresses the bearing according to Niemann [3]

$$\Delta d_1 = p = F \cdot V\left(\tfrac{1}{3} + \ln\left(\frac{d}{b}\right)\right) = F \cdot c(F) \quad , \quad V = \frac{4(1-\nu^2)}{\pi E L} \tag{14}$$

where

$$d = \frac{d_1 d_2}{d_1 - d_2} \quad , \quad d_1 = d_2 + 2\Delta r \quad , \quad b = 1.08\left(\frac{Fd}{EL}\right)^{1/2} \tag{15}$$

Now for the restoring force F a simpler model $F = k_F \cdot p$ with $k_F = 1 \cdot 10^9 = 10000/2 \cdot k_r$ was used. Here E is elastic modulus of the bearing ring. The shaft is steel, ν is poisson's ratio and L is length of the bearing in axial direction.

Now p is known and the normal force N is calculated as

$$N = if \quad p > 0 \quad then \quad F \quad else \quad 0 \tag{16}$$

Momentary contact is at angle β relative from the global X axis

$$\tan\beta = \frac{R_Y}{R_X} \quad \dot{\beta}\frac{1}{\cos^2\beta} = \frac{\dot{R}_Y}{R_X} - \frac{R_Y \dot{R}_X}{{R_X}^2} = \frac{\dot{R}_Y R_X - R_Y \dot{R}_X}{{R_X}^2}$$
$$\dot{\beta} = \frac{\dot{R}_Y R_X - R_Y \dot{R}_X}{{R_X}^2 + {R_Y}^2} \tag{17}$$

The components of the external force **f** on the rotor are

$$f_X = N\left(-\cos\beta + \mu Z_{rad} \sin\beta\right)$$
$$f_Y = N\left(-\sin\beta - \mu Z_{rad} \cos\beta\right) \tag{18}$$

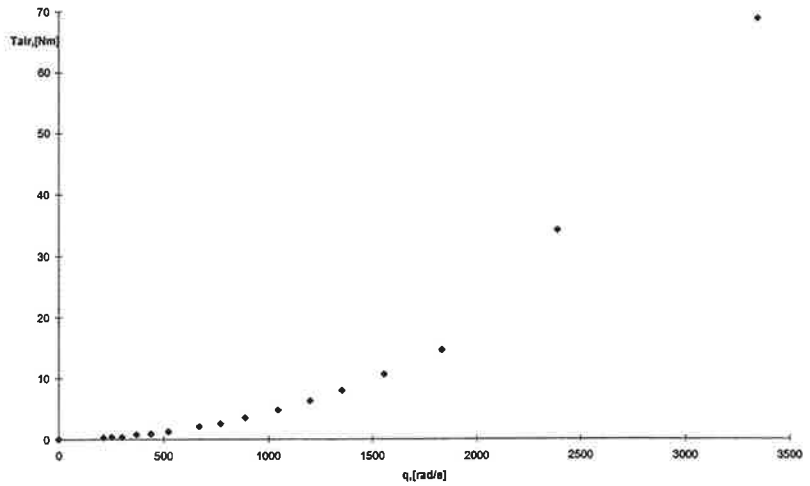


FIGURE 2: The measured air torque T_{air} vs. angular velocity $\dot{q}$ model.

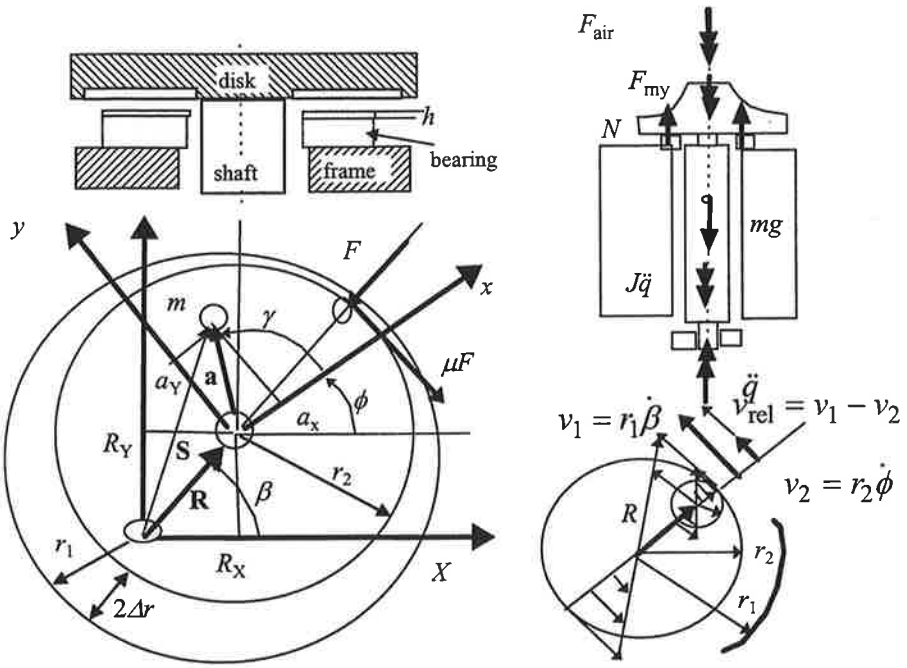


FIGURE 3: The rotor-shaft model. a) Jeffcot model b) geometry of motion

Some results of Simnon [2] simulation are shown in Fig.4.

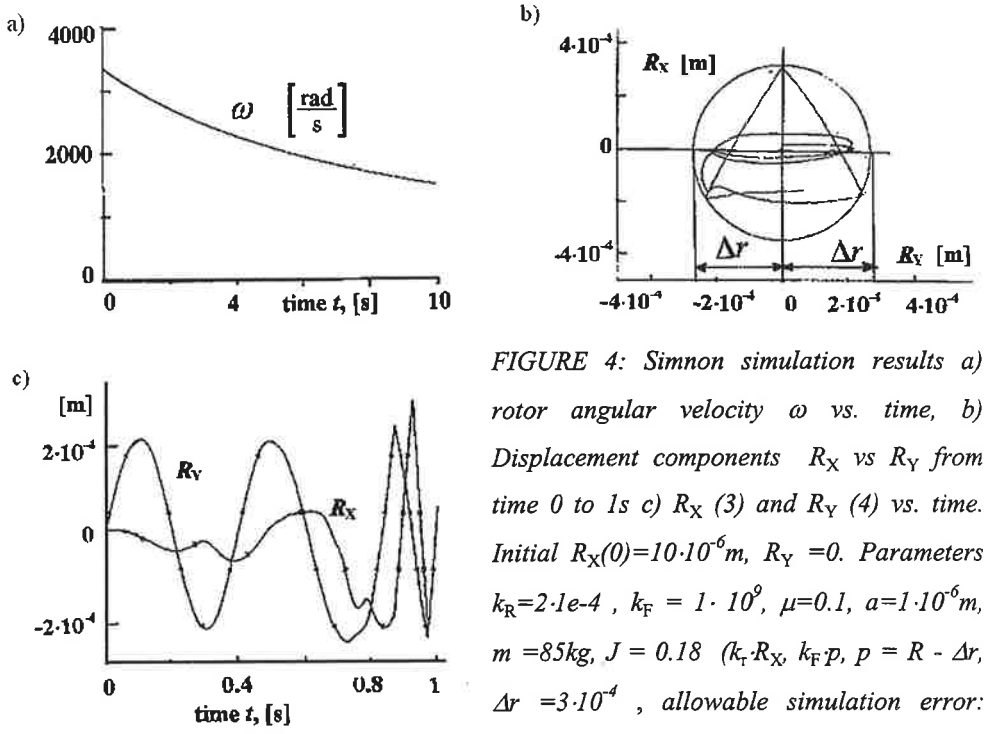


FIGURE 4: Simnon simulation results a) rotor angular velocity ω vs. time, b) Displacement components R_X vs R_Y from time 0 to 1s c) R_X (3) and R_Y (4) vs. time. Initial $R_X(0)=10\cdot10^{-6}$m, R_Y =0. Parameters k_R=2·1e-4 , k_F = 1· 10^9, μ=0.1, a=1·10^{-6}m, m =85kg, J = 0.18 ($k_r \cdot R_X$, $k_F p$, p = R - Δr, Δr =3·10^{-4} , allowable simulation error: 0.5·10^{-8}

3. SIMULATION OF ROTOR DYNAMICS WITH A PROGRAM

The Working model 3D of version 2 was used (Pre-release version) [4]. Some results are shown in Figs.5-14. A three dimensional model was made using AutoCAD, Fig.10. This model was then transferred to WM. The coefficient of friction between the bearing and the shaft was estimated as $\mu = 0.1$. The coefficients of restitution were for the rotor $e = 0.3$ and for the bearings $e = 0.2$. Initial rotational velocity was 20000 r/min. The fast initial drop in ω in Fig.5 partiably probably due to calculational accuracy.

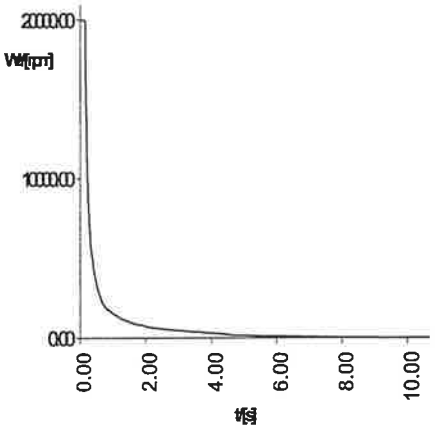


FIGURE 5: Angular velocity W_z [r/min] of rotor.

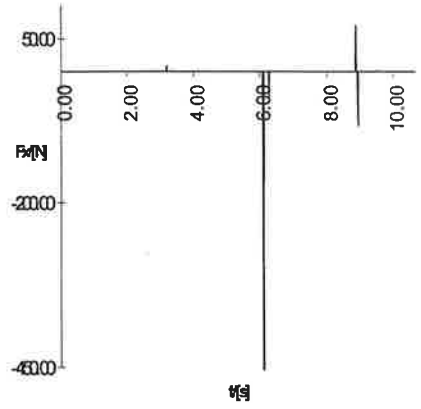


FIGURE 6: Contact force F_x between bearing and rotor.

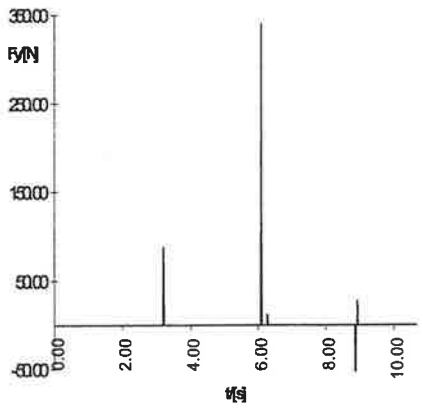


FIGURE 7: Contact force F_y between the bearing and the rotor.

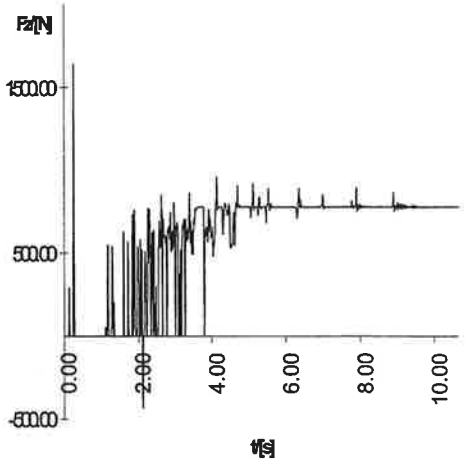


FIGURE 8: Contact force F_z between the bearing and the rotor.

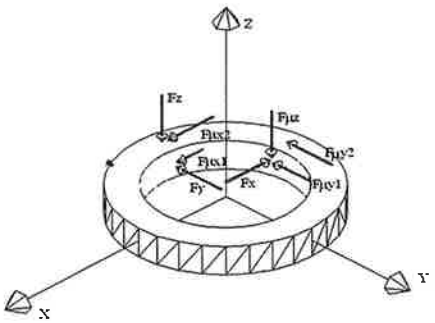


FIGURE 9: Coordinates and forces of WM3D model.

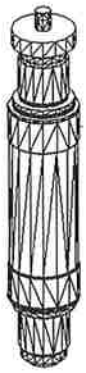


FIGURE 10: Autocad model of the rotor.

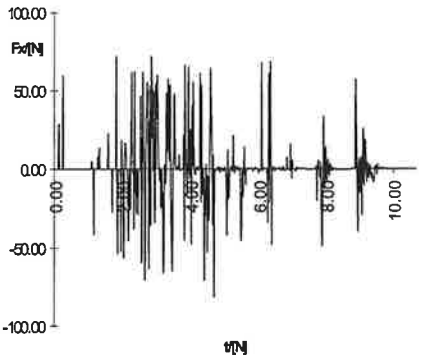


FIGURE 11: Friction force $F_{\mu x}$ between the bearing and rotor due to forces F_z and F_x.

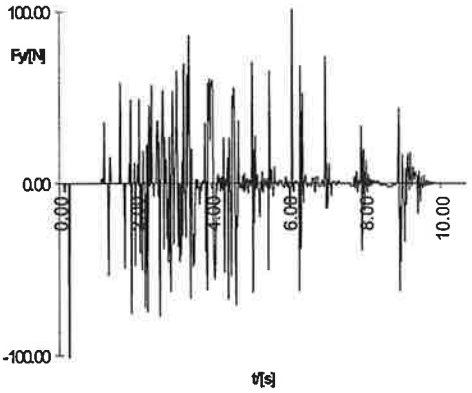


FIGURE 12: Friction force $F_{\mu y}$ between bearing and rotor due to forces F_z and F_y

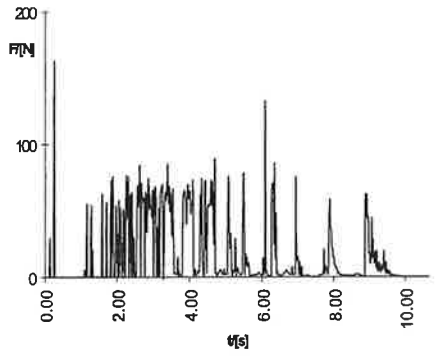


FIGURE 13: Resultant force F of friction forces $F_{\mu x}$ and $F_{\mu y}$

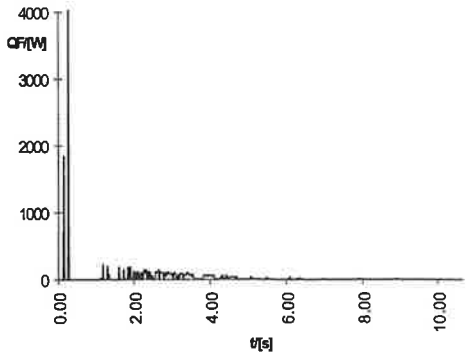


FIGURE 14: Total friction power Q_F between the bearing and the rotor.

4. WEAR MODELS

In analysing a tribological situation the following aspects are regarded as important :
1 Load, 2 speed, 3 vibration and dynamic loading, 4 temperature, 5 presence of loose abrasives, 6 nature of loose abrasives, 7 nature of gaseous environment, 8 contaminants, 9 lubrication and 10 damage in manufacture or assembly.

4.1 The linear wear model

This model gives the wear volume as, Fig.15a

$$\frac{V}{S} = Z\frac{F_n}{H} \quad , \quad V = A_n h \quad , \quad h = \dot{h}t \quad , \quad S = vt \quad , \quad F_n = pA_n \quad \rightarrow$$
$$\frac{A_n \dot{h} t}{vt} = Z\frac{pA_n}{H} \quad \rightarrow \quad \dot{h} = pv\frac{Z}{H} \tag{19}$$

where V is the worn out volume [m^3], F_n is normal force [N], A_n is nominal area [m^2], p is nominal pressure [Pa], h is thickness of the worn out layer [m], t is time [s], S is sliding distance [m], v is sliding velocity [m/s], Z is wear coefficient, $Z = k$ and H is hardness on the wearing surface, $H[\text{MPa}] = 9.81 \cdot H[\text{kp}/\text{mm}^2]$.

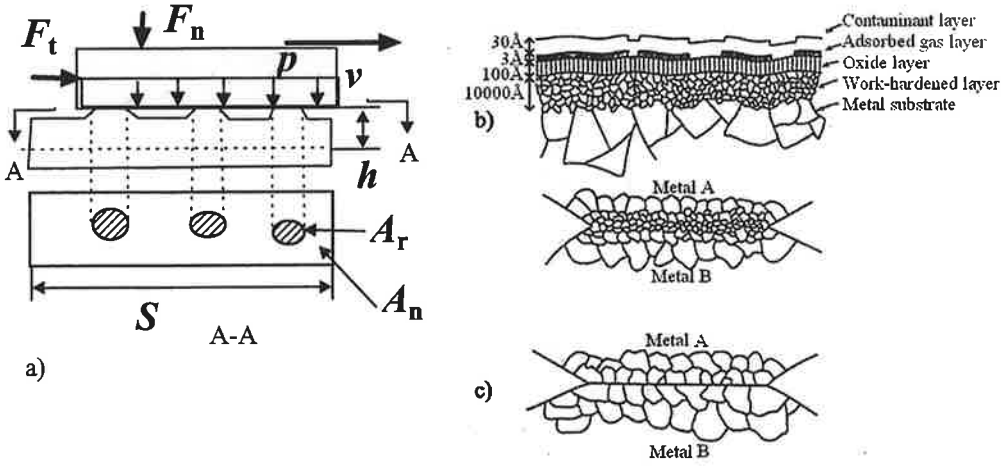


FIGURE 15: Wear models. a) Illustration of wear model, b) schematic view of films on a metal surface, c) a typical metallurgical and adhesional joint [5].

4.2 Hardness and mechanical properties

Wear rate is generally inversely proportional to hardness. The young's modulus of a polycrystalline metal is a structure-insensitive property. The yield strength of a metal is markedly structure dependent, however. Rabinowics [5] has plotted data from Metals handbook (1961) whenever possible data has been chosen for the pure metal in fully work hardened condition. This is the state that a metal surface might achieve after repeated sliding. The average relation for many materials is

$$\sigma_y = 0.003E \quad , \quad \sigma_y = \varepsilon_y E \quad , \quad H = 3\sigma_y$$
$$T_{melt} = \log(E)/A_k \quad , \quad A_k = \log(2.1\cdot 10^{11})/1500^o C \quad , \quad E\alpha^2 = 15\frac{N}{m^2}\cdot^o C^2 \tag{20}$$

where α is the coefficient of thermal expansion. Thermal stresses will appear with a thermal difference ΔT across the surface layer. $\sigma = E\alpha\Delta T$.

4.3 The adhesive wear rate models for metals and nonmetals

According to Rabinowicz [5] the wear coefficient $k = Z$ depends on the friction coefficient for metals (m) and nonmetals (n) as shown in Fig.16. Essentially, friction is not a temperature dependent quantity.

$$\dot{h}_m = pv\frac{Z_m}{H_m} \quad , \quad Z_m = A_m\mu^{B_m} \quad , A_m = 3.7\cdot 10^{-3} \quad , \quad B_m = 3.7 \quad H_m = H$$
$$\dot{h}_n = pv\frac{Z_n}{H_n} \quad , \quad Z_n = A_n\mu^{B_n} \quad , A_n = 26\cdot 10^{-6} \quad , \quad B_m = 2.05 \quad H_n = 0.2H \tag{21}$$

Rabinowicz gives also data [5] (p.160) for the following combinations. For WC on WC and for tool steel tool steel

$$\dot{h}_{WCWC} = pv\frac{Z_{WCWC}}{H_m} \quad , \quad Z_{WCWC} = 1.\cdot 10^{-6} \quad , \quad \dot{h}_{stst} = pv\frac{Z_{stst}}{H_m} \quad , \quad Z_{stst} = 1.\cdot 310^{-4} \tag{21}$$

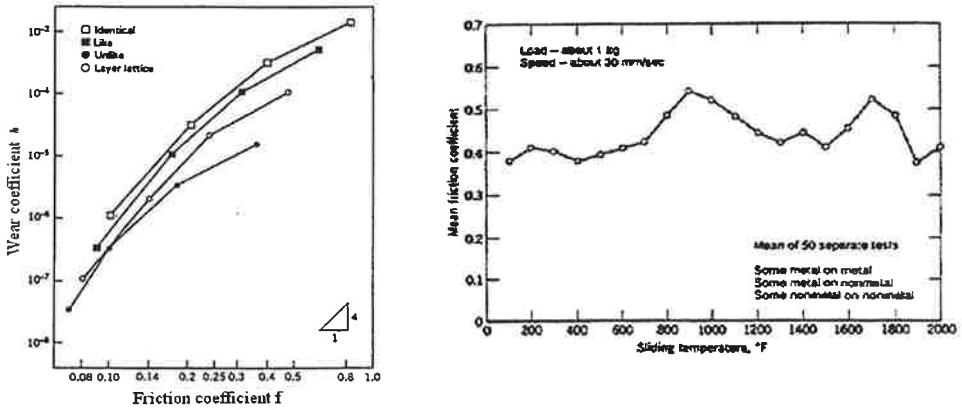
4.4 Abrasive wear

The abrasive wear rate equation looks similar to the adhesive equation. Rabonowicz gives [5] (p. 194) the following experimental values for the abrasive wear coefficient

$$\dot{h}_{abr} = pv\frac{Z_{abr}}{H_m} \tag{22}$$

TABLE 1: Abrasive wear coefficient values, [5] (Rabinowicz, p. 194)

lubrication	File	Abrasive paper , new	Loose abrasive grains	Coarse polishing
Dry surface	$500 \cdot 10^{-3}$	$10 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$0.1 \cdot 10^{-3}$
Lubricated	$1000 \cdot 10^{-3}$	$20 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$0.2 \cdot 10^{-3}$



a) *b)*

FIGURE 16: Wear models. a) The wear coefficient for metals k = Z vs. friction coefficient, b) mean friction coefficient vs. sliding temperature as derived from 50 widely varying material combinations [5].

4.5 Wear under the deceleration stage

The model is shown in Fig.3. The rotor is in normal operation supported by magnetic bearings in axial directions. To slow it down an axial retaining bearing is used. It also gives radial guidance to the rotor.

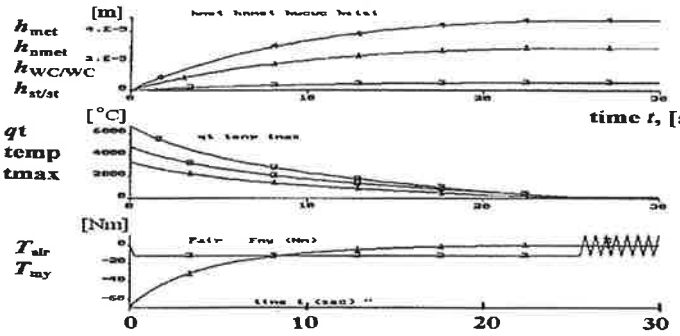


FIGURE 17: Wear and dynamical models. a) Weared out surface h for hmet(1) : for metals, hnmet(2) : for nonmetals, hwcwc(3) for WC/WC pairs (close to abscissa), hstst(4) for steel/steel pairs, b) qt(1) is rotational speed, temp(2) is mean temperature, tmax(3) is flash temperature, c) $F_{air}(1) = T_{air}$ air friction torque, $F_{my}(2) = T_{my}$ friction torque on the axial bearing surface.

The dynamic equation of motion of the rotor in the braking stage is

$$\begin{aligned} \ddot{q} &= \frac{T_{\text{air}} + T_{\text{my}}}{J} \\ T_{\text{air}} &= -sign(\dot{q}) k \dot{q}^2 \\ T_{\text{my}} &= if \quad abs(\dot{q}) < q_{\text{tdr}} \quad or \quad abs(q) < q_{\text{dr}} \quad then \quad T_{\text{m}} \quad else \quad T_{\text{mmy}} \end{aligned} \tag{23}$$

where the decelerating friction torque is

$$\begin{aligned} T_{\text{mmy}} &= -sign(v) \quad \mu N r \quad , \quad N = mg \quad , \quad v = \dot{q} r \\ T_{\text{m}} &= -T_{\text{myy}} \frac{q}{q_{\text{dr}}} \quad , \quad T_{\text{myy}} = X \mu N r \end{aligned} \tag{24}$$

Here $J = 0.18$ kg m^2 is the mass moment of inertia, $m = 85$ kg is the mass of the rotor, T_{air} is the torque resisting the rotation due to air flow. Experimentally it was determined by a power measurement as

$$P = T_{\text{air}} \omega = k\omega^2 \cdot \omega \quad ,230kW = k(3351 rad/s)^3 \quad \rightarrow k = 6.112 \cdot 10^{-6} \tag{25}$$

An approximate solution is obtained as follows with numerical estimates

$$\begin{aligned} J\ddot{q} &= -\mu mgr \qquad \ddot{q} = -K = -\frac{\mu mgr}{J} = -\frac{0.35 \cdot 85 \cdot 10 \cdot 0.04}{0.18} = -66 \\ \dot{q}(t) &= \dot{q}(0) - Kt \qquad T = \frac{\omega_0}{K} = \frac{3351}{66} = 50 \text{sec} \\ \dot{h} &= pv\frac{Z}{H} = (\omega_0 - Kt) rp \frac{Z}{H} \qquad h(t) = \left(\omega_0 t - \tfrac{1}{2} K t^2\right) rp \frac{Z}{H} \\ h(T) &= \tfrac{1}{2} \frac{{\omega_0}^2}{K} rp \frac{Z}{H} = \tfrac{1}{2} \frac{3351^2}{66} 0.04 \cdot 0.61 \cdot 10^6 \frac{10^{-5}}{1000 \cdot 10^6} = 2 \cdot 10^{-5} \end{aligned} \tag{26}$$

Substituting here further a model describing the dependence of the wear coefficient on the coefficient of friction gives the thickness of the worn out layer at time T to stop.

$$\begin{aligned} Z &= A_{\text{k}} \mu^{B_{\text{k}}} \\ h(T) &= \tfrac{1}{2} \frac{{\omega_0}^2}{K} rp \frac{Z}{H} = \frac{\tfrac{1}{2} J {\omega_0}^2}{A} \frac{rmg}{rmg\mu} \frac{Z}{H} = \frac{\tfrac{1}{2} J {\omega_0}^2}{A} \frac{A_{\text{k}} \mu^{B_{\text{k}}-1}}{H} = \frac{KE}{Area} \frac{A_{\text{k}} \mu^{B_{\text{k}}-1}}{Hardness} \end{aligned} \tag{27}$$

The allowed wear per one deceleration is h = 0.1 mm = $10 \cdot 10^{-5}$m. The calculated wear at room temperature for steel/steel pair is about $5 \cdot 10^{-5}$m, Fig.17a. But when the temperature of a steel workpiece increases from 0.2 to 0.4 T_{melt}, the hardness decreases by about a factor of about 0.3. Thus wear h increases about three times or more than allowed.

5. BEARING MATERIAL SELECTION USING FRICTION ENERGY INPUT AND COMPARISON OF CALCULATED AND ALLOWED TEMPERATURES

5.1 Bearing temperature estimates with simple stationary models

When surfaces slide together, almost all the energy is dissipated in friction and appears in the form of heat at the interface. The surfaces make contact not over the nominal area A_n but over only a few isolated junctions whose area is the real area of contact A_r, Fig.15a. During sliding these junctions are broken and their temperature is fairly even at flash temperature. The mean temperature of the layers is lower. Several models may be described. At the present case the pressure is 0.5 MPa in axial surfaces, sliding veloity is 100 m/s, ambient temperature is high 150 °C. No lubricants may be used.

Model A. At moderate speed v the interface attains an equilibrium mean temperature rise T_m above the rest of the material given by

$$T_m = \frac{\mu W v}{4 J r_j (\lambda_1 + \lambda_2)} \tag{28}$$

Here μ is friction coefficient, r_j is the radius of the junction, J is the mechanical equivalent of heat, W is the load carried by a single junction, and λ_1 and λ_2 are the thermal conductivities of the two contacting materials.

Model B. Rabinowics [5] gives the following simplified model

$$T_m = cv, \quad T_m\left[{}^{\circ}C\right], \quad v\left[\mathrm{m/s}\right], \quad c = 50\,(1/3...3)\left[{}^{\circ}\mathrm{C}/(\mathrm{m/s})\right] \tag{29}$$

Typical values are : $c = 50$ generally, $c = 9$ with brass on nylon, $c = 13$ with steel on bronze, $c = 32$ with steel on nylon, $c = 140$ steel on steel.

Model C. A simplified model for the mean temperature of the friction surface is based on the assumption that the friction power in conducted to the two contacting layers under steady state heat flow conditions

$$Q_F = Q_1 + Q_2$$
$$\mu p v A = A\left(\lambda_1 \frac{T - T_w}{h_1} + \lambda_2 \frac{T - T_w}{h_2}\right) \quad T = T_w + \frac{\mu p v}{\lambda_1 / h_1 + \lambda_2 / h_2} \tag{30}$$

5.2 Friction power estimate with stationary models

One method of design a retainer bearing is based on ability of the bearing to dissipate heat. In this approach a pv value is computed using the equation

$$pv = \frac{k(T_B - T_A)}{\mu} \tag{31}$$

where p is load per unit of projected bearing area [kPa], v is surface velocity of journal relative to bearing surface [m/s], T_A is ambient air temperature [°C], T_B is bearing bore temperature [°C], μ is coefficient of friction.

The constant k in Eq.(31) depends upon the ability of the bearing to dissipate heat. Table 2 shows some of the materials commonly used under dry or mixed-film conditions. It is to be noted that all quantities listed are maximum values. However, they cannot all be maximum at the same time. [7] (p. 463)

TABLE 2: Maximum normally allowable pressure, temperature, sliding velocity and pv values of some materials [7] .

MATERIAL	MAXIMUM pressure, MPa	MAXIMUM temperature, °C	MAXIMUM speed, m/s	MAXIMUM *pv* value, kPa * m/s
Cast bronze	31.0	165	7.5	1750
Porous bronze	31.0	65	7.5	1750
Porous iron	55.0	65	4.0	1750
Phenolics	41.0	95	13.0	530
Nylon	7.0	95	5.0	100
Teflon	3.5	260	0.5	35
Reinforced teflon	17.0	260	5.0	350
Teflon fabric	410.0	260	0.3	900
Delrin	7.0	80	5.0	100
Carbon-graphite	4.2	400	13.0	530
Rubber	0.4	65	20.0	...
Wood	14.0	65	10.0	530

First it is essential to represent the theory of dissipated heat generated from friction at the interfaces. Friction power that releases at sliding surfaces is

$$Q_F = Q_1 + Q_2 \tag{32}$$

where Q_1 represents heat dissipation to retainer bearing (Fig.18) and Q_2 heat dissipated to rotor shaft and impeller back ring.

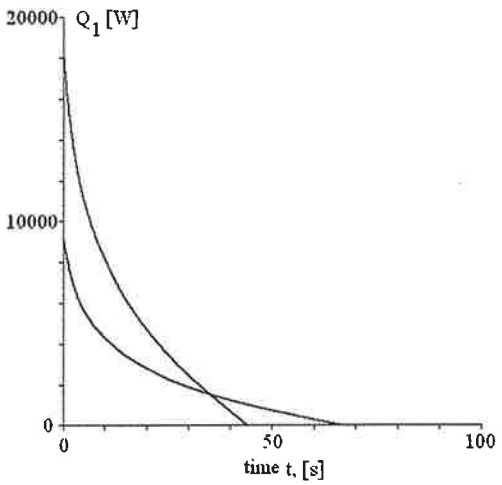


FIGURE 18: Heat power [W] dissipated to the retainer bearing vs. time [s] due to sliding friction at interfaces for friction coefficients 0.1(lower curve) and 0.2(upper curve).

5.3 Transient temperatures at the bearing

In order to select the best candidate material pairs for field tests the highest transient temperature loads are compared with melting temperatures or allowed limit temperatures. The load temperatures were calculated using the following approach.

a The partial differential equations of heat conduction in time and in thickness direction were discretized first, Fig.19c, Eq.33.

b The finite difference method with Euler application [6] was used to solve transient temperatures at nodes.

c Then these were solved using Simnon "discrete system" option.

$$\frac{dT_1}{dt} = \frac{T_1^{n+1} - T_1^{n}}{\Delta t} \qquad \frac{dT_1}{dy} = \frac{T_2^{n} - T_1^{n}}{h_1} \tag{33}$$

Here n and $n+1$ are sequential time step numbers, Δt is time step, T_A is ambient temperature in bearing housing, α [W/(m^{2}K)] is the combined coefficient of radiative and convective heat transfer, λ [W/(mK)] is coefficient of conductive heat transfer, h_i are layer thicknesses, ρ [kg/m^3] is density of the material, C_p [J/(kgK)] is specific heat and Q_1 is dissipated heat flow. In this model two material layers (h_1, h_2, ρ_1, ρ_2, C_{p1}, C_{p2}, λ_1 and λ_2) can be used.

Energy balance of node 1 gives (Fig.19c)

$$Q_1(t) = Q_{C\,1} + Q_{12} \qquad Q_1(t) = \rho_1 c_{p1} \frac{h_1}{2} \frac{T_1^{n+1} - T_1^n}{\Delta t} - \lambda_1 \frac{T_2^n - T_1^n}{h_1} \tag{34}$$

It is assumed that heat dissipation to the bearing is 1/2 times friction power

$$Q_1(t) = \frac{\frac{1}{2} Q_F}{A_n} \tag{35}$$

Here A_n is nominal area of axial surface of retainer bearing. Total friction power at sliding surfaces is Q_F. Energy balance of node 2 gives

$$Q_{12} = Q_{C\,12} + Q_{C23} + Q_{23} \tag{36}$$

$$-\lambda_1 \frac{T_2^n - T_1^n}{h_1} = \rho_1 c_{p1} \frac{h_1}{2} \frac{T_2^{n+1} - T_2^n}{\Delta t} + \rho_2 c_{p2} \frac{h_2}{2} \frac{T_2^{n+1} - T_2^n}{\Delta t} - \lambda_2 \frac{T_3^n - T_2^n}{h_2}$$

Energy balance of node 3

$$Q_{23} = Q_{C\,3} + Q_H \tag{37}$$

$$-\lambda_2 \frac{T_3^n - T_2^n}{h_2} = \rho_2 c_{p2} \frac{h_2}{2} \frac{T_3^{n+1} - T_3^n}{\Delta t} + \alpha\left(T_3^n - T_A(t)\right) \tag{38}$$

In this application we chose α = 30 W/(m^{2}K) according to Shigley [7] and assumed constant ambient temperature T_A = 150 °C. The following equations are obtained

$$T_1^{n+1} = (1 - 2F_1) T_1^n + 2F_1 T_2^n + 2\frac{h_1}{\lambda_1} F_1 Q_1^n \tag{39}$$

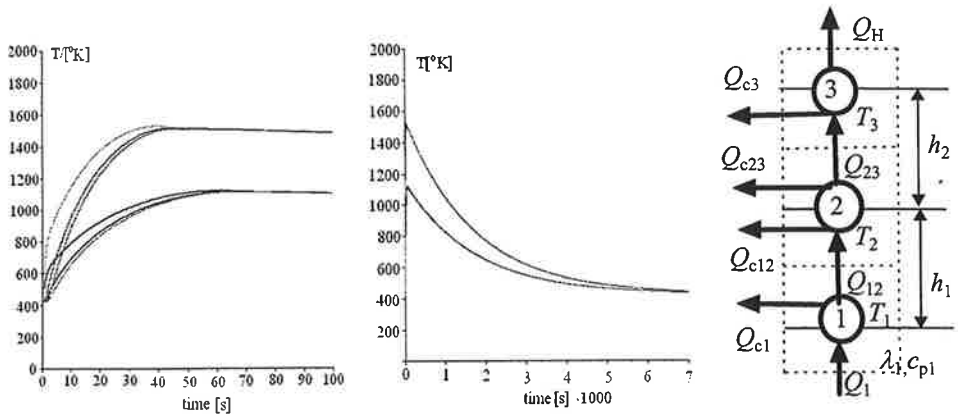
$$T_2^{n+1} = 2\frac{SF_1F_2}{F_1 + SF_2} T_1^n + \left(1 - 2\frac{F_1F_2(S+1)}{F_1 + SF_2}\right) T_2^n + 2\frac{F_1F_2}{F_1 + SF_2} T_3^n \tag{40}$$

$$T_3^{n+1} = 2F_2 T_2^n + \left(1 - 2F_2\left[1 + Bi_2\right]\right) T_3^n + 2F_2 Bi_2 T_A \tag{41}$$

Where

$$F_1 = \frac{\lambda_1 \Delta t}{\rho_1 C_{p_1} h_1^2} \qquad F_2 = \frac{\lambda_2 \Delta t}{\rho_2 C_{p_2} h_2^2} \qquad S = \frac{\lambda_1 h_2}{\lambda_2 h_1} \qquad Bi_2 = \frac{\alpha h_2}{\lambda_2} \tag{42}$$

Then these were solved using Simnon "discrete system" option. The results in Fig.19 show temperature vs. time curves at three nodes of the model for the retainer bearing made of iron with low friction surfaces using two assumed friction values μ = 0.1 and 0.2.



a) b) c)

FIGURE 19: Temperature rise of iron bearing. Friction coefficients are 0.1, 0.2, melting temperature 1473 °K. Temperature [°K] as function of time [s] at three node points, a) time scale 0 ...100 s , b) time scale 0...7000 s and c) a finite difference model used.

TABLE 3: Temperature rise peaks and melting temperatures [°K] of some material pairs with two coefficients of friction 0.1 and 0.2.

Material sliding on steel	Temperature at $\mu = 0.1$	Temperature at $\mu = 0.2$	Melting temperature or limit temperature
Bronze	1170	1610	1283
Steel	1120	1540	1623
Iron	1120	1520	1473
Stellite 6	980	1360	1533...1630
PI 60% + graphite 40%	4400	7300	glass trans. temp. 638
Silicon carbide	1450	2050	3073

6. DISCUSSION

Successful design of high speed machinery requires interdisciplinary approach and synergical combination expertises of rotor dynamics simulation, tribology and materials science and heat transfer. This approch gave satisfactory results. One drawback is the need to use of several design tools. Reliability of complex simulations need to be checked with simple models and also the sensitivity of results to error limit choices. An integrated design tool is required which is user friendly and gives reliable results cost-effectively.

7. CONCLUSIONS

Customers using high speed machinery often need to decelerate them. One choice is mechanical braking system based on dry friction. The goal in study was explore ways to optimize this system to endure sufficiently many decelerations. This goal consisted of the following subgoals.The first goal was analytical modelling and simulation solution of the

deceleration dynamics of the rotor bearing system.This was achieved satisfactorily and simulation results were nearly quantitative.This model can be used for optimization and sensitivity "what if studies". The second goal was to simulate the deceleration dynamics using a 3D multibody dynamics simulation program Working Model. The model geometry was created using AutoCAD and imported to WM. The results were reasonable and detailed. The third goal was to use tribology, material models and simple analytical dynamics for selecting the optimal wear resistant material pairs. The results showed that the wear resistance of structural steels was not sufficient. The fourth goal was to calculate the transient temperatures of the bearing. The result was safety margin of peak temperature load vs. endurance or now the melting temperature. This was needed in material selection. The fifth goal was to use the previous results to aid the design the testing program. This was achieved satisfactorily at start up stage of testings. It is recommended that more emphasis is laid to expertise on physical simulation modelling and testing and maintaing availability of expertise for use in future tasks cost-effectively.

ACKNOWLEDGEMENTS

Special thanks are expressed to Dr. Associate Professor Jaakko Larjola, Lappeenranta University of Technology, LUT, Dept. of Energy Technology, and Professor Tapani Jokinen at the Laboratory of Electromechanics at Helsinki University of Technology and Mr. Jukka Seppänen, High Speed Tech Ltd Oy for financial arrangements of this project. The contributions of the staff of the Laboratory of Energy Technology at LUT and Associate Professors Jaakko Larjola, Timo Talonpoika and Professor Pertti Sarkomaa are gratefully appreciated. The excellent contribution of Mr. Artturi Salmela at Dept. of Mechanical Engineering in Working Model 3D simulations is gratefully acknowledged.

REFERENCES

1. D.Childs, *Turbomachinery Rotordynamics, Phenomena, Modeling , and Analysis*, John Wiley & Sons, Inc., New York (1993).
2. *Simnon*, SSPA systems,Gothenburg, 1990
3. G. Niemann, *Maschinenelemente Band I.* Springer-Verlag Berlin, Heidelberg 1981
4. *Working Model.* Knowledge Revolution, San Mateo, California, USA, V. 4.0, 1997
5. E. Rabinowicz, *Friction and wear of materials*, second edition, John Wiley & Sons, New York, 1995
6. G. E. Myers, *Analytical methods in conduction heat transfer,* McGraw-Hill, 1971
7. J. E. Shigley, *Mechanical Engineering Design*, McGraw-Hill Book Company, 1986

BIFURCATION STRUCTURE OF A DRIVEN VAN DER POL-TYPE EQUATION

R. VON HERTZEN[1], O. KONGAS[1,2] AND J. ENGELBRECHT[2]
[1]Helsinki University of Technology, Otakaari 1, FIN-02150 Espoo, Finland
[2]Institute of Cybernetics, Akadeemia tee 21, EE-0026 Tallinn, Estonia

ABSTRACT

The bifurcation structure and bifurcation diagrams of a periodically driven van der Pol-type nerve pulse equation are presented. The bifurcation maps display 1:1 phase locked and period doubling regions, a classical region of Arnold tongues and quasiperiodic behaviour and an interesting transition region with periodic, quasiperiodic and chaotic solutions. The response exhibits Neimark-Sacker, period doubling and saddle-node bifurcations, *N:M*-type phase-locked states, Farey organization and chaotic behaviour. The rich variety of calculated arrhythmias and conduction blocks agrees well with measured behaviour of dog and sheep cardiac Purkinje fibers.

1. INTRODUCTION

Already in 1920's mathematical and experimental models were developed to explain cardiac arrhythmias. These models, based on a difference equation [1] and on coupled nonlinear electric oscillators [2], were used to display the generation of different rhythms related to atrioventricular heart block as a function of the system parameters. These studies are the earliest ones to emphasize the role of bifurcations of a nonlinear model to understand the qualitative behaviour in biological systems. Later studies, accounting also for spatial aspects, were performed utilizing the concept of an excitable medium [3]. These studies associated ventricular tachycardia and fibrillation with rotating spiral waves in cardiac tissue. An important contribution was the Hodgking-Huxley model based on nonlinear partial differential equations describing the excitation propagation in squid giant axon [4]. The model was developed to explain voltage clamp studies of the ion currents in squid nerves. Thereafter, an enormous amount of work has been done to explain the functioning of heart [5].

The idea that human disease may sometimes be associated with bifurcations in the dynamics of living organisms was originally proposed by Mobitz [1] and van der Pol and

van der Mark [2], and made more explicit by Mackey and Glass [6] with the notion of 'dynamical disease' to denote abnormal dynamics in physiological systems associated with changes in system parameters. Of most interest is the understanding of the bifurcations and topological features of the nonlinear equations under parametric changes. This topological approach considers the identification of disease as a problem of understanding the bifurcations in an appropriate underlying model system. They suggested that one basis for therapy is to manipulate the physiological parameters back into their normal ranges.

Many biological rhythmic processes can be modeled by nonlinear differential equations exhibiting limit cycle behaviour. Recently, periodically driven and coupled oscillators have been subject to considerable interest. The basic analytical model for the periodically perturbed biological oscillator was given by Guevara and Glass [7]. Many important systems with relaxation oscillations have been described by the harmonically driven van der Pol equation, being a basic model of driven self-excited oscillations in physics, electronics and biology. The driven van der Pol equation is one of the most intensely studied equations in nonlinear dynamics (see [8]). Much less studied are the *asymmetric* van der Pol and the *quiescent* nerve pulse equations. Nonlinear dynamics of the heartbeat was modeled by two coupled nonlinear oscillators using an analog electrical circuit with an external voltage source by West *et al.* [9]. A Bonhoeffer - van der Pol (BvP) model with self-sustained oscillations, exposed to periodic pulse trains, was used to describe the influence of periodic inhibitory trains on a crayfish pacemaker neuron [10]. Bonhoeffer - van der Pol equation has also been used to describe the cAMP signalling system in the cellular slime mold *Dictyostelium discoideum* and to model the cell cycle [11]. Many coupled oscillating systems, such as the primary and secondary pacemakers of the heart, have been modeled by the standard circle map.

A system can also contain nonlinear, spontaneously quiescent, excitable threshold elements (or neurons) paced by the oscillating parts of the system. The aperiodic response of non-spontaneously active cardiac Purkinje fibers and ventricular muscle cells to rhythmical stimuli from their surroundings was studied by Chialvo and Jalife [12] and Chialvo *et al.* [13]. A bifurcation analysis of the non-oscillating BvP equation stimulated periodically was given by Braaksma and Grasman [11] and Sato and Doi [14]. A mathematical model for periodically driven neurons and an analytical treatment for their firing frequency, explaining the experimental results of the artificial (transistor) neurons [15], were presented by Nagumo and Sato [16].

Nerve pulse propagation has attracted attention since Hodgkin and Huxley explained the mechanism of ion currents governing the pulse motion. Due to the complexity of the phenomenological Hodgkin-Huxley model, the simpler FitzHugh-Nagumo (FHN) model is widely used [17,18]. Still another approach, based on the full hyperbolic telegraph equations, leads finally to a Liénard-type nerve pulse equation (NPE) for the stationary wave profile of the transmembrane action potential [19,20]. The harmonically driven NPE is equivalent to the BvP equation except for a missing Duffing-type cubic term. Depending on

the parameter values, the NPE exhibits relaxation oscillations and can be considered as an (asymmetric) van der Pol equation, or the relaxation oscillations cease to exist [11,20] and the NPE becomes a quiescent excitable nerve pulse equation (or a quiescent van der Pol equation).

The cardiac electric conduction system can be considered as a network of selfoscillating pacemakers and quiescent, excitable, His-bundle and Purkinje fibers. The sino-atrial (SA) node, being the primary pacemaker, and the atrioventricular (AV) node, a secondary pacemaker, can be modeled as a pair of coupled relaxation oscillators [2,9,21]. Under normal conditions, the intrinsically faster SA node appears to entrain the slower secondary pacemaker resulting in a one to one phase-locking of the pacemakers [22]. However, perturbation of this system may lead to a complex dynamic interaction. The AV node, on the other hand, acts as a drive for the His bundle and Purkinje network, considered as non-pacemaking excitable media.

In this work the pacemakers of the heart in an entrained mode are modeled by a periodic train of Dirac delta spikes. These act as a drive for the cardiac conducting tissues (His-Purkinje network) modeled by the NPE. This pulse equation intrinsicly includes the refractory period of the nerve cells and no book-keeping of the expression or quenching of the pulse conduction is needed.

The paper is organized as follows. We first consider some basic features of heart physiology in view of the electric rhythm generation and the voltage transmission along the nerves. We then present the main results for the NPE and numerical results for the NPE driven by a periodic train of Dirac delta spikes, simulating real transmembrane action potential measurements. Finally, the conclusions are drawn.

2. PHYSIOLOGICAL BACKGROUND

Heart Dynamics

The rhythm of the heart is set by a small region of specialized myocardium in the right atrium, called the sinoatrial (SA) node, which generates a spontaneous electrical rhythm associated with the flow of ions, principally sodium, potassium and calcium, across the cell membrane. The rate of this sinuous pacemaker can be affected by nerval activity from sympathetic nerves which speed up the heart, and the vagus nerve, which slows down the heart. The electrical current spreads across the atria, and this in turn leads to a contraction of the atrial muscle. In normal individuals, the atria and ventricles are electrically coupled only by way of a small strip of specialized tissue, the atrioventricular (AV) node. The rest of the tissue that separates the atria and ventricles is nonconducting fibrous tissue. Specialized fibers, the Purkinje fibers, rapidly conduct the electrical activation to the ventricular subendocardium. Activation of the muscle cells is then completed within about 0.06s in

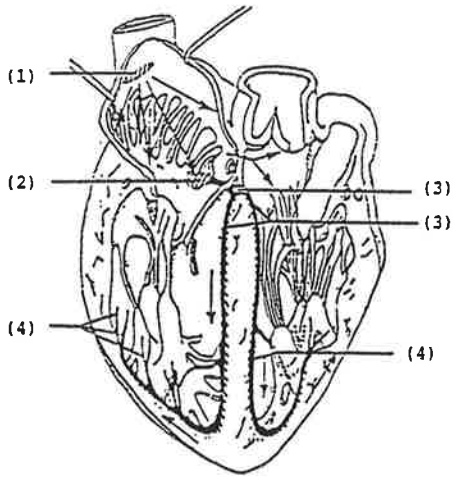


FIGURE 1. Main elements of electric action potential generation and propagation in heart: (1) SA node (2) AV node (3) His bundle with branches (4) Purkinje fibers.

humans ensuring nearly synchronized contraction of the massive ventricular muscle. The basic units of contractile material within the muscle cells are the sarcomeres which, when switched on by the calcium ion fluxes, triggered by the action potential, produce force in the direction of the muscle fibers. All the mechanical events of the heart as a pump are driven by the repetitive electric signal, the action potential, which starts from the SA node. The sequence of the main elements is the following [23]:

- SA *node* with its rhythmicity (a basic clock for the cardiac cycle)
- AV *node* introducing a delay which allows effective mechanical contraction of atria followed by ventricles
- *Bundle of His* and its branches carrying the action potential
- *Purkinje fibers* distributing the action potential over the myocardium
- *Myocardium* contracting as a result of the distributed action potential.

Nerve Pulse Transmission

The contemporary understanding of nerve pulse propagation is based on the membrane theory. The nerve pulse (voltage) is transmitted down the axoplasm core of a nerve which is surrounded by a cylindrical membrane. The currents through this membrane from the axoplasm to the intersticial fluid and inversely govern the nerve pulse propagation.

The process is the following [24]: Both the axoplasm (inside the nerve) and the intersticial fluid (surrounding the nerve) contain ions of sodium (Na^+) and potassium (K^+) as well as other ions. The relative concentration of sodium and potassium ions create the transmembrane potential. At equilibrium the value of this potential can be estimated by the Nernst equation [25]. Calculations show that the axoplasm contains a larger potassium and lesser sodium ion concentration than the intersticial fluid, resulting in a resting potential in the range (–50 mV, –100 mV).

If an electrical stimulus is applied to the nerve, the membrane acts in different ways depending on the value of the stimulus. If the stimulus is below a certain threshold value, the depolarization process of the membrane is reversible and the equililbrium state returns fast without any pulse propagating. If the stimulus is above the threshold, the permeability

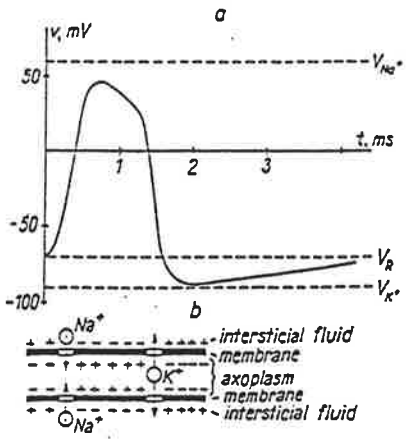


FIGURE 2. (a) Typical transmembrane action potential profile and (b) idealized nerve with ion currents.

of the membrane to sodium ions is increased and the inward flow of the sodium ions starts. The positive increase in the transmembrane potential causes then a further decrease of the permeability to these ions and the potential goes through a depolarizing phase resulting in a value of about +50 mV at the inside of the membrane. This process is followed by an increase in the potassium permeability which causes an outward flow of the potassium ions. When the potassium outward current equals the sodium inward current, the transmembrane potential peaks and the process returns to the equilibrium through an undershoot. As a result, an asymmetric solitary wave propagates in the nerve. This is depicted in Fig. 2.

Model for Nerve Pulse Propagation

To model the nerves we start from the Lieberstein [19] model

$$\pi a^2 C_a \frac{\partial v}{\partial t} + \frac{\partial i_a}{\partial x} + i = 0 ,$$
$$\frac{L}{\pi a^2}\frac{\partial i_a}{\partial t} + \frac{\partial v}{\partial x} + r i_a = 0 , \tag{1}$$

where v is the potential difference across the membrane, i_a the axon current per unit length, i the membrane current per unit length, a the axon radius, C_a the axon self-capacitance per unit area per unit length, L the axon specific self-inductance and r the resistance per unit length. Further, it is convenient to introduce the membrane current density $I = i/(2\pi a)$. Equations (1) are hyperbolic transmission line equations (telegraph equations).

For the current density I a FitzHugh-Nagumo type cubic expression

$$I = \alpha v + \beta v^3 + w \tag{2}$$

with one recovery variable w is used. The recovery variable obeys the equation

$$\frac{\partial w}{\partial t} = \gamma(v + v_0) . \tag{3}$$

Here α, β, γ and v_0 are constants. Starting from the above equations it can be shown [20,26] that the final equation for the transmembrane action potential becomes

$$\frac{\partial^2 v}{\partial \xi \partial s} + f(v)\frac{\partial v}{\partial \xi} + g(v) = 0 , \tag{4}$$

where

$$f(v) = a_0 + a_1 v + a_2 v^2 \quad , \quad g(v) = av \tag{5}$$

are polynomials with constant coefficients, s is distance along the nerve and the variable

$$\xi = c_0 t - s \tag{6}$$

has been used. Here t is time and c_0 the propagation velocity determined from the telegraph equations. Equation (4) together with the initial condition describes the full dynamics of the wave. However, when a stationary profile after the transient has been formed, the

corresponding ordinary differential equation may be utilized. Because the stationary wave does not propagate with the equilibrium velocity c_0, the transformation

$$\eta = s + \theta\xi \tag{7}$$

is introduced. This leads to the equation

$$\frac{d^2v}{d\eta^2} + f(v)\frac{dv}{d\eta} + \frac{1}{\theta}g(v) = 0, \tag{8}$$

where θ denotes a pseudovelocity determining the final velocity of the progressive wave through

$$c = \frac{\theta}{\theta - 1}c_0 \tag{9}$$

It is further shown that equations (4) and (8) exhibit a threshold, a possible amplification of the initial excitation and a formation of a stationary profile with a characteristic refractive part. It is also shown that the stationary wave profile of equation (8) is very similar to that of the FitzHugh-Nagumo equation.

We prefer writing equation (8) in the form

$$\frac{d^2v}{d\eta^2} + F(v - v_1)(v - v_2)\frac{dv}{d\eta} + v = 0, \tag{10}$$

where typical values for mammal nerves $F = 3.265$, $v_1 = 0.5$ and $v_2 = 1.9$ are used [26]. In this work these values for the nerve pulse equation are used unless otherwise stated.

A single stimulus acting on the nerve pulse equation (NPE) (10) in its resting state will be either attenuated or amplified resulting in a *sub-threshold* (small stimulus) or *supra-threshold* (larger stimulus) wave shown in Fig. 3.

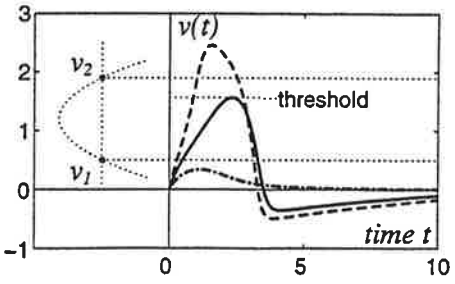


FIGURE 3. Threshold, sub- and supra-threshold wave profiles of equation (10).

The threshold solution on the $(v, dv/d\eta)$-phase plane touches tangentially the line of zero isoclins, which locates the inflection points of the curves $v = v(\eta)$. In the case of cardiac

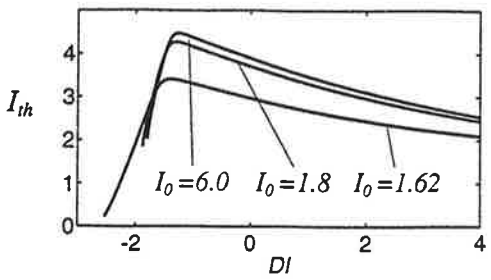


FIGURE 4. Threshold impulse I_{th} for the NPE for various initial conditions $(v_0, v_0') = (0, I_0)$.

Purkinje fibers the sub-threshold response is not able to fire the contraction in the myocardium cells resulting in a skip of one heart beat. The minimum impulse (velocity kick) $I_{th} = I_{th}(DI; I_0)$ needed to create a supra-threshold response, as a function of the impulse exertion time, is presented in Fig. 4. The curve displays a maximum indicating that the NPE under consideration behaves as a *supernormal neuron.*

3. RESULTS AND DISCUSSION

Simulation of Action Potential Experiments

In the following we study the nerve pulse equation (10) driven periodically by a train of Dirac delta spikes. This is of practical interest since many experimental transmembrane action potential measurements are performed using cell stimulation by short rectangular current pulses [12,13]. The emphasis is on a detailed study of the bifurcation diagrams of the response as a function of the drive frequency for different drive amplitudes and on the bifurcation map on the drive frequency-amplitude plane. The model equation for the ventricular conduction fibers driven rhythmically by (infinitely) short depolarizing current pulses thus reads

$$\frac{d^2v}{d\tau^2} + F(v - v_1)(v - v_2)\frac{dv}{d\tau} + v = I\sum_{n=0}^{\infty}\delta(\tau - nT), \tag{11}$$

where I accounts for the strength and $T=2\pi/\omega$ for the basic cycle length (period) of the stimulus, and $\delta(\cdot)$ is the Dirac delta function.

The overall bifurcation structure of equation (11) displaying the bifurcation lines of period one solutions on the (I,ω)-control plane is shown in Fig. 5. Here dotted and solid lines denote period doublings and Neimark-Sacker bifurcations, respectively. A detailed study of bifurcations up to period six solutions revealed the classic Farey organization and Arnold tongue structure for phase locking zones originating from the Neimark-Sacker lines within the quasiperiodic region, period doubling cascades and chaos originating from the two 'butterfly wings', and a complicated transition region between these main domains. A more detailed bifurcation structure within the zone $1 \le \omega \le 3$ is shown in Fig. 6. Here Neimark-

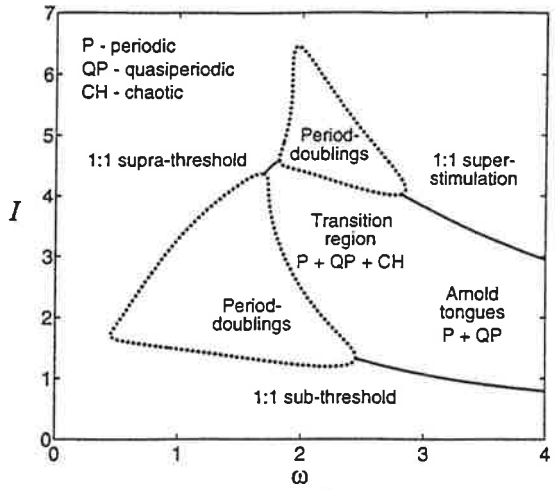


FIGURE 5. The skeleton of the bifurcation structure of eq. (11).

Sacker bifurcations are labeled by NS and saddle-node and period doubling bifurcations by $sn(m,n)$ and $pd(m,n)$, where m and n denote torsion and period numbers, respectively, as follows [8]:

$sn(m,n)$ one saddle and one node, both with period n and torsion m, coincide and disappear

$pd(m,n)$ an orbit with period $n/2$ and torsion $m/2$ bifurcates into an orbit with period n and torsion m.

The torsion number m equals the average number of windings per drive period of a neighboring orbit around the underlying attractor. For attractors with torsion number m and period number n the ratio $\rho = m/n$ defines the generalized winding number (GWN). It should be noted that the GWN is also well defined for systems with a broken or no invariant torus [27]. Let us consider Fig. 6(b). The bifurcation sequences $pd(1,2) \rightarrow$

$pd(1,4) \rightarrow$ ...chaos within the lower wing and $pd(1,6) \rightarrow pd(1,12) \rightarrow$ ...chaos and $pd(1,8) \rightarrow pd(1,16) \rightarrow$ ...chaos within period three and four Arnold tongues bordered by $sn(1,3)$ and $sn(1,4)$ lines, respectively, are found. The Arnold tongues originate as cusp-like forms from the points of resonances $R(1,3)$ and $R(1,4)$ at the Neimark-Sacker bifurcation line approximately at $\omega = 3$ and $\omega = 4$. The period doubling bifurcations inside the tonques are very close to the saddle-node bifurcations bordering the tongues. Note overlapping of tongues three and four and tongue three with the lower wing in the vicinity of chaotic regions. The higher period tongues do not display overlapping nor do they bifurcate, and together with the intervening quasiperiodic regions they form a typical Arnold tongue structure on the right hand side of the transition region. In the upper part of Fig. 6(b) a cusp pertaining to $sn(1,2)$ bifurcations appears very close to the $pd(1,4)$ line. This cusp creates a fold which leans over the period doubling region and continues up to higher values of I, changing to a $sn(1/2,2)$ bifurcation line above $I \cong 3.1$. Note also that the cusp is connected

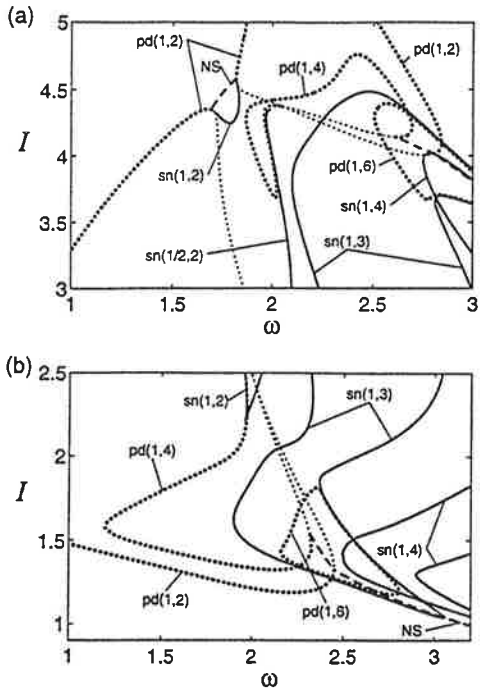


FIGURE 6. More detailed bifurcation structure around the transition region.

to the Neimark-Sacker line via the $pd(1,4)$ bifurcation line. This suggests that the cusp can be interpreted as a reminiscence of period two Arnold tongue. Upper part of the $1 \leq \omega \leq 3$ strip shown in Fig. 6(a) displays quite a similar structure with period doublings and Arnold tongues than the lower part. However, the $pd(1,4)$ bifurcation line extends below the upper

wing following the $sn(1/2,2)$ line down to $I \cong 3.7$. Note also that the period doubling wings are connected via a NS bifurcation line enclosing, together with the $sn(1,2)$ line, a closed area (lobe) displaying quasiperiodic behaviour. The lobe contains higher Arnold tongues and, therefore, this NS bifurcation line can be interpreted as a continuation of the NS bifurcation lines on the other side of the wings. The $pd(1,2)$ lines, which intervene the NS line, arise due to the 2:1 conduction block, which is primarily caused by the supernormality of the NPE under consideration.

In Figs 7-9 the bifurcation diagram using the Poincaré section points v_p, defined as the maximum values of v between consecutive Dirac delta spikes, largest Lyapunov exponent λ and generalized winding number ρ (GWN) for steady state solutions are shown as a function of ω for different drive amplitudes I. The definition of the GWN is based on the torsion of the local flow around a given attractor. Within a period-doubling cascade, the GWN exhibits an interesting stepwise structure [27] embodied in the sequence

$$\rho_n = \rho_\infty \pm \frac{(-1)^n}{3m_0 \cdot 2^n} \quad , \tag{12}$$

where

$$\rho_\infty = \rho_0 \mp \frac{1}{3m_0} \quad . \tag{13}$$

Here ρ_0 and m_0 are the GWN and the period of the first attractor in the cascade. Upper or lower signs in equations (12) and (13) must be chosen depending on which attractor is chosen as the first one. Equation (12) describes a sequence of alternatingly falling and rising steps with the step height halved within every bifurcation.

A typical bifurcation diagram for small drive amplitudes is shown in Fig. 7(a). For increasing ω the stable period one attractor experiences a NS bifurcation leading to quasiperiodic behaviour (see the Lyapunov exponent). The NS bifurcation occurs approximately (within a relative error of 0.15 %) when the average value of the drive $I/T = \omega I/2\pi$ equals the lower root $v_1 = 0.5$ at $\omega_{c1} \cong 4.19$. The amplitude spread in the Poincaré section seems at first to follow the typical square root law [28]. However, when the firing over the *instability strip* $v_1 < v < v_2$ starts, a sudden amplitude rise occurs. The periodic windows within the quasiperiodic region are narrow. When the drive average equals (within a relative error of 0.06 %) the value of the upper root $v_2 = 1.9$ at $\omega_{c2} \cong 15.9$, a reverse NS bifurcation occurs. This is preceded by a sudden amplitude decrease analogously to the lower end of the instability strip. At $\omega > \omega_{c2}$ a stable period one attractor resides above the instability strip (parts of the curve near the minima may penetrate into the instability strip). This corresponds to a *superstimulated* state where the repolarization of the neuron is prohibited. The GWN

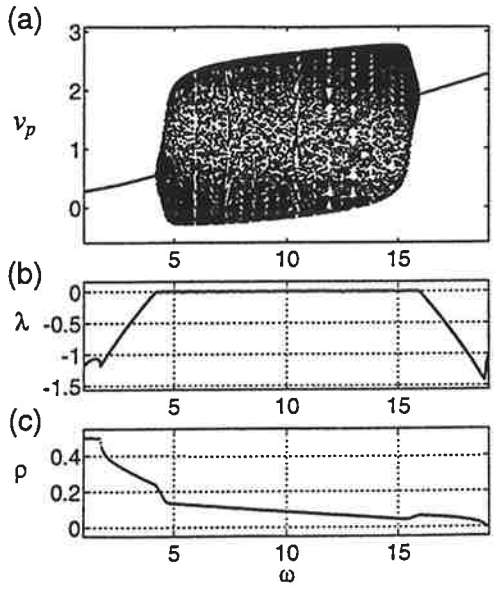


FIGURE 7. (a) Bifurcation diagram, (b) largest Lyapunov exponent and (c) generalized winding number as a function of ω for $I = 0.75$.

displays a regular devil's staircase-type behaviour throughout the ω-interval within the quasiperiodic region.

When the drive amplitude approaches the firing threshold ($I_{th} = 1.62$) for a single stimulus, the bifurcation diagram experiences strong exchanges. This is displayed in Fig. 8 for the value $I = 1.5$. A period doubling cascade and chaotic region now precede the quasiperiodic regime and replace the NS bifurcation (see Fig. 5). The period doublings occur at ω-values corresponding to a drive average lower than the root v_1. Consequently, it is natural that no NS bifurcation occurs at this stage. Nevertheless, the drive amplitude $I = 1.5$ is quite near the firing threshold for a single stimulus. This means that the system phase point penetrates deep into the instability strip, even at low drive frequencies, resulting in a strong interplay between the drive and the NPE leading to the observed period doublings and chaotic behaviour. For increasing ω the drive average assisted oscillations of the NPE become strong enough to compete with the external drive. This results in the quasiperiodic behaviour starting at $\omega \cong 3.3$ via a saddle-node bifurcation from the period four window. As before, the reverse NS bifurcation and the appearance of the period one superstimulated state occur for still larger values of ω. The small period windows at the left end of the bifurcation diagram are widened compared to the case of lower I. A closer examination shows that the dotted regions around the period doubling points in Fig. 8(c) indeed form the stepwise

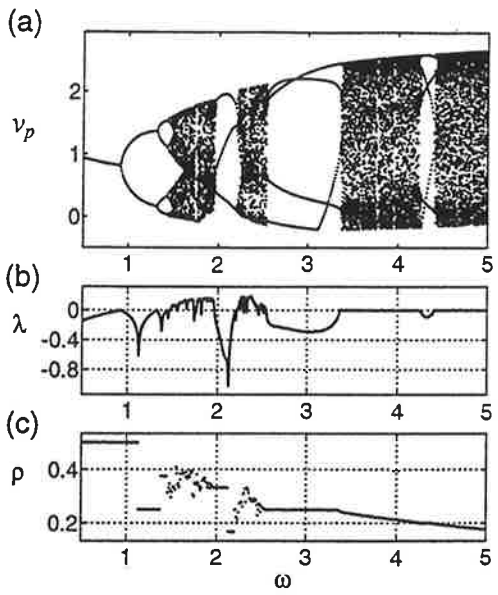


FIGURE 8. Same as Fig. 7 for $I = 1.5$.

structure determined by equations (12) and (13). The distinction between period doubling and saddle-node bifurcations is easily made solely on the basis of the GWN. Note also that the jump in the GWN occurs at the superstable solution (minimum of the Lyapunov exponent) indicating qualitative changes in the local flow around the attractor at this point.

The case for $I = 2.43$ is shown in Fig. 9. The drive amplitude has exceeded the single stimulus threshold and the response is supra-threshold already at small driving frequencies. For increasing ω the response bifurcates only once and then becomes directly quasiperiodic via a saddle-node bifurcation. The chaotic region is now omitted and the bifurcation diagram displays a beautiful sequence of alternating quasiperiodic regimes and phase-locked periodic states, emerging and disappearing via saddle-node bifurcations. For $I \lesssim \pi(\nu_1 + \nu_2)/\omega$ the Arnold tongues increase in size as the drive amplitude increases. The behaviour is quite similar to that displayed by the sine circle map at values $K < 1$ [29]. For increasing ω the period of the widest windows increases by one from window to window while the width of the windows decreases gradually. At $\omega_{c2} \simeq 4.89$ a reverse NS bifurcation takes place. The widest phase-locked plateaus visible in Fig. 9 correspond to 1:1 , 2:1 ,..., 7:1 stimulus: response lockings. These plateaus, as well as the whole devil's staircase, are clearly displayed by the GWN. Between the 1:1 and 7:1 lockings, the GWN falls from 1 to 1/7. After the falling and rising steps given by equation (12) a region of gradual monotonic decrease in the GWN appears. Within the quasiperiodic region we also find higher periodic

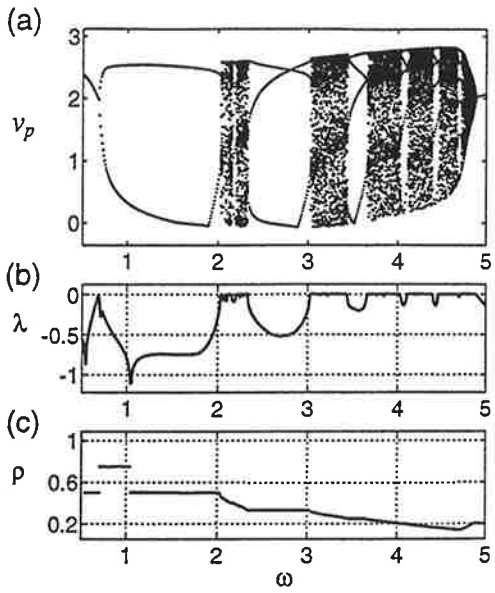


FIGURE 9. Same as Fig. 7 for $I = 2.43$.

windows ordered according to the Farey hierarchy (5:2 between 2:1 and 3:1, for example) and manifested by the devil's staircase as well. Besides the *N*:*M* locking ratio, the order of the supra- and sub-threshold responses (i.e. action potentials and dropped beats) in a sequence obtained by Farey-combination can also be predicted. For example, the 2:1 and 3:1 states can be represented by the patterns (1,0) and (1,0,0), respectively. Consequently, the pattern of the 5:2 state will be (1,0,1,0,0). A similar behaviour has been found by Chialvo and Jalife [12].

Low dimensional chaotic behaviour in sheep and dog cardiac Purkinje tissues has been demonstrated by Chialvo *et al.* [13], where period doubling bifurcations of the transmembrane action potential amplitude were shown to precede the irregular action potential behaviour. For a 1.5 × threshold drive amplitude and increasing drive frequency, the period doubling bifurcations associated with changes in the stimulus:response locking were manifested in the sequence 1:1 → 2:2, 2:1 → 4:2, 3:1 → 6:2 and 4:1 → 8:2 until irregular activity occurred at very brief cycle lengths. Careful exploration also revealed irregular dynamics between the 4:2 and 3:1 states. For a 2.6 × threshold drive amplitude, the corresponding sequence was 1:1 → 2:2 → 4:4 → irregular dynamics. Additional evidence to show that this irregular activity was in fact deterministic chaos was sought by plotting a return map for the action potential amplitude, which displayed a one-hump-type behaviour, typical of one-dimensional chaotic maps.

In order to compare with these experimental results, let us look at Figs 5 and 6. It can be seen from Fig. 6(b) that a horizontal cut at $I = 1.5 \times I_{th} \simeq 2.43$ results in a different sequence of attractors than the corresponding experimental one. However, a slight tilting down from the horizontal makes the cut pass through the sequence 1:1 $\rightarrow$ 2:2, 2:1 $\rightarrow$ 4:2 ... chaos, 3:1 $\rightarrow$ 6:2 ... chaos, 4:1 $\rightarrow$ 8:2 ... chaos, quasiperiodicity, 5:1 and so on. The consistency with the above experimental sequence is excellent. It can be seen that a slight raise of the cut makes it pass through the sequence 1:1 $\rightarrow$ 2:2, 2:1, quasiperiodicity, 3:1, quasiperiodicity, 4:1, quasiperiodicity, 5:1 and so on, i.e., through the classic Arnold tongue structure avoiding period doublings and chaotic regions. The real action potential measurements should be extended to these values to see if real Purkinje fibers also display the corresponding behaviour. It can be seen from Fig. 6(a) that a slightly tilted cut around $I = 2.6 \times I_{th} \simeq 4.21$ results in the sequence 1:1 $\rightarrow$ 2:2 $\rightarrow$ 4:4 $\rightarrow$ ... chaos $\rightarrow$ inverse period doubling sequence to the superstimulated state. Again, the fit to the corresponding experimental sequence is excellent. Note also that in the ω-direction the domain of the 2:1 solution in the lower wing is about two times that of the 2:2 solution in the upper wing, which also agrees well with the experimental results for sheep Purkinje fibers [13].

4. CONCLUSIONS

In conclusion, some characteristic features of a second order spontaneously quiescent Liénard-type nerve pulse equation were studied. This nerve pulse equation was proposed to model the transmembrane action potential propagation in the cardiac His-Purkinje nervous network. The threshold and sub- and supra-threshold responses were described. A closer examination of the firing threshold revealed that the nerve pulse equation under consideration is a supernormal neuron.

For comparison with experimental results for dog and sheep cardiac Purkinje fibers, the nerve pulse equation was stimulated by a periodic train of narrow spikes. For small stimulus intensities the response developed, for increasing drive frequency, from a 1:0 sub-threshold state via NS bifurcations and quasiperiodic behaviour to a 1:1 superstimulated state. For stimulus intensities near the single stimulus threshold value, period doubling bifurcations and chaotic behaviour replaced the NS bifurcation at lower driving frequencies. This development has no counterpart in the sine circle map and originates from the firing property of the NPE. However, for larger driving frequencies the quasiperiodic behaviour was developed via a saddle-node bifurcation from a periodic state leading finally, via a reverse NS bifurcation, to a 1:1 superstimulated state as above. Sequences of *N:M*-type phase-locked states displaying a Farey-tree and devil's staircase structure were found. For intermediate stimulus intensities the response was supra-threshold even at low drive frequencies and, for increasing ω, developed directly into quasiperiodic behaviour with intervening periodic windows. The behaviour resembled that of the sine circle map for $K <$ 1. For still larger stimulus intensities the tori started to break until only chaos interrupted by

periodic windows was present. The reverse NS bifurcation leading to the 1:1 superstimulated state was replaced by a reverse period doubling cascade. The whole bifurcation structure displayed an approximate symmetry with respect to the center line $I\omega = \pi(\nu_1 + \nu_2)$.

The calculated results agreed closely with the measured transmembrane action potential responses for supernormal dog and sheep cardiac Purkinje fibers. It was proposed that the measurements with real Purkinje fibers would be extended to also detect a possible quasiperiodic behaviour.

REFERENCES

1. Mobitz, W. 1924. "Über die unvollständige Störung der Erregungsüberleitung zwischen Vorhof und Kammer des menschlichen Herzens." *Z. Gesamte Exp. Med. **41**,* 180-237.
2. Van der Pol, B. and van der Mark, J. 1928. "The heartbeat considered as a relaxation oscillation and an electrical model of the heart." *Phil. Mag. Suppl. **6**,* 763-775.
3. Wiener, N. and Rosenblueth, A. 1946. "The mathematical formulation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle." *Arch. Inst. Cardiol. Mex. **16**,* 205-265.
4. Hodgkin, A.L. and Huxley A.F. 1952. "A quantitative description of membrane current and its application to conduction and excitation in nerve." *J. Physiol. **117**,* 500-544.
5. Glass, L.; Hunter, P.J. and McCulloch, A. (Editors) 1991. *Theory of Heart: Biomechanics, Biophysics, and Nonlinear Dynamics of Cardiac Function.* Springer-Verlag, New York.
6. Mackey, M.C. and Glass, L. 1977. "Oscillation and chaos in physiological control systems." *Science **197**,* 287-289.
7. Guevara, M.R. and Glass, L. 1982. "Phase Locking, Period Doubling Bifurcations and Chaos in a Mathematical Model of a Periodically Driven Oscillator: A Theory for the Entrainment of Biological Oscillators and the Generation of Cardiac Dysrhythmias." *J. Math. Biology. **14**,* 1-23.
8. Mettin, R., Parlitz, U. and Lauterborn, W. 1993. "Bifurcation Structure of the Driven van der Pol Oscillator." *Int. J. Bifurcation and Chaos **3**,* 1529-1555.
9. West, B.J., Goldberger, A.L., Rovner, G. and Bhargava, V. 1985. "Nonlinear Dynamics of the Heartbeat." *Physica **17 D**,* 198-206.
10. Nomura, T., Sato, S., Doi, S., Segundo, J.P. and Stiber, M.D. 1993. "A Bonhoeffer - van der Pol oscillator model of locked and non-locked behaviors of living pacemaker neurons." *Biol. Cybern. **69**,* 429-437.
11. Braaksma, B. and Grasman, J. 1993. "Critical dynamics of the Bonhoeffer - van der Pol equation and its chaotic response to periodic stimulation." *Physica D **68**,* 265-280.
12. Chialvo, D.R. and Jalife, J. 1987. "Non-linear dynamics of cardiac excitation and impulse propagation." *Nature **330**,* 749-752.
13. Chialvo, D.R., Gilmour, R.F, Jr. and Jalife, J. 1990. "Low dimensional chaos in cardiac

tissue." *Nature **343,*** 653-657.

14. Sato, S. and Doi, S. 1992. "Response Characteristics of the BVP Neuron Model to Periodic Pulse Inputs." *Math. Biosci. **112,*** 243-259.
15. Harmon, L.D. 1961. "Studies with Artificial Neurons, I: Properties and Functions of an Artificial Neuron." *Kybernetik **1,*** 89-101.
16. Nagumo, J. and Sato, S. 1972. "On a Response Characteristic of a Mathematical Neuron Model." *Kybernetik **10,*** 155-164.
17. FitzHugh, R. 1961. "Impulses and Physiological States in Theoretical Models of Nerve Membrane." *Biophys. J. **1***, 445-466.
18. Nagumo, J., Arimoto, S. and Yoshizawa, S. 1962. "An Active Pulse Transmission Line Simulating Nerve Axon." *Proc. IRE **50***, 2061-2070.
19. Lieberstein, H.M. 1967. "On the Hodgkin-Huxley Partial Differential Equation." *Math. Biosci. **1***, 45-69.
20. Engelbrecht, J. and Tobias, T. 1987. "On a model stationary nonlinear wave in an active medium." *Proc. R. Soc. Lond. **A411***, 139-154.
21. Katholi, C.R., Urthaler, F., Macy, J., Jr and James, T.N. 1977. "A Mathematical Model of Automaticity in the Sinus Node and AV junction Based on Weakly Coupled Relaxation Oscillators." *Comp. Biomed. Res. **10***, 529-543.
22. Goldberger, A.L., Bhargava, V., West, B.J. and Mandell, A.J. 1985. "Nonlinear Dynamics of the Heartbeat." *Physica **17 D***, 207-214.
23. Noble, M.I.M. 1979. *The Cardiac Cycle*. Blackwell Sci. Publ., Oxford.
24. Marsocci, V.A. 1982. "Electrical Network Modelling of Active Membranes of Nerves." *CRC Crit. Rev. in Biomed. Engng **8,*** 135-194.
25. Scott, A.C. 1977. *Neurophysics*. Wiley, New York.
26. Engelbrecht, J. 1986. "The Evolution of Nonlinear Waves in Active Media." *Wave Motion **8***, 93-100.
27. Parlitz, U. and Lauterborn, W. 1987. "Period-doubling cascades and devil's staircases of the driven van der Pol oscillator." *Phys. Rev. **A36,*** 1428-1434.
28. Hassard, B.D.; Kazarinoff, N.D. and Wan, Y-H. 1981. *Theory and Applications of Hopf Bifurcation*. London Math. Soc. Lecture Note Series No 41, Cambridge Univ. Press, Cambridge.
29. Jensen, M.H., Bak, P. and Bohr, T. 1984. "Transition to chaos by interaction of resonances in dissipative systems. I. Circle maps." *Phys. Rev. **A30***, 1960-1969.

ON VARIATIONAL PRINCIPLE APPROACH TO EIGENVALUE PROBLEMS OF NON-CONSERVATIVE MECHANICAL SYSTEMS.

J.KASKI
Department of Mechanical Engineering
Lappeenranta University of Technology
P.O. Box 20, 53851 Lappeenranta
E-mail: Juha.Kaski@lut.fi

ABSTRACT

In this paper, the concept of self-adjointness is generalized and extended in a certain sense to the problems which are non-selfadjoint in the classical sense. An extension of Hamilton's variational principle to those problems is also made. Theoretically, the work is based on the change-of-variable technique familiar from the theory of partial differential equations. The extensions may be utilized in developing new approximation procedures for the kind of problems considered here. As far to the author's knowledge, the theoretical ideas of the paper are brand new and have not been published in the literature before.

1. INTRODUCTION

The eigenvalue problem associated with nonconservative mechanical problems has been the concern of numerous studies during the last years. It would be injustice omitting to mention the numerous papers of Leipholz [e.g. 1,2,3,4,5,6] where a firm foundation for the concept of *generalized self-adjointness* is laid. As application examples, he considers systems which are truly nonconservative in the classical sense, and shows that certain systems are yet conservative with respect to functionals obtained through the use of the concept of generalized self-adjointness. Furthermore, he proves that there is an extension of Hamilton's variational principle to corresponding systems. One should note that the main shortcoming of Leipholz' otherwise distinguished work is that the velocity-dependent terms are almost totally omitted from the differential equation in the references op.cit..

Walker [7] as well as Inman & Olsen [8] include the velocity-dependent terms but limit their consideration to boundary conditions corresponding the pin-ended beams. Very interesting and useful results may then be derived which, however, are not feasible e.g. in the case of cantilevered fluid conveying pipes.

Tonti [9,10] gave a set of variational formulations for every (nonlinear) problem (sic!). Tonti's work is based on replacement of the "canonical" bilinear form with a more general one. Through the change of the bilinear form, the so called *extended variational formulations* are obtained. Alliney & Tralli [11] apply the theory presented by Tonti and give an extended variational formulation for the Beck's problem. In a certain sense, their formulation agrees with Leipholz formulations, but is little more unrestricted (in the spirit of Tonti). As to my opinion, Tonti's work is not very feasible for the problems considered in this paper.

One more different[1] viewpoint to the non-selfadjoint problems can be found in the paper of Auchmuty [12]. Combining his work with the results published by Tretter [13] gives good opportunities to treat a wide variety of eigenvalue problems. Unfortunately, the problems with velocity-dependent terms in the differential equation become little cumbersome, which is a drawback in their works.

In this paper, a new point of view is taken by applying a change-of-variable technique familiar from the theory of partial differential equations. This makes it possible to tackle e.g. the problem of cantilevered pipe conveying fluid and show that the problem becomes conservative with respect to a certain functional. Also, an extension of Hamilton's variational principle associated with the problem becomes possible. Furthermore, some statements concerning the completeness of eigenfunctions are easily made. The proposed method may offer a new basis even for the numerical solution procedures.

2. BACKGROUND

The linearized mechanical systems considered here can be mathematically described by a partial differential equation of the form

$$\ddot{w} + L_1\dot{w} + L_2 w = 0 \text{ on } \Omega \tag{1}$$

with boundary conditions

$$Bw = 0 \text{ along } \partial\Omega, \tag{2}$$

where $w = w(x,t)$ denotes the deflection in Ω. Problems with one spatial dimension only are involved here, thus $\tilde{x} \in \Omega = (0,1)$ with the boundary $\partial\Omega$ consisting of points $\tilde{x} = 0$ and $\tilde{x} = 1$. The dot denotes time differentiation, and L_1, L_2 are linear spatial differential operators. B is a linear spatial differential operator reflecting the boundary conditions.

[1] In fact, Auchmuty's work has a lot of similarities with [8].

The above boundary eigenvalue problem covers a wide range of linearized mechanical systems. Because of the preliminary nature of this study, a restricted subclass of (1) will be considered here with L_1, L_2 defined by

$$L_1 u = 2\beta\mu \frac{\partial u}{\partial \tilde{x}} \tag{3}$$

$$L_2 u = \mu^2 u'' + u'''' . \tag{4}$$

Then, the basic differential equation (e.g. Chen [14]) in a dimensionless form for a fluid conveying pipe is achieved as

$$\ddot{w} + 2\beta\mu\dot{w}' + \mu^2 w'' + w'''' = 0, \tag{5}$$

where β, μ are dimensionless parameters, β taking care of the mass of both the pipe and the fluid and, μ corresponds to the fluid velocity.

The boundary conditions are taken as those of a cantilevered pipe, i.e.

$$w(0,\tilde{t}) = 0 = w'(0,\tilde{t}) \tag{6}$$

$$w''(1,\tilde{t}) = 0 = w'''(1,\tilde{t}) . \tag{7}$$

The differential equation (5) with Coriolis-effect together with the boundary conditions (6), (7) describes a truly *nonconservative* mechanical system. In the language of mathematics, the corresponding expression were *non-selfadjoint*.

3. CHANGE OF INDEPENDENT VARIABLES

Let us write the differential equation (5) in the form $Lw + w'''' = 0$ using the operator L defined by

$$Lw = \ddot{w} + 2\beta\mu\dot{w}' + \mu^2 w'' . \tag{8}$$

The parameter β always satisfies $\beta < 1$ (see e.g. [14]). Then,

$$(\beta\mu)^2 - \mu^2 = (\beta^2 - 1)\mu^2 < 0, \tag{9}$$

which shows that L is an elliptic operator. This suggests the following change of independent variables[2],

$$\begin{cases} \xi(\tilde{x},\tilde{t}) = -\dfrac{1}{\theta}(\tilde{x} - \beta\mu\tilde{t}) \\ \\ \tau(\tilde{x},\tilde{t}) = \tilde{t} \end{cases} \tag{10}$$

where the shorthand notation θ is defined by $\theta^2 = (1-\beta^2)\mu^2$.

Substitution of (10) into the deflection function w allows one to write an identity

$$w(\tilde{x},\tilde{t}) \equiv v(\xi(\tilde{x},\tilde{t}),\tau(\tilde{x},\tilde{t})). \tag{11}$$

Partial differentiation on the both sides of the identity with respect to $\tilde{x}$ and $\tilde{t}$ as many times as needed in (5) gives the differential equation for v

$$\frac{\partial^2 v}{\partial \tau^2} + \frac{\partial^2 v}{\partial \xi^2} + \frac{1}{\theta^4}\frac{\partial^4 v}{\partial \xi^4} = 0. \tag{12}$$

Denote $\xi_0 \equiv \xi(0,\tilde{t})$, $\tau_0 \equiv \tau(0,\tilde{t})$, $\xi_1 \equiv \xi(1,\tilde{t})$ and $\tau_1 \equiv \tau(1,\tilde{t})$. The boundary conditions (6,7) now appears as

$$v(\xi_0,\tau_0) = 0 = \frac{\partial v}{\partial \xi}(\xi_0,\tau_0) \tag{13}$$

$$\frac{\partial^2 v}{\partial \xi^2}(\xi_1,\tau_1) = 0 = \frac{\partial^3 v}{\partial \xi^3}(\xi_1,\tau_1). \tag{14}$$

Although the equations (12-14) already give the necessary problem transformation into the form that accepts certain conservativeness result formulation as well as a possibility to extend Hamilton's variational principle to the nonconservative system involved, let us yet make a second change of variables (a kind of aesthetic change) affecting mainly on the boundary conditions.

Let the new variables x and t be defined by

---

[2] Let us denote that the change of variables is not unique but there are several (obviously infinitely many) possibilities. The only demand is that the new variables satisfy a certain set of first order partial differential equations which will not be reproduced here. The set of equations can be found in almost any classical treatment of partial differential equations.

$$\begin{cases} x(\xi,\tau)=\xi_1-\xi \\ \\ t(\xi,\tau)=\tau \end{cases} \tag{15}$$

Write an identity

$$v(\xi,\tau)\equiv u(x(\xi,\tau),t(\xi,\tau)) \tag{16}$$

and make the necessary differentiations on both sides of equation (16) as well as the substitutions into the boundary conditions, thus obtaining

$$\ddot{u}+u''+\frac{1}{\theta^4}u''''=0 \tag{17}$$

$$u(-\theta^{-1},\tilde{t})=0=u'(-\theta^{-1},\tilde{t}) \tag{18}$$

$$u''(0,\tilde{t})=0=u''''(0,\tilde{t}), \tag{19}$$

where the dot denotes differentiation with respect to t and the prime with respect to x. Next, we turn to the main subjects of the study.

4. ON THE COMPLETENESS OF "EIGENFUNCTIONS"

It is not very difficult to prove the completeness of eigenfunctions in certain Sobolev-like spaces for the problem (17-19), but as the proof is rather lengthy it will be omitted here. As a reference for the proof it is worth to mention again the paper by Tretter [13] which gives the necessary tools.

5. GENERALIZED SELF-ADJOINTNESS AND CONSERVATIVITY

Define an operator T by the relation

$$T\phi=\phi''\left(-\frac{1}{\theta}-x,t\right). \tag{20}$$

Instead of the "canonical" bilinear form (ψ,ϕ) we define a more useful bilinear form with respect to operator T by

$$\langle\psi,\phi\rangle=(\psi,T\phi)=\int_{-\frac{1}{\theta}}^{0}\psi(x,t)\phi''(-\theta^{-1}-x,t)\,dx \tag{21}$$

Through a direct calculation[3], it can be shown that the following results holds true:

$$\langle \psi, \phi \rangle = \langle \phi, \psi \rangle \tag{22}$$

$$\langle \psi'', \phi \rangle = \langle \phi'', \psi \rangle \tag{23}$$

$$\langle \psi'''', \phi \rangle = \langle \phi'''', \psi \rangle . \tag{24}$$

This means that the eigenvalue problem associated with equations (17-19) is self-adjoint with respect to the operator T.

Define now the functional F,

$$F = \langle \dot{u}, \dot{u} \rangle + \left\langle u'' + \frac{1}{\theta^4} u'''', u \right\rangle . \tag{25}$$

Differentiation of both sides with respect to t gives

$$\dot{F} = 2\left\langle \ddot{u} + u'' + \frac{1}{\theta^4} u'''', \dot{u} \right\rangle .$$

Application of (17) justifies that

$$\dot{F} = 0 \quad \Rightarrow \quad F = \text{constant}. \tag{26}$$

Obviously we may say that also the original system (5-7) is conservative in some (higher than the Leipholz' higher) sense. The exact conclusions will be made in a later paper.

Consider the functional H,

$$H = \int_{t_1}^{t_2} \left\{ \langle \dot{u}, \dot{u} \rangle - \left\langle u'' + \theta^{-4} u'''', u \right\rangle \right\} dt . \tag{27}$$

Exploiting the standard techniques of variational calculus and the above defined generalized self-adjointness, the first variation of H assumes the form

$$\delta H = 2 \int_{t_1}^{t_2} \left\{ \langle \dot{u}, \delta \dot{u} \rangle - \left\langle u'' + \theta^{-4} u'''', \delta u \right\rangle \right\} dt . \tag{28}$$

[3] For the reference, see e.g. Leipholz [3].

Applying integration by parts to the first bilinear form and imposing on the variation δu the common conditions

$$\delta u(t_1) = 0 = \delta u(t_2) \tag{29}$$

one achieves at

$$\delta H = -2\int_{t_1}^{t_2} \langle \ddot{u} + u'' + \theta^{-4} u'''', \delta u \rangle \, dt \,. \tag{30}$$

Hence, as the differential equation for the problem is satisfied, it follows that

$$\delta H = 0 \,. \tag{31}$$

But this means that in the space of admissible variations the functional H attains its stationary value.

6. CONCLUSIONS AND REMARKS

It has been shown that the dynamics of a cantilevered fluid conveying pipe obeys some conservativity features. The study also shows that an extension of Hamilton's variational principle is in some sense applicable for the problem.

Remark 1. Although the problem becomes self-adjoint with respect to a certain operator, it still remains indefinite. Thus, Rayleigh quotient cannot be applied in the eigenvalue problem solution.

Remark 2. In the study, the operator L_1 was chosen according to equation (3). The above described technique can also be applied in the case of operator L_1 defined by $L_1\dot{u} = \alpha\dot{u} + 2\beta\mu\dot{u}'$ (i.e. external damping included). But, if e.g. Kelvin-Voigt material model is assumed, then such a simple change-of-variable technique described above cannot be found.

REFERENCES

1. H. Leipholz, *On a Generalization of the Concepts of Self-adjointness and of Rayleigh's Quotient*, Mech. Res. Comm. **1** (1974), 67-72.
2. H. Leipholz, *On Conservative Elastic Systems of the First and Second Kind*, Ing. Arch. **43** (1974), 255-271.
3. H. Leipholz, *On an Extension of Hamilton's Variational Principle to Nonconservative Systems which are Conservative in a Higher Sense*, Ing. Arch. **47** (1978), 257-266.

4. H. Leipholz, *Variational Principles for Non-conservative Problems, a Foundation for a Finite Element Approach*, Comput. Meths. Appl. Mech. Engrg. **17/18** (1979), 609-617.
5. H. Leipholz, *On Principles of Stationarity for Non-selfadjoint Rod Problems*, Comput. Meths. Appl. Mech. Engrg. **59** (1986), 215-226.
6. H. Leipholz, *Stability of Elastic Systems*, Sijthoff and Noordhoff, Alphen aan den Rijn (1980).
7. J.A. Walker, *Energy-like Liapunov Functionals for Linear Elastic Systems on a Hilbert Space*, Q. Appl. Math. **30** (1973), 465-480.
8. D.J. Inman and C.L. Olsen, *Dynamics of Symmetrizable Nonconservative Systems*, J. Appl. Mech. **55** (1988), 206-212.
9. E. Tonti, *A General Solution of the Inverse Problem of the Calculus of Variations*, Hadronic J. **5** (1982), 1404-1450.
10. E. Tonti, *Variational Formulation for Every Nonlinear Problem*, Int. J. Engng. Sci. **22** (1984), 1343-1371.
11. S. Alliney and A. Tralli, *"Extended" Variational Formulations and F.E. Models in the Stability Analysis of Non-Conservative Elastic Systems*, Trans. Can. Soc. Mech. Eng. **10** (1986), 107-114.
12. G. Auchmuty, *Variational Principles for Eigenvalues of Compact Operators*, SIAM J. Math. Anal. **20** (1989), 1321-1335.
13. C. Tretter, *Nonselfadjoint Spectral Problems for Linear Pencils $N-\lambda P$ of Ordinary Differential Operators with λ-Linear Boundary Conditions: Completeness Results*, Integr. Equat. Oper. Th. **26** (1996), 222-248.
14. S.-S. Chen, *Flow-Induced Vibration of Circular Cylindrical Structures*, Springer-Verlag, Berlin (1987).

TYÖKONEEN KULJETTAJAN AKTIIVISEEN VÄRÄHTELYN VAIMENNUKSEEN TARKOITETTU ISTUIMEN RIPUSTUSMEKANISMIN SIMULOINTIMALLI MATLAB/SIMULINK-YMPÄRISTÖSSÄ

Matti Kangaspuoskari

Oulun yliopisto, Konetekniikan osasto
PL 444, 90571 Oulu

TIIVISTELMÄ

Metsätraktoreiden ja erilaisten maansiirtotöissä käytettävien työkoneiden kuljettajiin kohdistuu työskentelyn aikana matalataajuista (0-10 Hz) suuriamplitudista tärinää. Pysty- ja pituussuuntainen värähtely ei ole kovin haitallista kuljettajalle koska ihmisen keho kestää varsin hyvin nämä rasitukset. Sen sijaan sivusuuntainen heilahtelu on ongelmallinen. Kuljettajan selkärankaan kohdistuu voimakas rasitus joka ajan myötä alentaa työtehoa ja aiheuttaa väsymystä.

Värähtelyongelmien eliminoiminen on tapahtunut perinteisesti passiivisilla vaimentimilla joilla systeemistä poistetaan energiaa. Teoriassa värähtelyongelmien eliminoiminen on mahdollista myös aktiivisesti ohjatuilla järjestelmillä. Aktiivisessa vaimentimessa systeemiin tuodaan lisää energiaa, joka aikaansaa värähtelyä kumoavan vastavoiman. Helpointen toteutettavissa on matalataajuisen jäykän kappaleen liiketilan hallitseminen.

Tässä työssä on luotu aktiiviseen värähtelyjen vaimennukseen soveltuvan istuimen ripustusmekanismin ja sen säätöalgoritmin simulointimalli *MATLAB/SIMULINK-ympäristössä [1]. Valitulla kahden vapausasteen mekanismilla on mahdollista vaimentaa yleistä tasoliikettä. Simulointimallilla voidaan simuloida kuljettajan liiketilaa, kun työkoneeseen kohdistuu tunnettu heräte. Edelleen simulointimallilla on mahdollista simuloida ja kehittää aktiivivaimentimen säätöalgoritmeja [2].

1. JOHDANTO

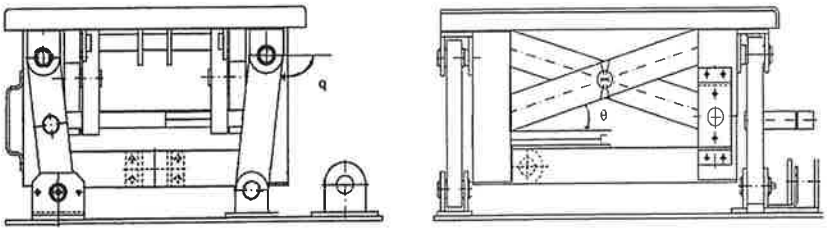
Metsätraktorien kuljettajiin kohdistuvan matalataajuisen tärinän aiheuttaja on maaston epätasaisuus. Tärinän voimakkuuteen vaikuttavat lisäksi ajonopeus, kuormaus ja kuljettajan ajotapa. Maastossa kulkeva traktori heilahtelee sekä pysty-, sivu- että pituussuunnassa. Pituus- ja pystysuuntaisen heilahtelun ihminen kestää varsin hyvin. Sivusuuntainen heilahtelu on ongelmallinen ja rasittaa erityisesti kuljettajan selkärankaa. Tämä aiheuttaa alentunutta työtehoa ja väsymystä.

Perinteinen ratkaisu värähtelyjen hallinnassa on kuljettajan istuimen eristäminen alustastaan jousen ja vaimentimen muodostamalla passiivisella vaimentimella. Tuloksen on mekaaninen alipäästösuodin, joka vaimentaa rajataajuutta korkeampia taajuuksia. Samalla suodin kuitenkin vahvistaa rajataajuutta matalampia taajuuksia. Rajataajuutta laskemalla yhä suurempi osa värähtelystä saadaan vaimennuksen piiriin, mutta samalla järjestelmän jäykkyys pienenee ja sitä kautta kuljettajan suhteellinen siirtymä ohjaamon suhteen kasvaa [3].

Aktiivisessa vaimentimessa käytetään ulkoista energiaa värähtelyjen vaimentamiseen. Tyypillisesti aktiivinen vaimennusjärjestelmä on servomekanismi, joka antureiden ohjaamana tuottaa alustaan nähden vastakkaisen liikkeen. Pelkästään takaisinkytkentään perustuva järjestelmä asettaa toimilaitteelle ja sen ohjaukselle suuret vaatimukset. Eräs kompromissi aktiivisen ja passiivisen järjestelmän välillä on ns. hitaaasti toimiva aktiivijousitus. Systeemissä siirtymää tuottava toimilaite on kytketty sarjaan passiivisen vaimentimen kanssa. Takaisinkytkennän avulla ohjataan toimilaitetta ja vaimennetaan passiivisen vaimentimen rajataajuuta alhaisempi värähtely. Koska aktiivinen vaimennin toimii vain matalilla taajuuksilla, ei toimilaitteelle ja sen ohjaukselle aseteta kovin suuria vaatimuksia [4].

2. ISTUIMEN RIPUSTUSMEKANISMI

Henry Jouppi on diplomityössään [5] esitellyt aktiiviseen värähtelynvaimennuksen soveltuvan kahden vapausasteen mekanismin jolla on mahdollista vaimentaa yleistä tasoliikettä.

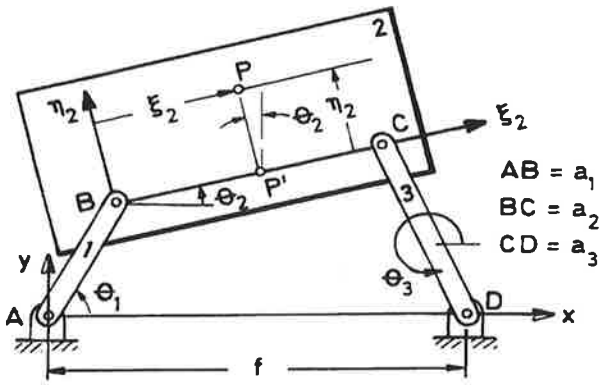


Kuva 1: Istuimen ripustusmekanismi.

Mekanismi muodostuu nelinivelmekanismista ja sen kanssa sarjaan kytketystä saksimekanismista. Nelinivelmekanismi tekee ympyränkaariliikettä jonka rotaatiokeskipiste on maanpinnan alapuolella sijaitsevassa traktorin kuvitteellisessa rotaatiokeskipisteessä. Saksimekanismi on suoravientimekanismi jolla vaimennetaan pystyliikettä. Näin muodostetulla kahden vapausasteen mekanismilla on mahdollista vaimentaa yleistä tasoliikettä.

3. KINEMATIIKKA

Tarkastellaan kuvan 2 mukaista kahden vapausasteen tasomekanismia jossa nelinivelmekanismin päällä on suoravientimekanismi.



Kuva 2: Ripustusmekanismin kinemaattinen malli.

Nelinivelmekanismille on voimassa karteesisessa Axy-koordinaatistossa yhtälö [6]

$$\begin{aligned} a_1 \cos\theta_1 + a_2 \cos\theta_2 + a_3 \cos\theta_3 - f &= 0 \\ a_1 \sin\theta_1 + a_2 \sin\theta_2 + a_3 \sin\theta_3 &= 0 \,. \end{aligned} \tag{1}$$

Valitaan nelinivelmekanismille vapausasteeksi vaakatasosta mitattu kulma $\theta_3 = q$. Kulmat θ_1 ja θ_2 voidaan ratkaista yhtälöstä (1) *MATLABin FSOLVE-funktiolla. Kulmien nopeudet ja kiihtyvyydet saadaan derivoimalla yhtälöä (1) ajan suhteen.

Saksimekanismin pisteelle P voidaan kirjoittaa

$$\begin{aligned} x_p &= a_1 \cos\theta_1 + \xi_2 \cos\theta_2 - \eta_2 \sin\theta_2 \\ y_p &= a_1 \sin\theta_1 + \xi_2 \sin\theta_2 + \eta_2 \cos\theta_2 \,. \end{aligned} \tag{2}$$

Valitaan saksimekanismille vapausasteeksi nelinivelmekanismin suhteen mitattu kulma Θ. Jos saksimekanismin sauvan pituus on L voidaan siis yhtälöön (2) sijoittaa $\eta_2 = L \sin\theta$. Pisteen P nopeus ja kiihtyvyys saadaan derivoimalla yhtälöä (2) ajan suhteen.

Nyt pisteen P liiketila, ts. paikka, nopeus ja kiihtyvyys on tunnettu.

4. LIIKEYHTÄLÖ

Etsitään liikeyhtälö käyttäen Lagrangen yhtälöitä vapaassa koordinaatistossa [7]

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} = Q_i , \qquad i = 1,2 . \tag{3}$$

Valitaan nelinivelmekanismille vapausasteeksi vaakatasosta mitattu kulma $q_1 = q$ ja saksimekanismille nelinivelmekanismin suhteen mitattu kulma $q_2 = \theta$. Q_i on vastaavassa toimilaitteessa vaikuttava yleistetty voima koordinaatin i suuntaan. Kokonaisliike-energia T on tyypillisesti muotoa

$$T(q_i,\dot{q}_i) = \frac{1}{2} f(q_i)\dot{q}_i^2 . \tag{4}$$

Sijoittamalla yhtälöön (1) saadaan

$$\begin{aligned} A\ddot{\theta} + B\ddot{q} + C\dot{q}^2 + D\dot{\theta}\dot{q} + E\dot{\theta}^2 &= Q_1 \\ F\ddot{\theta} + G\ddot{q} + H\dot{q}^2 + I\dot{\theta}\dot{q} + J\dot{\theta}^2 &= Q_2 , \end{aligned} \tag{5}$$

missä funktiot A,B,... J ovat funktioita koordinaateista q ja θ.

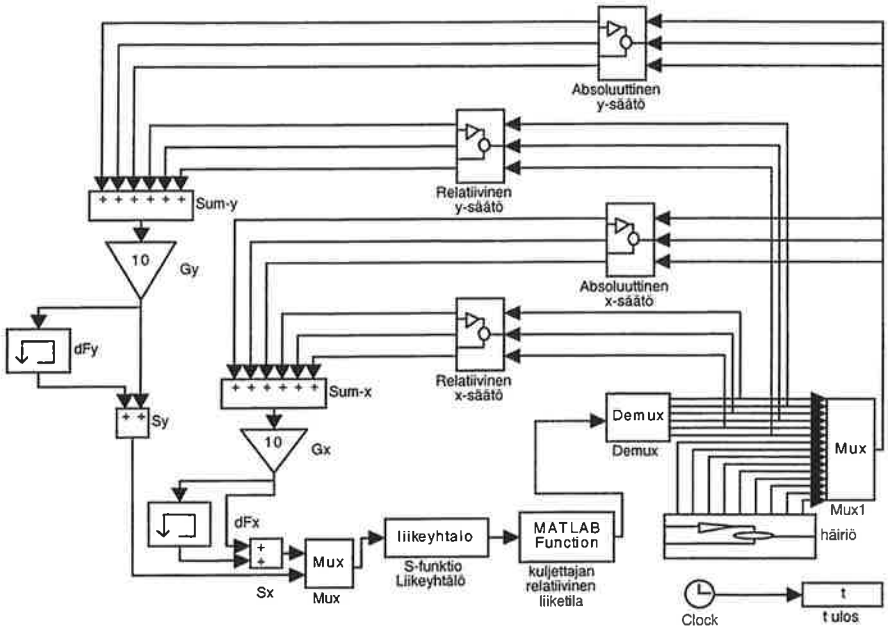
4. MEKANISMI *MATLAB/SIMULINK-YMPÄRISTÖSSÄ

Mekanismin simulointiin on käytetty *MATLAB/SIMULINK-ohjelmistoa. Liikeyhtälö (3) on voimakkaasti epälineaarinen, joten perinteiset integraalimuunnoksiin perustuvat menetelmät siirtofunktion löytämiseksi ovat vaikeita toteuttaa. Mekanismin simuloinnissa käytetään liikeyhtälön numeerista aikaintegrointia lähtien tunnetusta alkutilasta. Valmiita integrointi-algoritmeja löytyy *MATLAB-ohjelmistosta. Simulointia varten liikeyhtälö muunnetaan vektorimuotoiseksi ensimmäisen kertaluvun differentiaaliyhtälöksi. Tästä tilamuuttujaesityksestä kirjoitetaan edelleen ns. S-funktio, joka on muotoa

$$\begin{aligned} \dot{x} &= \underline{A}x + \underline{B}u \\ y &= \underline{C}x + \underline{D}u . \end{aligned} \tag{6}$$

Mekaniikan tehtävässä yleensä valitaan minimimäärä tilamuuttujia. Nyt tilamuuttujat vektorissa x esittävät vapausasteiden kulmaa ja kulmanopeutta. Vektorissa u ovat mekanismin toimilaitteissa vaikuttavat voimat. Funktio tulostaa vektorisssa y mekanismin liiketilan, ts. valittujen vapausasteiden kulman, kulmanopeuden ja kulmakiihtyvyyden. $\underline{A}$, $\underline{B}$, $\underline{C}$ ja $\underline{D}$ ovat kerroinmatriiseja.

*MATLABin laajennus *SIMULINK on dynaamisten systeemien simulointiin kehitetty ohjelma. *SIMULINK avaa *MATLAB-istuntoon erityisen lohkokaavioikkunan. Tutkittavan systeemin malli luodaan lohkokaavioikkunaan yksinkertaisesti vetämällä hiirellä `ikoneita` valikoista.



Kuva 3: Simulointimalli *MATLAB/SIMULINK-ympäristössä.

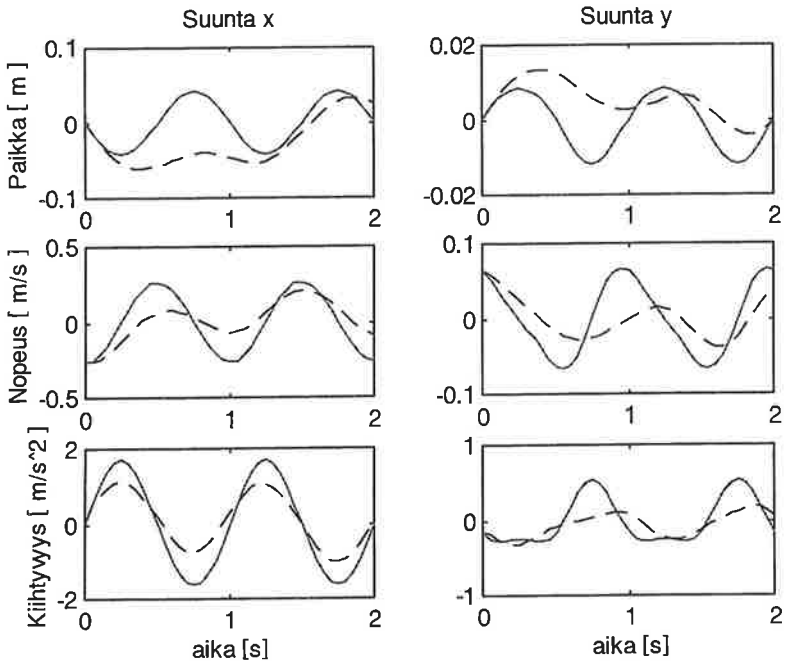
Kuvan 3 mallissa S-funktio `Liikeyhtälö` sisältää yhtälön (6) esittämän tilamallin. Kun mekanismin liiketila tunnetaan saadaan pisteen P liiketila kinemaattisilla yhtälöillä (2) (ja sen aikaderivaatoilla). Yhtälöt kirjoitetaan *MATLABin .m-funktioksi ja niitä kutsutaan funktiossa `kuljettajan relatiivinen liiketila`.

Häiriö tuodaan systeemiin Axy-koordinaatiston yleisenä tasoliikkeenä maahan nähden. Tällöin edellä kuvatut kinemaattiset yhtälöt kuvaavat pisteen P relatiivista liiketilaa liikkuvan Axy-koordinaatiston suhteen. Pisteen P absoluuttinen liiketila saadaan yleisesti tunnettujen relatiivisen liikkeen kinematiikkaa esittävien yhtälöiden avulla [8]. Liikeyhtälöä kirjoitettaessa Axy-koordinaatisto oletettiin inertiaalikoordinaatistoksi.

Kuvan 3 mallissa mekanismiin on liitetty säädin jossa pisteen P vaakasuoraa liiketilaa säädetään nelinivelmekanismilla ja pystysuoraa liiketilaa saksimekanismilla. Mallin säädin perustuu takaisinkytkentää jossa aiheutetaan erosuureeseen verrannollinen muutos toimilaitteessa vaikuttavaan voimaan. Erosuure voi olla lineaarinen funktio pisteen P absoluuttisesta ja relatiivisesta liiketilasta. Käytännön sovellutuksissa ilmeisen hyödyllinen on erosuure joka perustuu kuljettajan absoluuttiseen kiihtyvyyteen ja relatiivisen paikkaan.

5. TULOKSET

Tulosten käsittely on vaivatonta *MATLAB-ympäristössä. Tulosten graafista esittämistä varten on kirjoitettu *MATLABin .m-funktio. Tässä on tulostettu samaan kuvaan kuljettajan absoluuttinen liiketila mekanismin ollessa lukittu ja aktiivisen säädön ollessa päällä. Häiriönä on Axy-koordinaatiston 5° harmooninen rotaatio origonsa A ympäri.



Kuva 4: Kuljettajan absoluuttinen liiketila, häiriö 5° rotaatio pisteen A ympäri.
'____' Mekanismi lukittu, '-----' aktiivinen vaimennus.

Säätöparametrien optimiratkaisu sisältää kompromissin toisaalta kuljettajaan kohdistuvan kiihtyvyyden ja toisaalta ohjaamon suhteen tapahtuvan siirtymän suhteen. Lähteessä [9] on esitetty optimiratkaisuna säädintä joka minimoi neliöllisen suorituskykyindeksin Π. Laajennetaan suorituskykyindeksi yleiseen tasotapaukseen laskemalla vaaka- ja pystysuuntaiset osuudet yhteen

$$\Pi = \ddot{x}_a^2 + \ddot{y}_a^2 + \rho\,(x_r^2 + y_r^2)\ . \qquad (7)$$

Kuljettajan 'mukavuus' on siis funktio absoluuttisesta kiihtyvyydesta ja relatiivisesta paikasta ohjaamoon nähden. Painokerrointa ρ voitaisiin käyttää adaptiivisessa säädössä tai kuvaamaan eri kuljettajien käsitystä hyvästä vaimennuksesta. Tässä on käytetty arvoa $\rho = 1$.

Käsiteltävässä esimerkissä kuljettajan x-suuntainen kiihtyvyys pienenee 36 %, y-suuntainen kiihtyvyys pienenee 37 % ja mukavuusindeksi paranee 46 %.

KIITOKSET

Tekijä kiittää Prof. Kalervo Nevalaa ja Prof. Tatu Leinosta heidän järjestämästään rahoituksesta joka mahdollisti neljän kuukauden täysipäiväisen työskentelyn ongelman parissa.

LÄHTEET

1. Simulink User's Guide. The Math Works Inc. (1992).

2. K. Nevala, M. Kangaspuoskari and T. Leinonen, *Development of an active a suspension mechanism for the seat vibration damping,* - Proceedings of the Fourth IASTED International Conference, Robotics and Manufacturing, August 19-22, (1996), Honolulu, Hawaii, USA. Ed. by R.V. Mayorga. Anaheim, USA (1996), s. 337-339.

3. V. Iivonen, *Työkoneen istuimen aktiivinen vaimennus sähköhydraulisella servojärjestelmällä*, Sähköhydrauliset servojärjestelmät II. Tampereen teknillinen korkeakoulu. Opetusmoniste **32**. (1977).

4. P.W. Claar II and J.M. Vogel, *A Review of Active Suspension Control for On and Off-Highway Vehicles*, SAE Transactions, Vol. **98** (1989), No.2, s. 557-568.

5. H. Jouppi. *Ripustusmekanismin kehittäminen työkoneen istuimen aktiiviseen värähtelyn vaimennukseen. Diplomityö*. Oulun Yliopisto (1996).

6. P. Burton, *Kinematics and dynamics of planar machinery*. Prentice-Hall, Inc., Englewood Cliffs, N.J. (1979).

7. T. Salmi, *Mekaniikka 3. Kinetiikka*. Kustannusyhtymä. Tampere (1980).

8. T. Salmi. *Mekaniikka 2. Kinematiikka*. Kustannusyhtymä.Tampere 1979.

9. J.K.Hedrick. *Active suspensions for ground transport vehicles - a state of the art rewiew. Mechanics of transportation Suspension Systems*, Trans. ASME AMD-Vol **15** (1975).