62 /2/2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS

WARSAW 2024, QUARTERLY, VOLUME 62, INDEX 365238, ISSN 1429-2955

POLISH SOCIETY OF THEORETICAL AND APPLIED MECHANICS ptmts

POLISH SOCIETY OF THEORETICAL AND APPLIED MECHANICS

JOURNAL OF THEORETICAL AND APPLIED MECHANICS

No. 2 • Vol. 62

Quarterly

WARSAW, APRIL 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
(until 1997 Mechanika Teoretyczna i Stosowana, ISSN 0079-3701)

Beginning with Vol. 45, No. 1, 2007, *Journal of Theoretical and Applied Mechanics* (JTAM) has been selected for coverage in Thomson Reuters products and custom information services. Now it is indexed and abstracted in the following:

- **Science Citation Index Expanded** (also known as SciSearch®)
- **Journal Citation Reports/Science Edition**

* * * * *

Editorial Office

Al. Armii Ludowej 16, room 650; 00-637 Warsaw, Poland
phone (+48) 664 099 345, e-mail: biuro@ptmts.org.pl

www.jtam.pl

* * * * *

Rozwój kwartalnika naukowego *Journal of Theoretical and Applied Mechanics*, ISSN 1429-2955, jest dofinansowany ze środków Ministra Edukacji i Nauki przyznanych z pomocy *de minimis* w ramach programu „Rozwój czasopism naukowych", umowa RCN/SN/0056/2021/1. Niniejszy numer został sfinansowany przez ministerstwo w ramach projektu: Doskonała nauka – Wsparcie konferencji naukowych. V Polski Kongres Mechaniki-25 Międzynarodowa Konferencja Metod Komputerowych Mechaniki, umowa DNK/SN/548136/2022.

FROM THE EDITORS

It is our honour to present the special issue (Vol. 62/2024, No. 2) of *Journal of Theoretical and Applied Mechanics* (*JTAM*). It is devoted to the 5th Polish Congress of Mechanics which was organized jointly with the 25th International Conference on Computer Methods in Mechanics (together as the PCM-CMM 2023) on September 4-7, 2023 in Gliwice, Poland.

The idea of the Polish Congress of Mechanics (PCM) dates back to the 21st International Congress of Theoretical and Applied Mechanics (ICTAM) that took place in 2004 in Warsaw, Poland. This very successful event was jointly organized by the Warsaw University of Technology and the Institute of Fundamental Technological Research of the Polish Academy of Sciences. It was chaired by professor Witold Gutkowski (†2019), Honorary Member of the Polish Society of Theoretical and Applied Mechanics (PSTAM). Apart of its international importance, ICTAM 2004 was a breakthrough event for the Polish community of mechanical engineering science. Following this, the initiative of PSTAM, together with joint organizational effort of the entire community, resulted in the first Polish Congress of Mechanics that took place in 2007 in Warsaw. Its success led to organizing its subsequent editions which were then held in four-year intervals cycle: 2011 in Poznań, 2015 in Gdańsk, 2019 in Cracow and the most recent – 2023 in Gliwice. Starting from 2015, the Congress has been organized as a joint event with the biennial International Conference on Computer Methods in Mechanics (CMM) and jointly supervised by PTMTS and the Polish Society of Computer Methods in Mechanics (PTMKM). This partnership the Congress international status and raise its scientific rank.

The 2023 edition of this joint event gathered 304 participants. The conference programme included 10 plenary lectures and 225 participants presentations selected by the Scientific Committee.

This post-congress issue of *JTAM* contains 20 articles reflecting research results presented at PCM-CMM 2023, either as plenary lectures or in other forms of presentation. Their topics include solid mechanics, structural dynamics and control, damage analysis, coupled fields issues; many of them treated with particular attention paid to advanced numerical methods supporting the solution of problems analysed. All of the contributions have been reviewed within the standard editorial procedure of the Journal.

We would like to thank all the authors of articles in this issue for submitting their post-congress papers and thus contributing to the publishing output of *JTAM*. We are also grateful to the Ministry of Science and Higher Education of Poland for financial support of organization of the congress and publication of the selected articles.

Piotr Kowalczyk
Editor-in-Chief of *JTAM*

Włodzimierz Kurnik
Vice-President of PCM-CMM 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 1, pp. 193-206, Warsaw 2024
https://doi.org/10.15632/jtam-pl/183493

ABILITY OF LOCALIZING GRADIENT DAMAGE TO DETERMINE SIZE EFFECT IN CONCRETE BEAMS[1]

Adam Wosatko, Jerzy Pamin, Andrzej Winnicki
Cracow University of Technology, Faculty of Civil Engineering, Cracow, Poland
e-mail: adam.wosatko@pk.edu.pl; jerzy.pamin@pk.edu.pl; andrzej.winnicki@pk.edu.pl

The objective of the paper is to demonstrate the potential of the localizing gradient damage model in size effect simulations. Three different gradient activity functions for variable internal length scale are considered. Numerical simulations for an unnotched beam under three-point bending are referred to the experiment performed by Grégoire *et al.* (2013). A confrontation with the conventional gradient damage model as well as mesh sensitivity studies are also presented. It is proved that the localizing gradient damage model with different variants of the gradient activity function can reproduce the size effect quite reasonably.

Keywords: size effect, concrete, localizing gradient damage, finite element method

1. Introduction

The size effect is connected with a change of the material response, which is observed for structural elements or laboratory specimens with different volumes. In quasi-brittle materials like concrete, it is observed that the nominal strength decreases when the size of the considered specimen enlarges. An analogical relation is also observed for equilibrium paths in the post-peak regime, taking into account material brittleness. In fact, the size of the fracture process zone (FPZ) in quasi-brittle materials like concrete does not correspond to the dimensions of structural elements, instead it is related to the material length scale. The main cause of the size effect is deterministic and related to the rate of energy dissipation in FPZ and evolving cracks, see e.g. (Bažant and Planas, 1998).

The above statements are proven in many experiments, hence the deterministic size effect is one of crucial features examined for quasi-brittle materials. Together with the development of fracture and damage theories, the knowledge about the size effect laws has also been improved, see e.g. (Bažant and Le, 2017; Bažant and Planas, 1998). When the size effect is analyzed numerically, standard local models are not able to capture it properly. Therefore, correct computational models for concrete should be equipped with a localization limiter, i.e. contain an internal length scale. There are several approaches to ensure mesh-objective results for continuum models. The first option, followed in this paper, is to use a non-local formulation via integral or gradient-type averaging. The second concept is to introduce a rate-dependence into the constitutive relation. The simplest approach is the crack band theory, proposed first by Bažant and Oh (1983), which however is not a proper localization limiter since it alleviates only the mesh sensitivity of load--displacement diagrams. A complementary overview of these issues can be found, for instance, in (Bažant and Jirásek, 2002).

In this paper, the damage model is enhanced by the presence of higher-order gradients via an averaging equation in the formulation based on continuum damage mechanics. The gradient

[1]Paper presented during PCM-CMM 2023

damage model was first suggested by Peerlings *et al.* (1996). In the description of finite elements (FEs), two types of degrees of freedom are distinguished, i.e. an averaged strain measure is approximated next to the standard displacement field. The zone of localization represents concrete cracking in the model and it is controlled by a constant internal length scale. The interpretation of the internal length scale as constant in quasi-brittle materials can be connected with the maximum aggregate size as a counterpart of the width of the FPZ, see e.g. (Bažant and Planas, 1998). The conventional gradient damage (CGD) model ensures mesh-objective results, but Geers (1997) demonstrated that artificially expanding damage zones could occur, hence versions of the gradient damage model with evolving internal length scale have been proposed as more correct. In other words, the issue of spuriously widened damage zones is mitigated when the internal length scale becomes a variable. In this case, the model needs a definition of the so-called gradient activity function. The early concept is that the gradient activity increases as a function of an equivalent strain, see e.g. (Geers, 1997; Saroukhani *et al.*, 2013). However, if a localization phenomenon is observed during the loading process, then the interaction region of diffuse microcracks diminishes and tends to the formation of one macrocrack. From this point of view, the localizing gradient damage (LGD) model, proposed first by Poh and Sun (2017), where the gradient activity function decreases with damage growth, describes the change of the internal length scale in a more proper way. Nowadays, the LGD model is employed in many applications, e.g. it can be coupled with three-surface cap plasticity, as derived by Zhao *et al.* (2023) or used for simulations with impact loading, see (Wosatko, 2022).

In this paper, attention is focused on simulations of the size effect for beams subjected to three-point bending. Grégoire *et al.* (2013) performed experiments for unnotched (Type 1) and notched (Type 2) concrete beams using four sizes of specimens, and next analyzed them numerically by means of a non-local integral-type damage model. Other experimental tests of concrete beams under three-point bending were studied by Hoover *et al.* (2013), where four different sizes and five different options of notch depth were considered. Experiments for eccentrically notched beams (the notch is not located directly under the load) together with corresponding simulations using the discrete crack model with interface FEs were discussed by García-Álvarez *et al.* (2012). The aforementioned experimental research on the size effect was comprehensively verified in computations. A phase-field damage model equipped with additional approximation to regularize a crack surface was investigated by Feng and Wu (2018). An isotropic damage model with the crack width determined by the so-called Irwin's characteristic length was demonstrated by Barbat *et al.* (2020). In the model, a mixed FE formulation with the interpolation of displacement and strain fields as well as a stabilization strategy were employed. The size effect has also been explored using different versions of gradient damage models. For example, the size effect can help one to estimate characteristic parameters of the CGD model as shown by Carmeliet(1999). Size effect simulations given by Zhang *et al.* (2021) in confrontation with the experiments (Grégoire *et al.*, 2013; Hoover *et al.*, 2013) presented the applicability of the LGD model. The analysis of the energy dissipation during the loading process for the CGD and LGD models was highlighted there. The size effect can also be predicted using the stress-based LGD model (Negi *et al.*, 2021).

In this paper, the numerical analysis is limited to unnotched beams under three point bending, i.e. size effect Type 1 is simulated. The results are referred to the experiment (Grégoire *et al.*, 2013), where four different sizes of specimens were taken into account. The LGD model with different functions of gradient activity is considered and additionally compared with the CGD model. Both models are implemented by the authors in the FEAP package (Taylor, 2001). Section 2 describes briefly both versions of the gradient damage model, but definitions of the gradient activity function (including a new polynomial one) are characterized in detail. Section 3 shows the numerical analysis of the unnotched beam, where the simulation data, mesh sensitivity and size effect studies are respectively presented. The results for the LGD model with three

different gradient activity functions are discussed in the context of its ability to determine the size effect properly. A comparison with the results for the CGD model is also made. Conclusions are summarized in Section 4.

2. Overview of applied gradient damage models

2.1. Essentials of conventional gradient damage (CGD)

The standard boundary value problem (BVP) for statics is considered, where the equilibrium equation with corresponding boundary conditions is taken into account. Small strains are assumed. The model employed in this paper is based on the continuum damage mechanics theory, where in the nonlocal formulation an averaging equation with gradient terms is added to guarantee a mesh-independent solution, see (Peerlings *et al.*, 1996). The thermodynamic framework leads to weak forms of both the mentioned equations, and finally to a matrix system. More details of different variants of the gradient damage model can be found in many works, e.g. (Geers, 1997; Peerlings *et al.*, 2004; Poh and Sun 2017; Negi *et al.*, 2021; Wosatko, 2022). Below only the most crucial elements of the theory are recalled. Voigt's notation (called also matrix-vector notation) is used.

The real and effective (fictitious) configurations of a damaging body are distinguished. The concept of strain equivalence is adopted, i.e. the actual and effective strain tensors are equivalent $\boldsymbol{\epsilon} = \widehat{\boldsymbol{\epsilon}}$. The effective stress tensor $\widehat{\boldsymbol{\sigma}}$ (introduced in a vector form) affects the undamaged material skeleton, so the stress tensor $\boldsymbol{\sigma}$ corresponding to the real material is reduced by the presence of damage ω

$$\boldsymbol{\sigma} = (1-\omega)\widehat{\boldsymbol{\sigma}} \qquad\qquad \widehat{\boldsymbol{\sigma}} = \mathbf{D}\boldsymbol{\epsilon} \tag{2.1}$$

where $\mathbf{D}$ is Hooke's operator. In the model, ω is a scalar measure which changes from 0 for the undamaged material to 1 for its complete failure. This elastic stiffness degradation is a proper description for quasi-brittle materials like concrete. The damage activation function F^d is defined in the strain space

$$F^d(\boldsymbol{\epsilon}, \kappa^d) = \widetilde{\epsilon}(\boldsymbol{\epsilon}) - \kappa^d \tag{2.2}$$

where κ^d is a damage history parameter and $\widetilde{\epsilon}$ is an equivalent strain measure. The function $\widetilde{\epsilon}(\boldsymbol{\epsilon})$ describes the loading process and can be defined according to the modified von Mises formula (de Vree *et al.*, 1995)

$$\widetilde{\epsilon}(\boldsymbol{\epsilon}) = \frac{(k-1)I_1^{\boldsymbol{\epsilon}}}{2k(1-2\nu)} + \frac{1}{2k}\sqrt{\Big(\frac{k-1}{1-2\nu}I_1^{\boldsymbol{\epsilon}}\Big)^2 + \frac{12kJ_2^{\boldsymbol{\epsilon}}}{(1+\nu)^2}} \tag{2.3}$$

where $I_1^{\boldsymbol{\epsilon}}$ and $J_2^{\boldsymbol{\epsilon}}$ are strain invariants, ν is Poisson's ratio and $k = f_c/f_t$ is the ratio of uniaxial compressive and tensile strengths, which enables different responses of the concrete model in tensile and compressive regimes, even though the scalar description is employed. Damage ω is a function of the history parameter κ^d and can be defined as (Mazars and Pijaudier-Cabot, 1989)

$$\omega(\kappa^d) = 1 - \frac{\kappa_o}{\kappa^d}\Big(1 - \alpha + \alpha \mathrm{e}^{-\eta(\kappa^d - \kappa_o)}\Big) \tag{2.4}$$

where κ_o is the damage threshold. This formula holds when $\kappa^d > \kappa_o$, and then damage ω asymptotically grows to 1 according to the exponential function, which has been observed in the experiments by Hordijk (1991). The parameter α sets the level of the residual stress $(1-\alpha)E\kappa_o$, where E is Young's modulus. In this way, the total loss of material stiffness can be excluded.

The parameter η defines material brittleness in the post-peak stage and is related to concrete fracture energy G_f.

In the conventional gradient damage (CGD) model, the damage activation function defined in Eq. (2.2) takes the following form

$$F^d(\boldsymbol{\epsilon}, \kappa^d) = \overline{\epsilon}(\widetilde{\epsilon}(\boldsymbol{\epsilon})) - \kappa^d \tag{2.5}$$

and the averaged (nonlocal) strain $\overline{\epsilon}$ is a function of the equivalent strain $\widetilde{\epsilon}$ via the following differential equation (Peerlings *et al.*, 1996)

$$\overline{\epsilon} - \varphi \nabla^2 \overline{\epsilon} = \widetilde{\epsilon} \tag{2.6}$$

The BVP problem becomes regularized by the presence of $\overline{\epsilon}$ together with its second gradient in this averaging equation. For a domain $\mathcal{B}$, the natural boundary condition $\mathcal{N}^T \nabla \overline{\epsilon} = 0$ holds on the boundary $\partial\mathcal{B}$ ($\mathcal{N}$ is the outward normal to the surface of domain $\mathcal{B}$). It is assumed that the gradient is scaled by $\varphi > 0$. This quantity is constant in the CGD model and denoted in this paper by the parameter φ_s, which is equivalent to c and related to the square of internal length scale l

$$\varphi_s = c = \frac{1}{2} l^2 \tag{2.7}$$

The internal length scale sets the localization band width (Geers, 1997; Peerlings *et al.*, 1996).

2.2. Localizing gradient damage (LGD) and gradient activity functions

In the localizing gradient damage (LGD) model, originally given by Poh and Sun (2017), the quantity φ becomes a variable and is called the gradient activity function. The averaging equation is rewritten as follows

$$\overline{\epsilon} - \nabla(\varphi \nabla \overline{\epsilon}) = \widetilde{\epsilon} \tag{2.8}$$

The above equation can be derived from a microforce balance, but here additional effects of micro-macro scale interaction, connected with the definition of a coupling stress, are not considered. The thermodynamic framework of the LGD model can be found e.g. in (Negi *et al.*, 2021; Poh and Sun, 2017; Wosatko, 2022). More detailed derivations and different aspects of implementation of this model are discussed, for example, by Wang *et al.* (2002) and Wosatko (2022). After discretization of the weak form and linearization of the BVP, it turns out that an additional matrix operator has to be computed in the matrix system of equations, where the derivative of φ is needed. Some proposals of gradient activity functions together with their derivatives are listed below.

When the LGD model is employed, the gradient activity is a function of damage ω. It is illustrated by Poh and Sun (2017), Wosatko (2022) that the influence of nonlocal interactions in the localization region should decrease with the increase of damage. It is observed that the crack band width gradually reduces and the model tends to the local one, so its localizing character reveals. The first formula for the gradient activity function is defined by Poh and Sun (2017) and includes exponential terms

$$\varphi_e(\omega) = c_{max} \frac{(1-R)\exp(-n\omega) + R - \exp(-n)}{1 - \exp(-n)} \tag{2.9}$$

In Eq. (2.9), c_{max} is the maximum internal length scale squared, R is the (minimum) residual level of nonlocal interaction and n is the power which changes the rate of decrease of the inter-

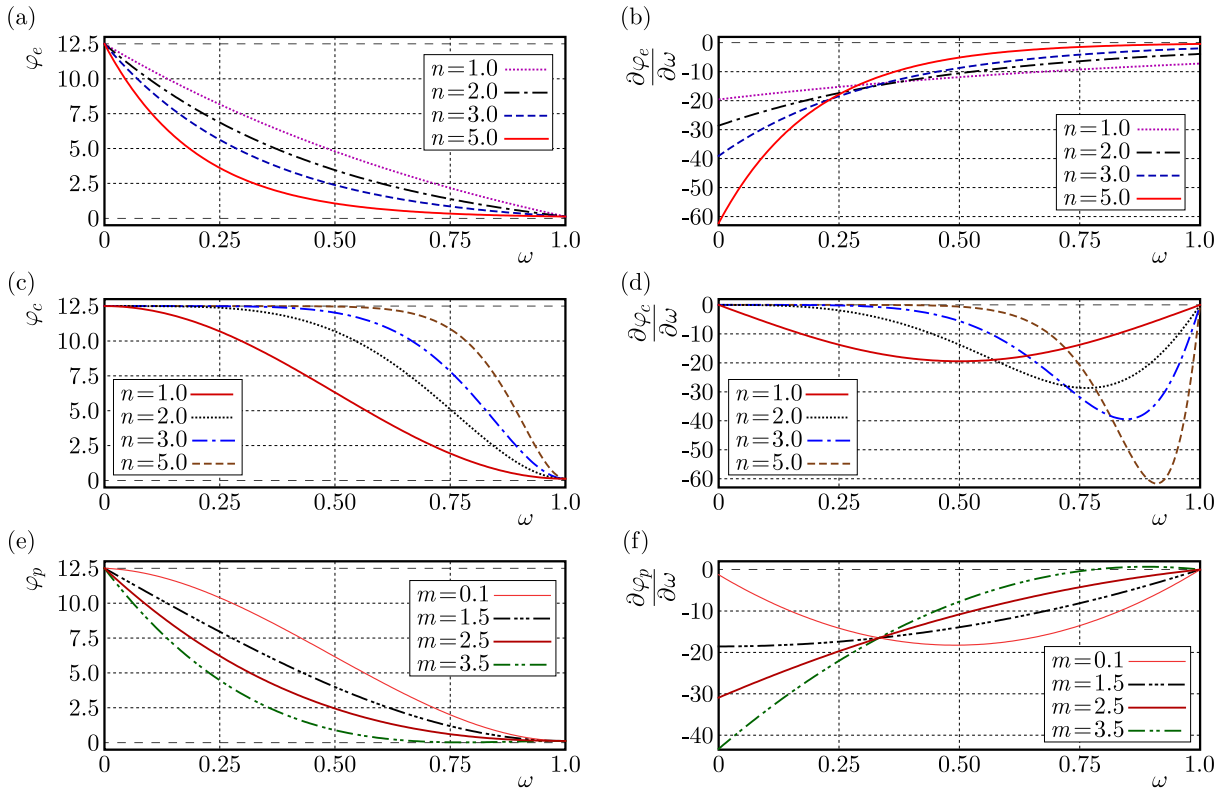


Fig. 1. Gradient activity functions and their derivatives for different values of n or m: (a) function φ_e, (b) derivative $\partial\varphi_e/\partial\omega$, (c) function φ_c, (d) derivative $\partial\varphi_c/\partial\omega$, (e) function φ_p, (f) derivative $\partial\varphi_p/\partial\omega$

action, which can be shortly called the intensity parameter. This function is depicted in Fig. 1a in diagrams with different n ($c_{max} = 12.5\,\mathrm{mm}^2$ and $R = 0.01$). The derivative of function φ_e is

$$\frac{\partial\varphi_e}{\partial\omega} = c_{max}\frac{(R-1)n\exp(-n\omega)}{1-\exp(-n)} \tag{2.10}$$

Figure 1b presents diagrams of this derivative for analogical cases. There are possible alternative definitions of the gradient activity function. The relation $\varphi(\omega)$ and its derivative can be defined by means of cosine and sine functions as proposed by Wosatko (2022)

$$\begin{aligned}\varphi_c(\omega) &= c_{max}\Big[\frac{1}{2}\big(\cos(\pi\omega^n)+1\big)(1-R)+R\Big]\\ \frac{\partial\varphi_c}{\partial\omega} &= \frac{1}{2}\pi c_{max}n(R-1)\omega^{(n-1)}\sin(\pi\omega^n)\end{aligned} \tag{2.11}$$

Figures 1c and 1d illustrate both the definitions for $c_{max} = 12.5\,\mathrm{mm}^2$ and $R = 0.01$. It should be noticed that if $n > 1.0$ then the start of the decreasing interaction process is postponed. For example, for intensity $n = 5.0$, the value of φ_c is effectively reduced only after $\omega > 0.5$. The intentional retardation of this reduction is introduced by Wang *et al.* (2022) using function $\varphi_e(\omega)$ from Eq. (2.9), governed by an additional threshold for damage, so that the cracking in fiber reinforced ultra-high performance concrete beams can be simulated. The change of the interaction area within the localization region should be delayed for special concrete materials. As shown in Fig. 1d, for each n, the derivative $\partial\varphi_c/\partial\omega$ starts from the value 0.0 for $\omega = 0.0$ as well as it is equal to 0.0 for $\omega = 1.0$ at the end. It seems that the derivative should be zeroed especially in the final stage of failure ($\omega = 1.0$), when further damage increment is not possible. Another formula comes from the phase-field approach (Borden, 2012; de Borst and

Verhoosel, 2016), where the gradient activity function with polynomial terms is written based on the so-called degradation function

$$\varphi_p(\omega) = c_{max}\{[(m-2)(1-\omega)^3 + (3-m)(1-\omega)^2](1-R) + R\} \tag{2.12}$$

where m is a weighting factor of the polynomials. The derivative of function φ_p is

$$\frac{\partial \varphi_p}{\partial \omega} = c_{max}(1-R)[(6-3m)\omega^2 + (4m-6)\omega - m] \tag{2.13}$$

Figure 1e presents the function from Eq. (2.12) in diagrams for $c_{max} = 12.5\,\mathrm{mm}^2$, $R = 0.01$ and with different factors m, while in Fig. 1f the corresponding derivatives defined in Eq. (2.13) are drawn. When the value of m is 0.1 or smaller (not presented here), then the function φ_p is similar to φ_c with $n = 1.0$. In fact, it resembles a cosine function. On the other hand, when $m = 1.5$ or larger, then the function φ_p is similar to the function φ_e given in Eq. (2.9). From this point of view, it seems that the function φ_p has the most universal form. It should be noted that red curves in Fig. 1 represent cases employed in the computations in the next Section.

3. Numerical study of unnotched beam under three point bending

3.1. Geometry and material model data

The numerical example discussed in this Section is based on the experiment conducted by Grégoire *et al.* (2013). The left symmetric half of the domain is taken into account. Figure 2 depicts configuration of the beam subjected to three point bending. Mesh M3, which is applied in the size effect study with identical density for each specimen, is also illustrated in Fig. 2. The dimensions of four specimens are summarized in Table 1. The thickness $T = 50\,\mathrm{mm}$ is the same for all considered cases. The following mesh densities are used: mesh M1 includes 1260 nodes and 1065 FEs, M2 – 4610 nodes and 4230 FEs, M3 – 17615 nodes and 16860 FEs, M4 – 65153 nodes and 63740 FEs. Mesh M4 is prepared only for the LGD model. Four-noded FEs with linear interpolation of the displacement field and the averaged strain measure are adopted.

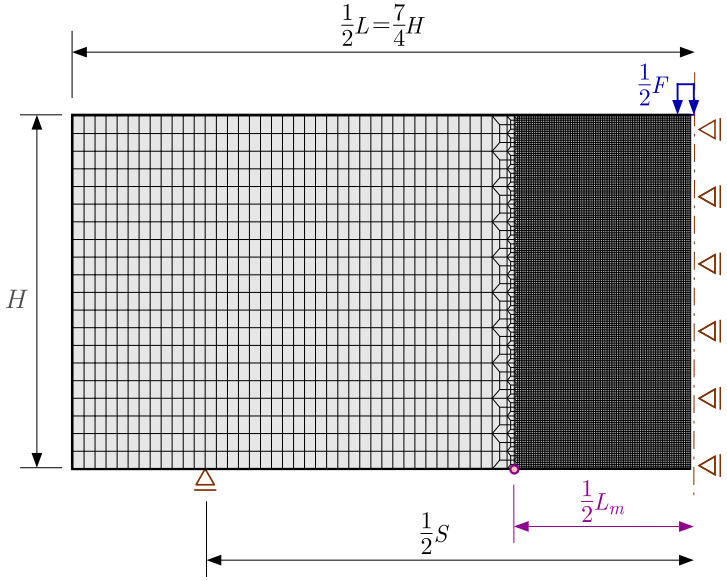


Fig. 2. Configuration of the symmetric half of the unnotched beam in three point bending and mesh M3

Plane stress conditions hold. Young's modulus $E = 37000\,\mathrm{MPa}$ and Poisson's ratio $\nu = 0.21$ are assumed for concrete. The threshold $\kappa_o = 0.0000946$ for the damage growth function in

Table 1. Geometry of specimens

Specimen	Length L [mm]	Height H [mm]	Span S [mm]	Measurement base L_m [mm]
D1	1400	400	1000	400
D2	700	200	500	200
D3	350	100	250	100
D4	175	50	125	50

Eq. (2.4) corresponds to the tensile strength $f_t = 3.5\,\text{MPa}$. The modified von Mises definition given in Eq. (2.3) is determined with the ratio $k = 12.086$, which means that the compressive strength is indirectly defined as $f_c = 42.3\,\text{MPa}$. The parameter $\alpha = 0.99$ is adopted for all computations. It is known that the same value of parameter η provides different behaviour for CGD and LGD models. The results for the LGD model give a much more brittle response, see e.g. (Poh and Sun, 2017), so the value of η connected with the rate of damage growth should be several times smaller than for the CGD model. Respectively, values 300 and 85 are applied. All the cases considered in the computations are listed in Table 2. Acronyms are connected with the CGD or LGD models and the choice of the gradient activity function φ. The parameters R and n or m are suitable for the given function. The internal length scale squared $c_{max} = 12.5\,\text{mm}^2$ is used in each case, but as a constant in the CGD model or as the maximum in the LGD model. For the CGD model, it simply means that $l = 5\,\text{mm}$. The options mentioned in Table 2 for the LGD model coincide with the red curves depicted in Fig. 1.

Table 2. Computational cases for size effect analysis

¢Case	Model				
	η	Function φ	Gradient activity	R	n or m
CGD	300	φ_s	constant	–	–
LGD-e	85	$\varphi_e(\omega)$	exponential	0.01	5.0
LGD-c	85	$\varphi_c(\omega)$	cosine	0.01	1.0
LGD-p-01	85	$\varphi_p(\omega)$	polynomial	0.01	0.1
LGD-p-25	85	$\varphi_p(\omega)$	polynomial	0.01	2.5

3.2. Mesh sensitivity study

Firstly, the numerical analysis is focused on the demonstration of the mesh-objective solution for both versions of the gradient damage model. Only specimen D3 is considered in this study. Figure 3 presents the diagrams of force F applied to the beam versus horizontal displacement u_{hor}, called also a pseudo-CMOD (crack mouth opening displacement), measured at the bottom edge between two points specified over the base L_m. A half of this base together with one point marked by a purple circle is illustrated in Fig. 2.

It is visible in Fig. 3a that all curves for the CGD model overlap, but simultaneously they deviate from the experiment. The contour plots for damage ω at the final stage are depicted in Fig. 4. It is seen that the same representation is obtained for each mesh. The most damaged region, where $\omega \to 1.0$, is illustrated by the black colour. All distributions of damage for the CGD model are quite spread. Therefore, the problem of too strongly broadened damage zone in the CGD model is confirmed, see also (Geers, 1997; Poh and Sun, 2017; Wosatko, 2022; Zhang *et al.*, 2021).

Figure 3b shows the equilibrium paths only for case LGD-p-25 with $\varphi_p(\omega)$ and $m = 2.5$. The mesh sensitivity study for the LGD model with functions φ_e and φ_c can be found in (Wosatko,

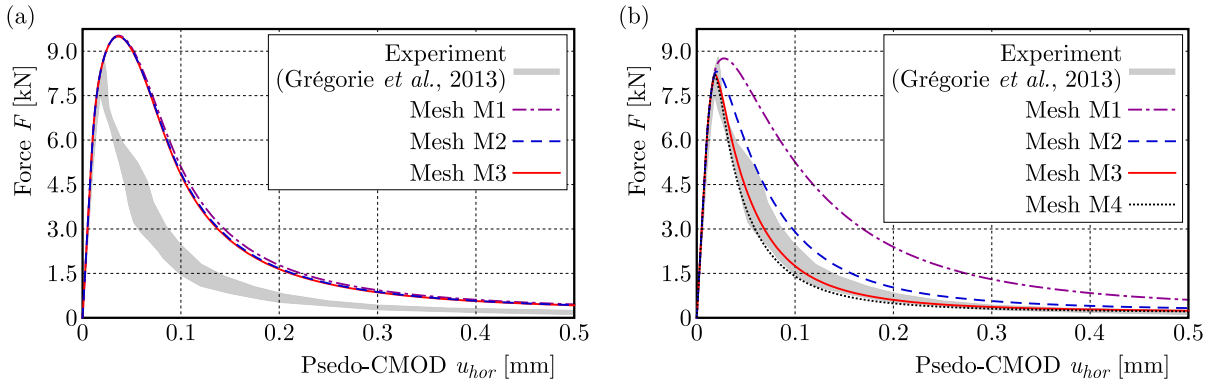


Fig. 3. Load vs. pseudo-CMOD diagrams, mesh-sensitivity study: (a) CGD, (b) LGD-p-25

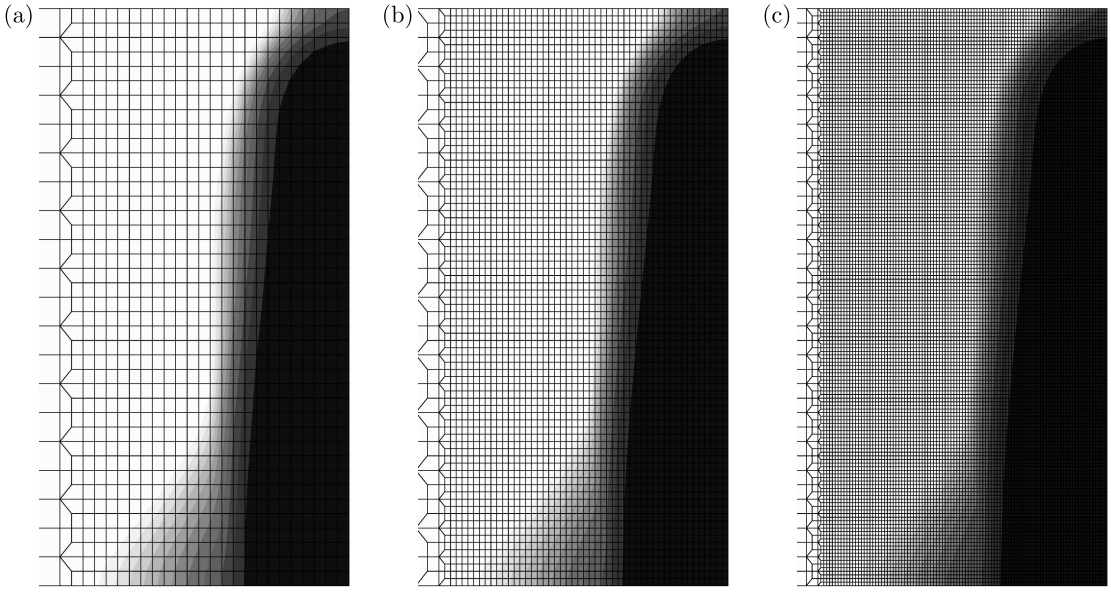


Fig. 4. Contour plots of damage ω for CGD, mesh-sensitivity study: (a) mesh M1, (b) mesh M2, (c) mesh M3

2022). It can be noticed here that the diagrams differ from each other, but the responses for M3 and M4 almost coincide. Starting from the diagram for mesh M1, next for M2, M3 and finally M4, it is observed that the load peaks get smaller and tend to the load-carrying capacity obtained in the experiment. An analogical order of the results is seen after the peak for softening. The solutions are closer and closer to the experimental response. Moreover, together with the mesh density growth, the differences between the diagrams decline. It is known that the LGD model requires a well-refined discretization (Wosatko, 2022) or a smart mesh densification near the expected cracking region (Negi *et al.*, 2021). Indeed, three meshes are sufficient to show the mesh-independent results for the CGD model. In the case of LGD model, the fourth mesh M4 has to be employed to prove that the consecutive solutions converge. Figures 5 and 6a depict the final distributions of ω for case LGD-p-25. The crack patterns represented by damage have a similar character. Now the damage zone is clearly narrowed, so artificial widening of the damage distribution is eliminated, and the solution remains mesh-objective. Figure 6 contains enlarged plots for M4 to provide a better visibility against the background of this very dense mesh. Figure 6b shows the distribution of the gradient activity function φ_p in a reversed scale, i.e. the black colour indicates the smallest values. The shape of this distribution is slightly wider, but generally coincides with the damage distribution presented on the left in Fig. 6a.

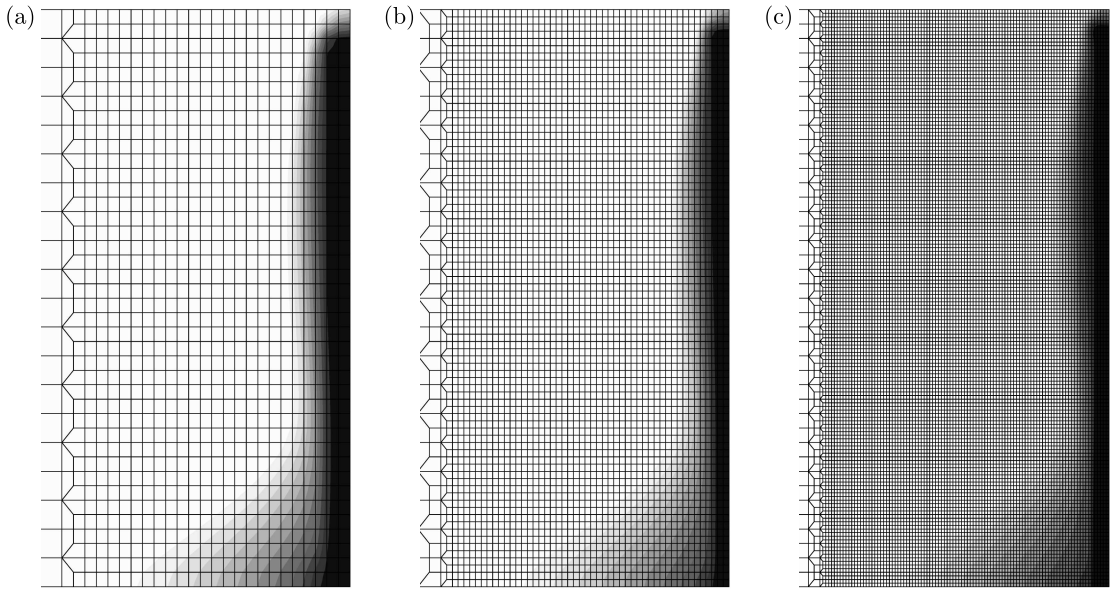


Fig. 5. Contour plots of damage ω for LGD-p-25, mesh-sensitivity study: (a) mesh M1, (b) mesh M2, (c) mesh M3

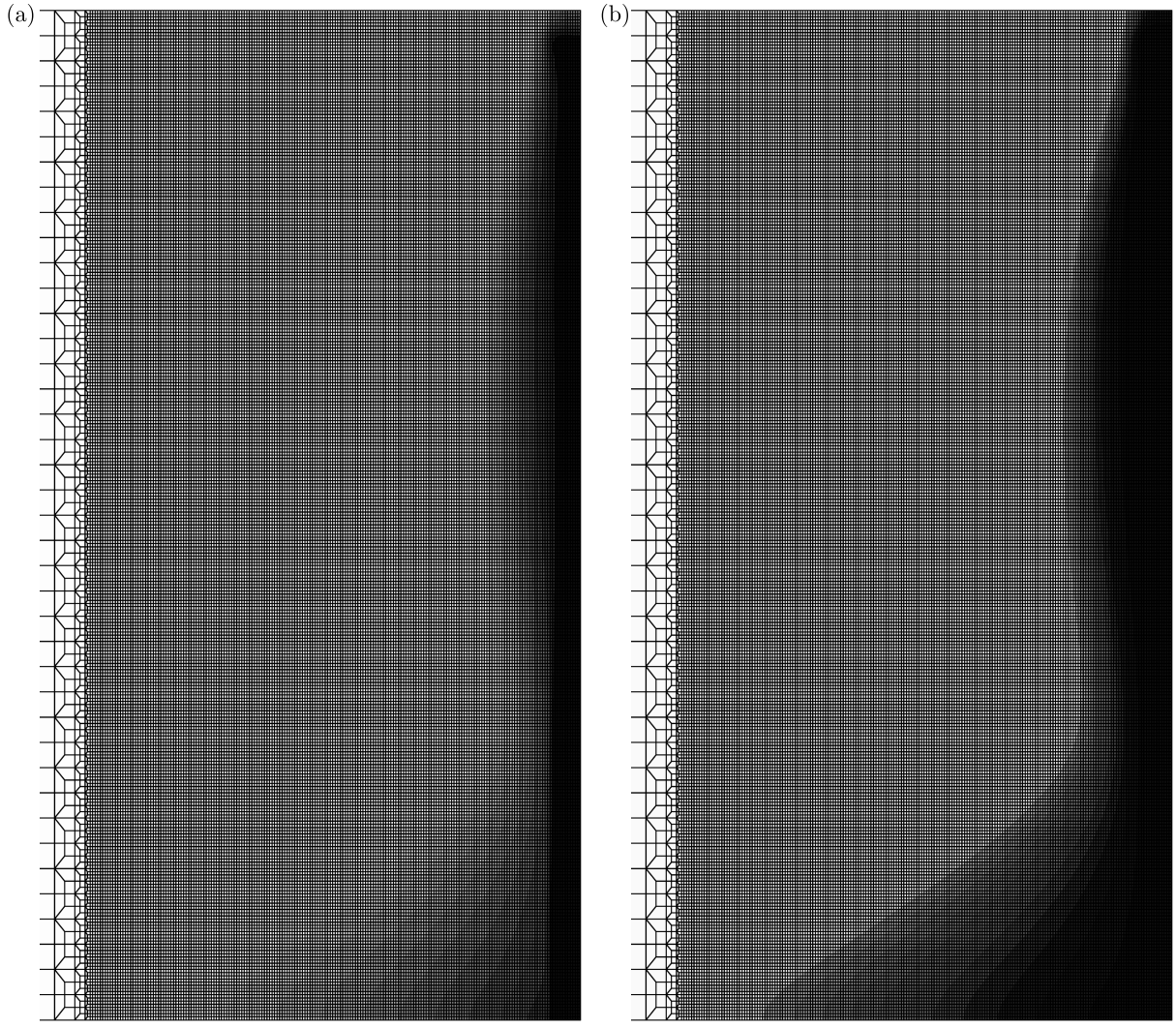


Fig. 6. Results for LGD-p25 and mesh M4: (a) damage ω, (b) gradient activity $\varphi_p(\omega)$ (reversed scale)

3.3. Size effect study

The results of the size effect study are discussed in this part of Section 3. Mesh M3 is selected for the computations for different specimen sizes, see Table 1. It should be reminded here that all experimental results are taken from Grégoire *et al.* (2013). Figure 7 juxtaposes the diagrams of the force F against the pseudo-CMOD u_{hor} for each beam, so that the confrontation between the experiment and responses for the cases defined in Table 2 can be carried out. The diagrams in Fig. 7a for the largest beam D1 are similar and differ only near the peak, however the curve after the peak for CGD gives a more brittle response. These equilibrium paths are over the limit of the gray zone coming from the experiment. The above observation changes together with reducing specimen dimensions. In Fig. 7b for beam D2, the response for the CGD model is more ductile than the others. The results for the LGD model are on the border of the gray region from the experiment. It is shown in Figs 7c and 7d for specimens D3 and D4 that CGD produces an exaggerated response, while the curves for options of the LGD model mostly fit the experimental results. They are different only for the maximum value of F, but in the same order for each beam size. Moreover, the cases LGD-c and LGD-p-01 overlap. The value of F for the case LGD-p-25 is below the previous two. The smallest F is obtained for the case LGD-e.

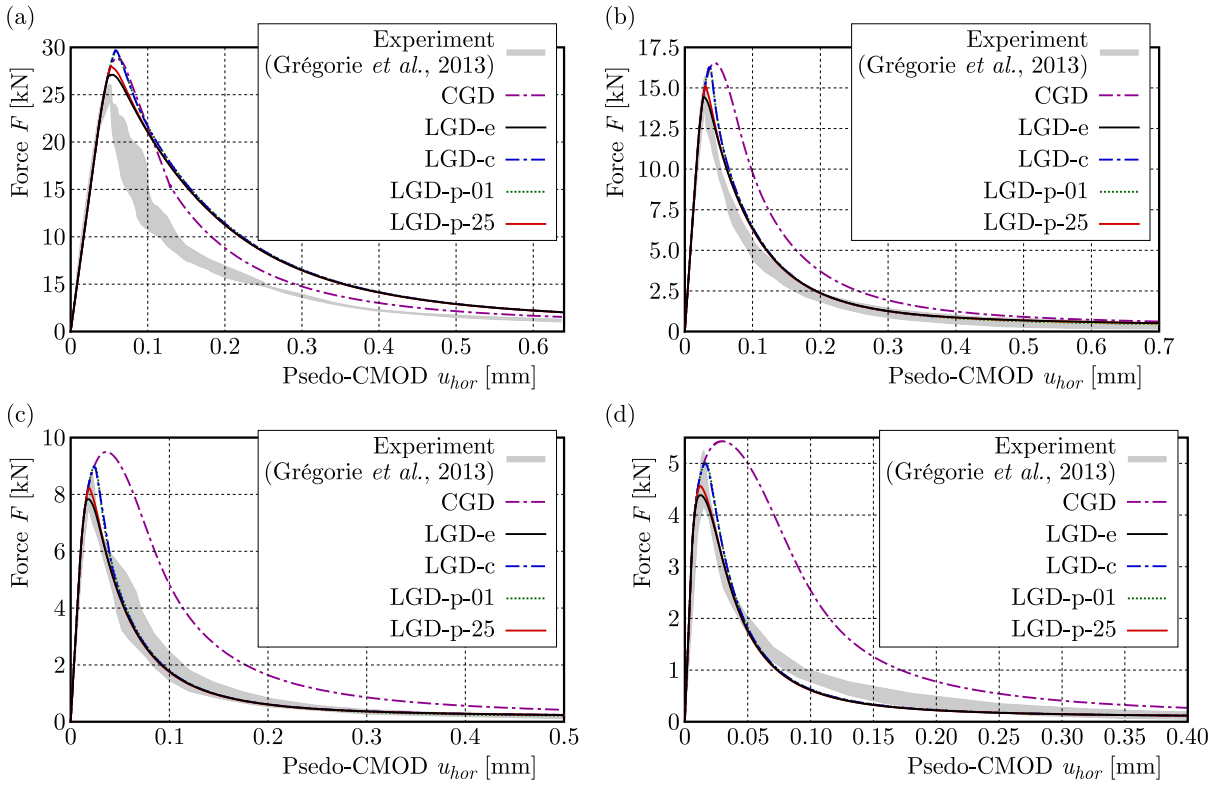


Fig. 7. Load vs. pseudo-CMOD diagrams – comparison between the employed models and experiment: (a) specimen D1, (b) specimen D2, (c) specimen D3, (d) specimen D4

Figures 8-11 present the FE meshes with final damage distributions for all cases given in Table 2 and, respectively, for all analyzed beams from Table 1. The undamaged areas, where $\omega \approx 0.0$, are represented by white colour, the gray scale shows the progress of cracking, and the total damage $\omega \to 1.0$ is depicted by the black colour. It is illustrated for specimen D1 in Fig. 8 that the active damage is limited to a quite narrow band along the symmetry axis of the beam for each computed case, but the zone for CGD is slightly wider. It seems that the effect of excessive broadening for the CGD model intensifies for smaller sizes of the beam. It should

be noted that the crack zone widths should be similar, while the sizes of FEs change and are proportional to the growing beam dimensions. In other words, the beam size should have minor influence on the size of the FPZ. In fact, the issue of spuriously widened damage zone for the CGD model is visible, see Figs. 10a and 11a. This drawback does not reveal for the LGD model. Of course, the visualized crack band widths in the contour plots increase from specimen D1 (largest) to D4 (smallest), but the damage zone widths for the LGD model are relatively quite small and the increase of the widths does not seem proportional to the size reduction. It can be observed that the damage patterns for LGD-c and LGD-p-01 are almost identical, so it is confirmed that the function $\varphi_c(\omega)$ with $n = 1.0$ conforms with the function $\varphi_p(\omega)$ with $m = 0.1$. On the other hand, similar damage distributions are obtained for cases LGD-e and LGD-p-25, i.e. the results for function $\varphi_e(\omega)$ with $n = 5.0$ and function $\varphi_p(\omega)$ with $m = 2.5$ are comparable.

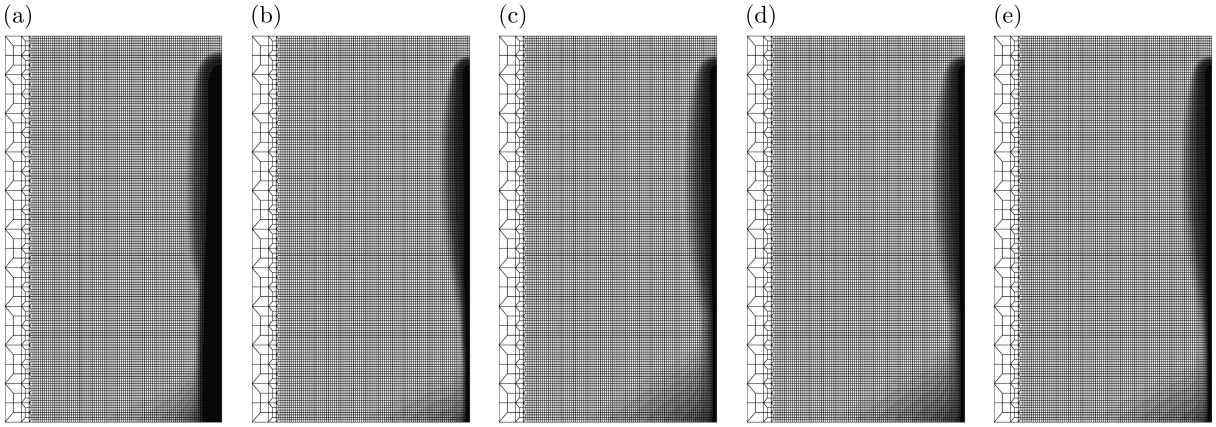


Fig. 8. Contour plots of damage ω for specimen D1: (a) CGD, (b) LGD-e, (c) LGD-c, (d) LGD-p-01, (e) LGD-p-25

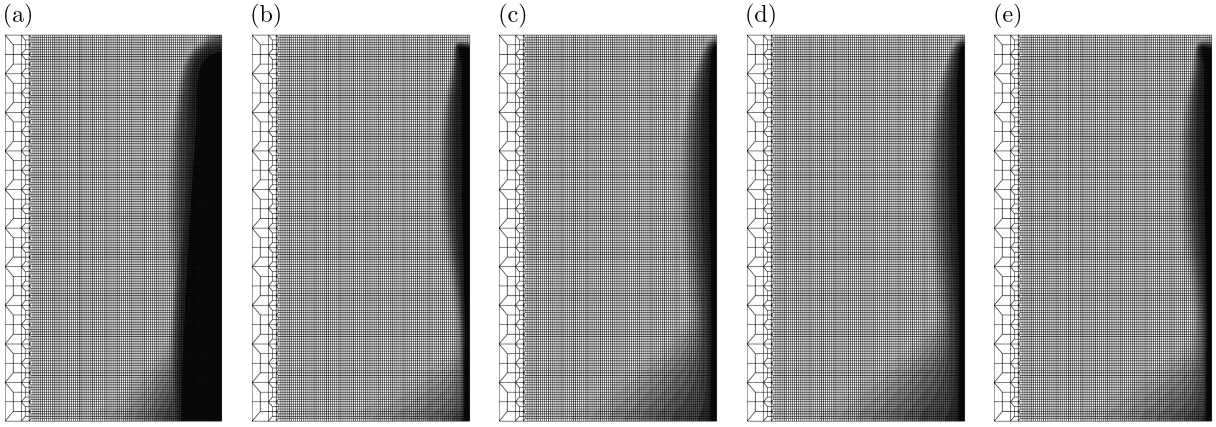


Fig. 9. Contour plots of damage ω for specimen D2: (a) CGD, (b) LGD-e, (c) LGD-c, (d) LGD-p-01, (e) LGD-p-25

Figure 12 shows the diagrams of nominal stress σ_{nom} versus the horizontal strain ε for cases CGD and LGD-p-25. Both quantities are calculated in the following way. The nominal stress is

$$\sigma_{nom} = \frac{3}{2}\frac{FS}{TH^2} \tag{3.1}$$

and the horizontal strain is $\varepsilon = u_{hor}/L_m$. It is seen that the value of the nominal stress grows together with the decrease of the beam size. The response becomes also less brittle when the

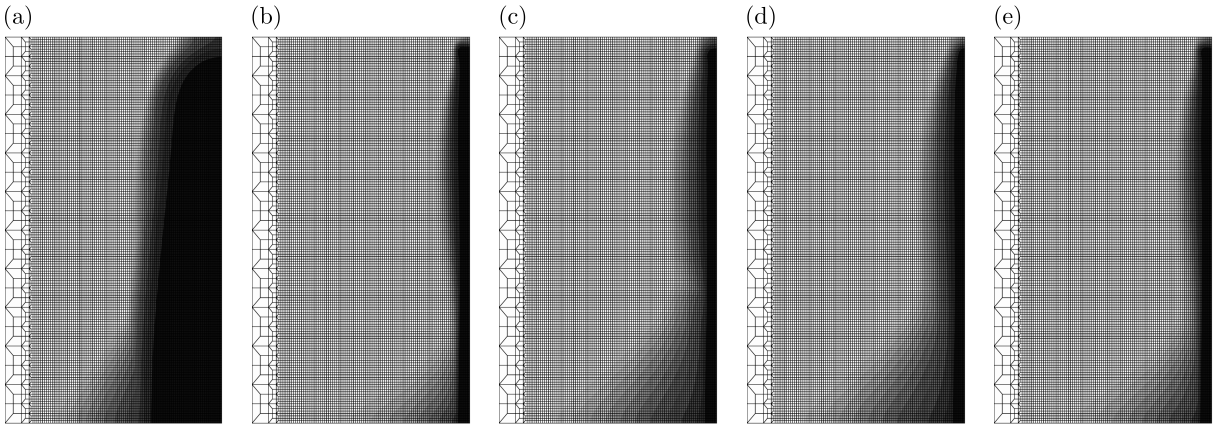


Fig. 10. Contour plots of damage ω for specimen D3: (a) CGD, (b) LGD-e, (c) LGD-c, (d) LGD-p-01, (e) LGD-p-25

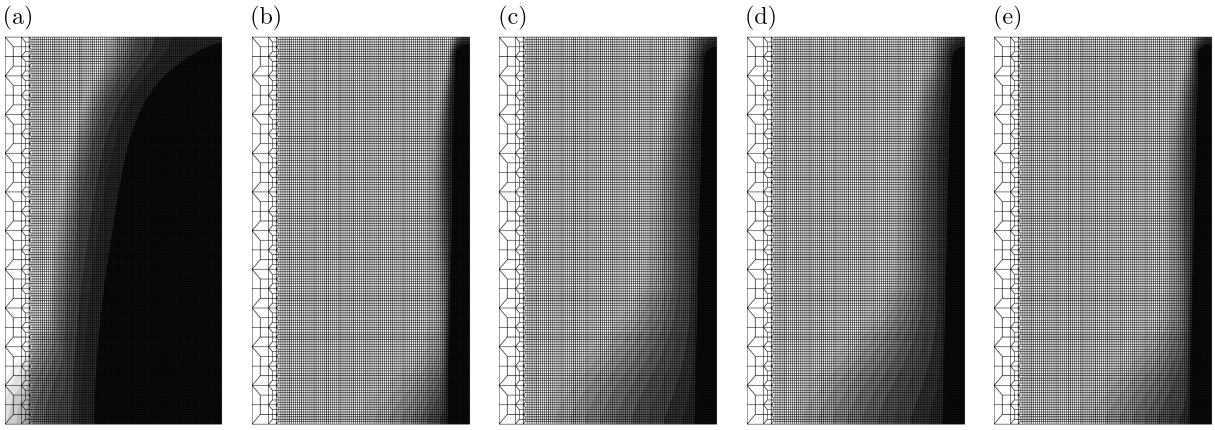


Fig. 11. Contour plots of damage ω for specimen D4: (a) CGD, (b) LGD-e, (c) LGD-c, (d) LGD-p-01, (e) LGD-p-25

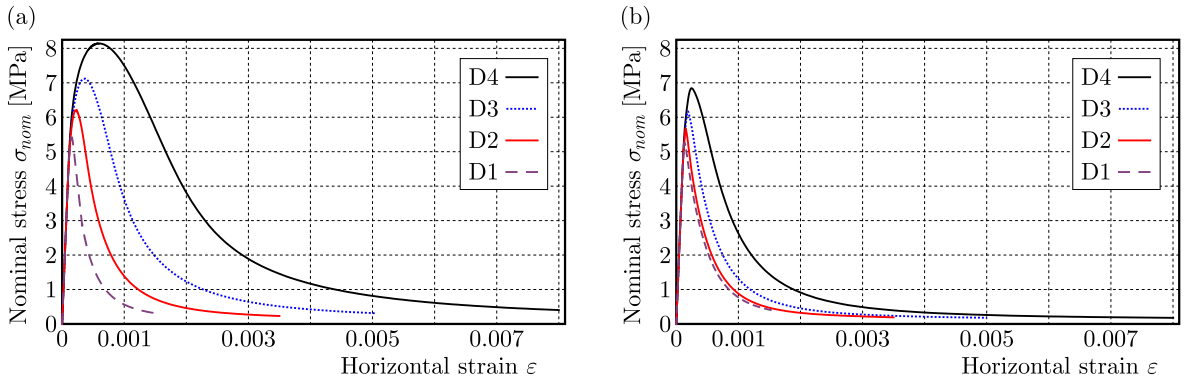


Fig. 12. Nominal stress vs horizontal strain diagrams – size effect study: (a) CGD, (b) LGD-p-25

specimen gets smaller. The CGD model demonstrates a much stronger size effect than the LGD-p-25. The size effect can also be verified based on Fig. 13, which is prepared in logarithmic scale for both axes. The nominal stress σ_{nom} is normalized by the tensile strength f_t, while the horizontal axis is determined by the proportion of beam heights H^i to the height H for specimen D4. The size effect is clearly visible for each case, but the results for CGD are over the zone repesenting the experiment.

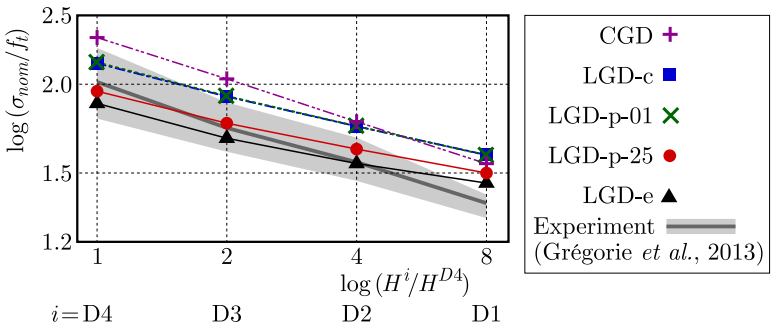


Fig. 13. Size effect plot – comparison between the employed models and experiment

4. Conclusions

In general, non-local finite element models should be able to simulate the deterministic size effect, because a localization limiter introduces an internal length scale which sets the size of the FPZ, independent of the specimen size. Gradient damage models are equipped with such a internal length scale, which can be a constant parameter as for the conventional gradient damage (CGD) model, or a variable represented by a gradient activity function as for the localizing gradient damage (LGD) model.

In the paper, three definitions of the gradient activity function are described and compared in the simulations of the size effect. The function with exponential terms is known from Poh and Sun (2017). The function defined by cosine terms is suggested by Wosatko (2022). The function with polynomial terms, proposed by Borden (2012), de Borst and Verhoosel (2016), in the context of the phase-field model as a degradation function, is used for the first time in the LGD model. This function seems to have the most universal character.

The numerical analysis has been focused on a concrete beam under three point bending. The mesh sensitivity study has confirmed that the CGD model exhibits an issue of spuriously widened damage zone, and the LGD model is able to simulate a properly narrow localization band. The results of the numerical size effect study are compared with the experimental results provided by Grégoire *et al.* (2013). The obtained results are similar to those presented by Zhang *et al.* (2021) and Negi *et al.* (2021) but here, the analysis is focused on the employment of various gradient activity functions. Based on the results for the unnotched beam, it is demonstrated that the LGD model properly simulates the size effect and, for carefully adopted model parameters, the differences in the results for different gradient activity functions are small.

References

1. Barbat G.B., Cervera M., Chiumenti M., Espinoza E., 2020, Structural size effect: Experimental, theoretical and accurate computational assessment, *Engineering Structures*, **213**, 110555
2. Bažant Z.P., Jirásek M., 2002, Nonlocal integral formulations of plasticity and damage: Survey of progress, *Journal of Engineering Mechanics – ASCE*, **128**, 11, 1119-1149
3. Bažant Z.P., Le J.-L., 2017, *Probabilistic Mechanics of Quasibrittle Structures. Strength, Lifetime, and Size Effects*, Cambridge University Press, Cambridge.
4. Bažant Z.P., Oh B., 1983, Crack band theory for fracture of concrete, *RILEM Materials and Structures*, **16**, 155-177
5. Bažant Z.P., Planas J., 1998, *Fracture and Size Effect in Concrete and Other Quasibrittle Materials*, CRC Press, New York
6. Borden M.J., 2012, Isogeometric analysis of phase-field models for dynamic brittle and ductile fracture, Ph.D. Thesis, The University of Texas at Austin, Austin, Texas

7. Carmeliet J., 1999, Optimal estimation of gradient damage parameters from localization phenomena in quasi-brittle materials, *Mechanics of Cohesive-Frictional Materials*, **4**, 1, 1–16
8. de Borst R., Verhoosel C.V., 2016, Gradient damage vs. phase-field approaches for fracture: Similarities and differences, *Computer Methods in Applied Mechanics and Engineering*, **312**, 78-94
9. de Vree J.H.P., Brekelmans W.A.M., van Gils M.A.J., 1995, Comparison of nonlocal approaches in continuum damage mechanics, *Computers and Structures*, **55**, 4, 581-588
10. Feng D.-C., Wu J.-Y., 2018, Phase-field regularized cohesive zone model (CZM) and size effect of concrete, *Engineering Fracture Mechanics*, **197**, 66-70
11. García-Álvarez V.O., Gettu R., Carol I., 2012, Analysis of mixed-mode fracture in concrete using interface elements and a cohesive crack model, *Sadhana*, **37**, 1, 187-205
12. Geers M.G.D., 1997, Experimental analysis and computational modelling of damage and fracture, Ph.D. Thesis, Eindhoven University of Technology, Eindhoven
13. Grégoire D., Rojas-Solano L.B., Pijaudier-Cabot G., 2013, Failure and size effect for notched and unnotched concrete beams, *International Journal for Numerical and Analytical Methods in Geomechanics*, **37**, 10, 1434-1452
14. Hoover C.G., Bažant Z.P., Vorel J., Wendner R., Hubler M.H., 2013, Comprehensive concrete fracture tests: Description and results, *Engineering Fracture Mechanics*, **114**, 92-103
15. Hordijk D.A., 1991, Local approach to fatigue of concrete, Ph.D. Thesis, Delft University of Technology, Delft
16. Mazars J., Pijaudier-Cabot G., 1989, Continuum damage theory – application to concrete, *Journal of Engineering Mechanics – ASCE*, **115**, 2, 345-365
17. Negi A., Singh U., Kumar S., 2021, Structural size effect in concrete using a micromorphic stress-based localizing gradient damage model, *Engineering Fracture Mechanics*, **243**, 107511
18. Peerlings R.H.J., de Borst R., Brekelmans W.A.M., de Vree J.H.P., 1996, Gradient enhanced damage for quasi-brittle materials, *International Journal for Numerical Methods in Engineering*, **39**, 19, 3391-3403
19. Peerlings R.H.J., Massart T.J., Geers M.G.D., 2004, A thermodynamically motivated implicit gradient damage framework and its application to brick masonry cracking, *Computer Methods in Applied Mechanics and Engineering*, **193**, 30, 3403-3417
20. Poh L.H., Sun G., 2017, Localizing gradient damage model with decreasing interaction, *International Journal for Numerical Methods in Engineering*, **110**, 6, 503-522
21. Saroukhani S., Vafadari R., Simone A., 2013, A simplified implementation of a gradient-enhanced damage model with transient length scale effects, *Computational Mechanics*, **51**, 6, 899-909
22. Taylor R., 2001, *FEAP – A Finite Element Analysis Program, Version 7.4, User Manual*, University of California at Berkeley, Berkeley
23. Wang J., Poh L.H., Guo X., 2022, Mixed mode fracture of geometrically similar FRUHPC notched beams with the localizing gradient damage model, *Engineering Fracture Mechanics*, **275**, 108843
24. Wosatko A., 2022, Survey of localizing gradient damage in static and dynamic tension of concrete, *Materials*, **15**, 5, 1875
25. Zhang Y., Shedbale A.S., Gan Y., Moon J., Poh L.H., 2021, Size effect analysis of quasi-brittle fracture with localizing gradient damage model, *International Journal of Damage Mechanics*, **30**, 7, 1012-1035
26. Zhao D., Yin B., Tarachandani S., Kaliske M., 2023, A modified cap plasticity description coupled with a localizing gradient-enahnced approach for concrete failure modeling, *Computational Mechanics*, **72**, 787-801

Manuscript received October 30, 2023; accepted for print November 27, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 207-218, Warsaw 2024
https://doi.org/10.15632/jtam-pl/183697

CRYOPRESERVATION ANALYSIS CONSIDERING DEGREE OF CRYSTALLISATION USING FUZZY ARITHMETIC[1]

Alicja Piasecka-Belkhayat, Anna Skorupa
Department of Computational Mechanics and Engineering, Silesian University of Technology, Gliwice, Poland
e-mail: anna.skorupa@polsl.pl

This article presents numerical modelling of the heat transfer process in a sample during cryopreservation by vitrification in a microfluidic system. Single-phase flow of the working fluid in the microchannels during warming was considered, while two-phase flow during cooling. The mathematical model is based on the Fourier equation with a source term that takes into account the degree of ice crystallisation. Fuzzy thermophysical parameters were assumed in the model. The problem was solved by the finite difference method and the fourth-order Runge-Kutta algorithm, using the concept of α-cuts. The results of numerical simulation were compared with the results from the literature.

Keywords: cryopreservation, crystallisation, vitrification, trapezoidal fuzzy numbers, α-cuts concept

1. Introduction

During modelling physical phenomena occurring in biological samples, uncertain parameters are often used. These variables are imprecise because they are determined experimentally and depend on the age, sex, and condition of the examined organism. As a consequence, physical processes in biological samples or in engineering systems are often simulated with deterministic models that introduce at the same time some assumptions and simplifications (Wang and Matthies, 2021).

On the other hand, uncertain quantities present in physical phenomena can be predicted by probabilistic or non-probabilistic approaches. Stochastic methods based on probabilistic algorithms are successfully suitable for describing uncertainties with a known database containing measurements of a given parameter. From these input data, a probability distribution function can be prepared to describe uncertainty characteristics. Unfortunately, for some engineering problems, probabilistic techniques are ineffective due to limited access to measurement data (Wang and Matthies, 2021).

Alternative approaches to the modelling of uncertain quantities are fuzzy set theory and interval set theory, which are qualified as non-probabilistic methods (Lü *et al.*, 2017). The concept of fuzzy sets was suggested by Lofti and Zadeh in 1965 (Zadeh, 1965). According to this theory, a membership function is defined for each element of the set, which takes values in the range from 0 to 1. The membership function determines whether a given element belongs to the set completely, partially, or is external to it (Hanss, 2005; Skorupa, 2023). There are different types of membership functions, where the simplest include linear functions, such as triangular and trapezoidal ones, and the most complex functions are Gaussian or bell curves (Caniani *et al.*, 2011). It is worth mentioning that fuzzy numbers are often performed using α-cuts concept (Giachetti and Young, 1997; Guerra and Stefanini, 2005).

An other effective method is the interval set theory introduced by Ramon and Moore in 1966 (Moore, 1966). In that case, uncertain variables are described by a set for which the upper and

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

lower bounds are specified. The membership function is equal to 1 when the element belongs to the set, or 0 when it is not in the given interval (Lü *et al.*, 2017; Skorupa, 2023).

The subject of this paper focusses on the use of fuzzy set theory to model phenomena that occur during cryopreservation. This is a process that involves reducing biological activity of a biological material and then storing it at a low temperature. The essence of cryopreservation is to prevent samples from damage after restoring them to physiological temperature (Jang *et al.*, 2017; Zhao and Fu, 2017).

One of the phenomena present during cryopreservation is crystallisation, which involves the formation of ice crystals from the water contained in the biological sample. The crystallisation process is initiated by nucleation caused by the metastable homogeneous phase. Then around the nucleus, growth of ice crystals occurs (Tan *et al.*, 2021).

The formation of ice crystals can damage biological samples. To prevent these problems, during cryopreservation, chemical compounds called cryoprotectants (CPAs) are used, and the cooling rate is regulated (Skorupa, 2023). On the basis of this, different cryopreservation methods can be distinguished. Vitrification, for example, involves overcooling the sample at a high cooling rate, thus vitrifying the solution without ice crystal formation (Jang *et al.*, 2017).

To predict the potential damage to the cryopreserved sample, the cryopreservation process is modelled mathematically and numerically. For this, it is necessary to consider various transport phenomena, such as heat and mass transfer or osmotic transport (Skorupa, 2023). The governing equation for estimating the thermal distribution is the Fourier equation (Fourier, 1882). Furthermore, this relationship is coupled to the degree of crystallisation determined, for example, from the non-isothermal kinetic equation proposed by Boutron and Mehl (1990). For preparing a macroscopic model of the liquid solidification process, different approaches are applied, for example, the uncoupled method, the Stefan model (sharp interface method) and the zone model (Skorupa, 2023; Song *et al.*, 2010; Zhou *et al.*, 2013).

In the literature, it is possible to find examples of mathematical calculations of thermal processes during freezing coupled to the degree of crystallisation with deterministic models (Shi *et al.*, 2018; Song *et al.*, 2010; Zhang *et al.*, 2017; Zhou *et al.*, 2013). Song *et al.* (2010) presented an example of modelling vitrification process modelling performed by the droplet-based method. The same technique for cell aggregates was also examined by Shi *et al.* (2018). Zhang *et al.* (2017) studied vitrification, in which the tube with the sample was immersed directly into the working fluid. That numerical calculation was supplemented by a model of the probability of intracellular ice formation. Zhou *et al.* (2013) investigated the vitrification process, which was analysed in a microfluidic system.

The interesting position is the article (Piasecka-Belkhayat and Skorupa, 2023) published in 2023, which is devoted to modelling thermal processes, including the degree of crystallisation, using the interval set theory. Further research on applying non-probabilistic methods, including interval numbers and triangular fuzzy numbers, can be found in the thesis (Skorupa, 2023).

This article presents the modelling of bioheat transfer and degree of crystallisation during cryopreservation in a microfluidic device with imprecise parameters. The uncertainties are considered by non-probabilistic methods, more specifically, algorithms for fuzzy numbers defined by a triangular and a trapezoidal membership function. The results obtained are compared with data from the literature (Piasecka-Belkhayat and Skorupa, 2023; Skorupa, 2023; Zhou *et al.*, 2013). The problem is specified mathematically by the Fourier equation and the non-isothermal kinetic equation, while simulations are conducted with the finite difference method (FDM) and the Runge-Kutta algorithm.

2. Governing equations

The paper analyses the task of one-dimensional heat transfer, taking into account that the cell suspension is maintained as a thin layer on the chip. Figure 1 shows the microfluidic system model based on the concept presented in (Tuckerman and Pease, 1981; Zhou *et al.*, 2013). The diagram also includes the marked nodes A, B and C where the corresponding boundary conditions are given.

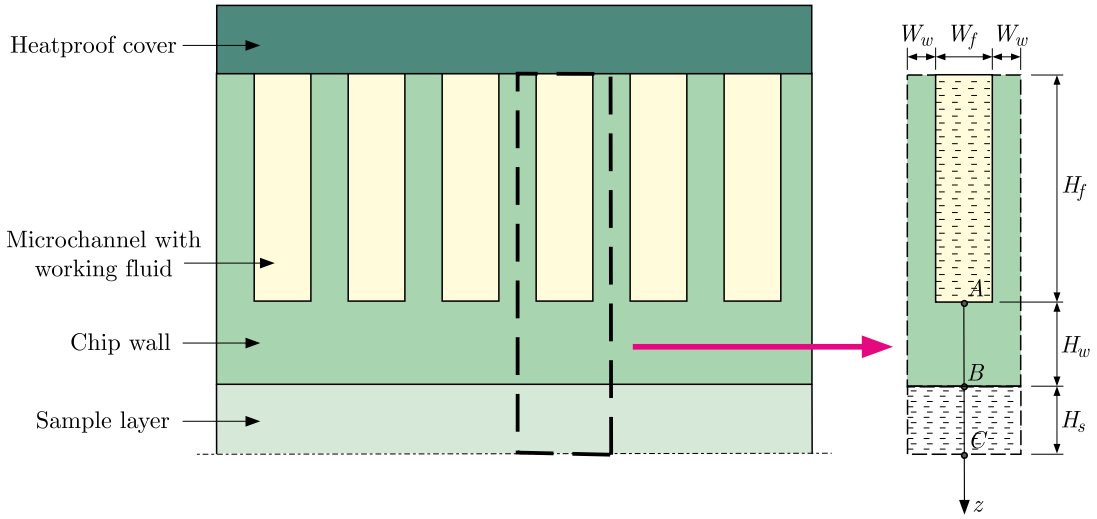


Fig. 1. Model of the microfluidic system

The energy equation describing the temperature distribution within a one-dimensional microfluidic chip can be formulated as the fuzzy Fourier equation (Fourier, 1882)

$$\frac{\partial(\widetilde{c}(\widetilde{T})\rho\widetilde{T}(z,t))}{\partial t} = \frac{\partial}{\partial z}\Big(\widetilde{\lambda}(\widetilde{T})\frac{\partial\widetilde{T}(z,t)}{\partial z}\Big) + \widetilde{S}_h \tag{2.1}$$

where $\widetilde{T}$ is the fuzzy temperature, $\widetilde{\lambda}$ is the fuzzy thermal conductivity, $\widetilde{c}$ is the fuzzy specific heat, $\widetilde{S}_h$ is the fuzzy heat source term, while ρ is the density.

The fuzzy source component $\widetilde{S}_h$ for the sample layer is expressed by the following relation (Shi *et al.*, 2018)

$$\widetilde{S}_h = \rho_h L_h \frac{\partial\widetilde{\chi}}{\partial t} \tag{2.2}$$

where $\widetilde{\chi}$ is the fuzzy degree of ice crystallisation belonging to the interval $(\widetilde{0},\widetilde{1})$, ρ_h is the density of water, and L_h is the latent heat of water. In contrast, for the chip wall, this component is equal to zero.

In the above equation, values for water were implemented instead of the thermophysical parameters of the biological sample due to the fact that the thermal properties of biological cells are similar to those of their prevalent component, namely water (Shi *et al.*, 2018).

The crystallisation process is described by the non-isothermal Mehl-Boutron kinetic equation (Boutron and Mehl, 1990)

$$\frac{\partial\widetilde{\chi}}{\partial t} = \widetilde{\chi}'(\widetilde{\chi},T) = k_a\widetilde{\chi}^{\frac{2}{3}}(1-\widetilde{\chi})(T_m-\widetilde{T})\mathrm{e}^{-\frac{Q}{RT}} \tag{2.3}$$

where $\partial\widetilde{\chi}/\partial t$ means the growth rate of $\widetilde{\chi}$, while k_a is the characteristic coefficient depending on the solution composition, T_m is the freezing (melting) temperature, Q is the activation energy and R is the gas constant ($R = 8.314\,\mathrm{J\,mol^{-1}K^{-1}}$).

The boundary conditions must be attached to the mathematical model defined in this way. Due to the use of thermal insulation (see Fig. 1), the heat flux between the devices and the environment is neglected.

Node A, shown in Fig. 1, lies at the boundary between the model and the working fluid, which is the main source of thermal changes. The fuzzy heat flux $\widetilde{q}$ at this node is described by the boundary condition of the 3rd type extended by the microchannel geometry (Zhou *et al.*, 2013)

$$\widetilde{q}(z,t)(W_f + 2W_w) = \alpha_\Gamma(\widetilde{T}(z,t) - T_f)(W_f + 2\widetilde{\eta}H_f) \tag{2.4}$$

where W_f, W_w and H_f are the microchannel dimensions (see Fig. 1), T_f is the temperature of the working fluid, α_Γ is the external heat transfer coefficient, $\widetilde{\eta}$ is the fuzzy fin efficiency and subscripts w and f denote the chip wall and the working fluid, respectively.

The fuzzy heat flux is defined as follows (Mochnacki and Suchy, 1993)

$$\widetilde{q}(z,t) = -\mathbf{n}\widetilde{\lambda}\frac{\partial\widetilde{T}(z,t)}{\partial z} \tag{2.5}$$

where $\mathbf{n}$ is the normal vector.

The fuzzy fin efficiency is determined from the relationship (Zhou *et al.*, 2013)

$$\widetilde{\eta} = \frac{\tanh(\widetilde{m}H_f)}{\widetilde{m}H_f} \tag{2.6}$$

where $\widetilde{m}$ is the fuzzy fin parameter (Zhou *et al.*, 2013)

$$\widetilde{m} = \sqrt{\frac{2\alpha_\Gamma}{\widetilde{\lambda}_w(\widetilde{T})2W_w}} \tag{2.7}$$

where $\widetilde{\lambda}_w$ is the fuzzy thermal conductivity of the chip wall.

Let us now turn to node B which is located at the interface between the chip wall and the sample layer. At this node, a boundary condition of the 4th type is assumed with ideal contact (Mochnacki and Suchy, 1993)

$$-\mathbf{n}\widetilde{\lambda}_w\frac{\partial\widetilde{T}_w(z,t)}{\partial z} = -\mathbf{n}\widetilde{\lambda}_s\frac{\partial\widetilde{T}_s(z,t)}{\partial z} \qquad \widetilde{T}_w(z,t) = \widetilde{T}_s(z,t) \tag{2.8}$$

where subscript s denotes the sample layer domain.

On the other hand, an adiabatic condition was assumed at node C due to symmetry of the system under consideration (Mochnacki and Suchy, 1993; Zhou *et al.*, 2013)

$$\widetilde{q}(z,t) = -\mathbf{n}\widetilde{\lambda}\frac{\partial\widetilde{T}(z,t)}{\partial z} = \widetilde{0} \tag{2.9}$$

To complete the mathematical description, it is necessary to take into account the initial conditions in which the temperature and the degree of crystallisation in the sample domain were determined at time $t = 0$ (Zhou *et al.*, 2013)

$$\widetilde{T}(z,0) = T_0 \qquad \widetilde{\chi}(z,0) = \chi_0 \tag{2.10}$$

where T_0 and χ_0 are the initial values of temperature and degree of crystallisation, respectively.

3. Numerical model

The fuzzy finite difference method, which is a very good tool for solving nonlinear equations, was used to solve an unsteady heat transfer problem defined in this way, Eq. (2.1). The non-linearity present in Eq. (2.1) is related to both the varying values of thermophysical parameters $\widetilde{c}(\widetilde{T})$, $\widetilde{\lambda}(\widetilde{T})$ and the fuzzy source term $\widetilde{S}_h(\widetilde{T})$. The numerical model of thermal processes occurring in the microfluidic system is based on the finite difference method as presented in (Mochnacki and Suchy, 1993) supplemented by the rules of fuzzy arithmetics, in particular, the concept of α-cuts dedicated for triangular and trapezoidal fuzzy numbers (Piasecka-Belkhayat and Korczak, 2020; Skorupa, 2023).

According to the fuzzy number theory, any fuzzy number can be written as the sum of all its α-cuts (Piasecka-Belkhayat and Korczak, 2020; Skorupa, 2023)

$$\widetilde{a} = \sum_{\alpha\in[0,1]} \widetilde{a}_\alpha \tag{3.1}$$

where α-cuts are expressed as a set of closed intervals

$$\forall\alpha \in [0,1]: \qquad \widetilde{a}_\alpha = [a_\alpha^-, a_\alpha^+] \tag{3.2}$$

Considering the case of a triangular fuzzy number $\widetilde{\alpha} = (a^-, a_0, a^+)$, we can define the α-cut as a set of closed intervals of the following form (Skorupa, 2023)

$$\widetilde{a}_\alpha = [(a_0 - a^-)\alpha + a^-, (a_0 - a^+)\alpha + a^+] \tag{3.3}$$

where a_0 is the core of the number, while the values of a^- and a^+ denote the left and right ends of the fuzzy number respectively; while in the case of the trapezoidal fuzzy number $\widetilde{a} = (x_0, y_0, \sigma, \beta)$, the α-cut is represented as follows (Piasecka-Belkhayat and Korczak, 2020)

$$\widetilde{a}_\alpha = [x_0 - (1-\alpha)\sigma, y_0 + (1-\alpha)\beta] \tag{3.4}$$

where x_0 and y_0 are the defuzzifiers from the left and right side, respectively; σ and β are the left and right fuzzinesses. In this way, complicated arithmetic operations on fuzzy numbers are avoided by performing calculations on closed intervals, which are interval numbers, using the rules of interval arithmetics (Skorupa, 2023). The calculations carried out involved different values of the parameter α.

Using the finite difference method, a time grid with a constant step Δt and a geometric grid with a constant step h are introduced at the beginning. Figure 2 shows the three-point star idea, which is used to create the geometric mesh.

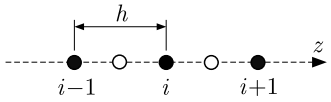


Fig. 2. Three-points star

The model assumes weak non-linearity of the specific heat. Therefore, the differential equation for internal nodes corresponding to the fuzzy Fourier equation written in the implicit scheme has the following form (Mochnacki and Suchy, 1993)

$$\widetilde{c}_i^{f-1}\rho_i\frac{\widetilde{T}_i^f - \widetilde{T}_i^{f-1}}{\Delta t} = \frac{2}{h^2}\frac{\widetilde{\lambda}_{i-1}^{f-1}\widetilde{\lambda}_i^{f-1}}{\widetilde{\lambda}_{i-1}^{f-1} + \widetilde{\lambda}_i^{f-1}}(\widetilde{T}_{i-1}^f - \widetilde{T}_i^f) + \frac{2}{h^2}\frac{\widetilde{\lambda}_{i+1}^{f-1}\widetilde{\lambda}_i^{f-1}}{\widetilde{\lambda}_{i+1}^{f-1} + \widetilde{\lambda}_i^{f-1}}(\widetilde{T}_{i+1}^f - \widetilde{T}_i^f) + (\widetilde{S}_h)_i^f \tag{3.5}$$

or it can be written as

$$A_i\widetilde{T}_{i-1}^f + B_i\widetilde{T}_i^f + C_i\widetilde{T}_{i+1}^f = \widetilde{T}_i^{f-1} + \frac{\Delta t}{\widetilde{c}_i^{f-1}(\rho_i)_i}(\widetilde{S}_h)_i^f \tag{3.6}$$

where coefficients A_i, B_i and C_i are calculated from the relationships

$$\begin{aligned} A_i &= -\frac{2\Delta t}{h^2\widetilde{c}_i^{f-1}\rho_i}\frac{\widetilde{\lambda}_{i-1}^{f-1}\widetilde{\lambda}_i^{f-1}}{\widetilde{\lambda}_{i-1}^{f-1}+\widetilde{\lambda}_i^{f-1}} \\ B_i &= \frac{2\Delta t}{h^2\widetilde{c}_i^{f-1}\rho_i}\Big(\frac{\widetilde{\lambda}_{i-1}^{f-1}\widetilde{\lambda}_i^{f-1}}{\widetilde{\lambda}_{i-1}^{f-1}+\widetilde{\lambda}_i^{f-1}} + \frac{\widetilde{\lambda}_{i+1}^{f-1}\widetilde{\lambda}_i^{f-1}}{\widetilde{\lambda}_{i+1}^{f-1}+\widetilde{\lambda}_i^{f-1}}\Big) + 1 \\ C_i &= -\frac{2\Delta t}{h^2\widetilde{c}_i^{f-1}\rho_i}\frac{\widetilde{\lambda}_{i+1}^{f-1}\widetilde{\lambda}_i^{f-1}}{\widetilde{\lambda}_{i+1}^{f-1}+\widetilde{\lambda}_i^{f-1}} \end{aligned} \tag{3.7}$$

while superscript f means a moment of time, the time step is $\Delta t = t^f - t^{f-1}$, $i = 2, \ldots, n_{el}-1$, where n_{el} is the number of nodes, and

$$\widetilde{c}_i^{f-1} = \widetilde{c}(\widetilde{T}_i^{f-1}) \qquad \widetilde{\lambda}_k^{f-1} = \widetilde{\lambda}(\widetilde{T}_k^{f-1}) \tag{3.8}$$

where k denotes the node number ($k = i-1, i, i+1$).

The obtained system of fuzzy equations supplemented with boundary-initial conditions can be solved using the Thomas method (Skorupa, 2023). It is important to note that the advantage of the implicit scheme is its stability and the lack of restrictions on allowable values of the time step (Mochnacki and Suchy, 1993).

Let us now turn to linearisation of the fuzzy source component $\widetilde{S}_h(\widetilde{\chi}_i^f)$, Eq. (2.2). The values of the fuzzy degree of ice crystallisation $\widetilde{\chi}$ and ice growth rate $\widetilde{\chi}'$ were calculated numerically using the fourth-order Runge-Kutta algorithm due to difficulty in calculating these relationships analytically (Piasecka-Belkhayat and Skorupa, 2023; Zhou *et al.*, 2013)

$$\chi^{f+1} = \widetilde{\chi}^f + (\widetilde{\chi}')^{f+1}\Delta t \tag{3.9}$$

where

$$(\widetilde{\chi}')^{f+1} = \frac{(\widetilde{\chi}')_1 + 2(\widetilde{\chi}')_2 + 2(\widetilde{\chi}')_3 + (\widetilde{\chi}')_4}{6} \tag{3.10}$$

while

$$\begin{aligned} &(\widetilde{\chi}')_1 = \widetilde{\chi}'(\widetilde{\chi}^f, \widetilde{T}^f) \qquad (\widetilde{\chi}')_2 = \widetilde{\chi}'\Big(\widetilde{\chi}^f + \frac{(\widetilde{\chi}')_1}{2}\Delta t, \widetilde{T}^f + \frac{\Delta\widetilde{T}^f}{2}\Big) \\ &(\widetilde{\chi}')_3 = \widetilde{\chi}'\Big(\widetilde{\chi}^f + \frac{(\widetilde{\chi}')_2}{2}\Delta t, \widetilde{T}^f + \frac{\Delta\widetilde{T}^f}{2}\Big) \\ &(\widetilde{\chi}')_4 = \widetilde{\chi}'(\widetilde{\chi}^f + (\widetilde{\chi}')_3\Delta t, \widetilde{T}^f + \Delta\widetilde{T}^f) \qquad \Delta\widetilde{T}^f = \widetilde{T}^{f+1} - \widetilde{T}^f \end{aligned} \tag{3.11}$$

The temperature dependence of thermal conductivity and specific heat of both the chip wall (made by silicon) and the sample layer (solution of ethylene glycol – EG and water) was assumed in the numerical model. These parameters were calculated as temperature-dependent polynomial functions using a linear regression method. In the case of silicon, the polynomial functions are of the form

$$\begin{aligned} \widetilde{\lambda}_w(\widetilde{T}) &= 1.3496\cdot 10^{-8}\widetilde{T}^5 + 1.1636\cdot 10^{-5}\widetilde{T}^4 + 0.0024\widetilde{T}^3 + 0.1416\widetilde{T}^2 - 2.0261\widetilde{T} + 54.3813 \\ \widetilde{c}_w(\widetilde{T}) &= 2.4923\cdot 10^{-7}\widetilde{T} + 9.1657\cdot 10^{-5}\widetilde{T}^3 + 0.0023\widetilde{T}^2 + 1.395\widetilde{T} + 677.6804 \end{aligned} \tag{3.12}$$

while for the EG solution they are expressed as

$$\begin{aligned}\widetilde{\lambda}_s(\widetilde{T}) &= -2.4041 \cdot 10^{-2}\widetilde{T}^2 - 17.741\widetilde{T} + 1442.8)\frac{1}{1000} \\ \widetilde{c}_s(\widetilde{T}) &= 2.8467\widetilde{T} + 2727.7\end{aligned} \tag{3.13}$$

It should be noted that in the case of silicon, the measurement points taken from the literature (Desai, 1986; Glassbrenner and Slack, 1964) in the temperature range 20-300 K (−253°C to 27°C) and coefficients $R^2 = 0.989$ for thermal conductivity and $R^2 = 0.999$ for specific heat were adopted. On the other hand, for the EG solution, the relationships given the producer MeGlobal$^{\text{TM}}$ (MeGlobalTM, 2008; Zhou *et al.*, 2013) were used. In addition, the fuzzy temperature appearing in Eqs. (3.12) and (3.13) should be expressed in °C.

4. Example of computations

The study analyses one-dimensional heat transfer in a microfluidic system (see Fig. 1) with the dimensions: $W_f = 5 \cdot 10^{-5}$ m, $W_w = 2.5 \cdot 10^{-5}$ m, $H_f = 3.5 \cdot 10^{-4}$ m and $H_w = H_s = 10^{-4}$ m (Piasecka-Belkhayat and Skorupa, 2023; Zhou *et al.*, 2013).

The microchannel contains the working fluid. During the cooling process, liquid nitrogen is applied, and because of the presence of the liquid form and particles of evaporated nitrogen, its flow is considered two-phase. In contrast, during warming, water is introduced into the system, whose flow is assumed to be single-phase. The working fluid is characterised by the following variables: for cooling $T_f = -196$°C, $\alpha_\Gamma = 1.048 \cdot 10^4\,\text{Wm}^{-2}\text{K}^{-1}$ and for warming $T_f = 40$°C, $\alpha_\Gamma = 4.74 \cdot 10^4\,\text{Wm}^{-2}\text{K}^{-1}$, respectively (Piasecka-Belkhayat and Skorupa, 2023; Zhou *et al.*, 2013).

The sample layer is examined as a solution of CPA (ethylene glycol, EG) and the biological sample is considered as water in proportions: 45% of EG and 55% of H_2O. For this solution, the vitrification parameters can be stated: $T_m = 243.5$ K, $Q = 4.187 \cdot 10^3\,\text{J mol}^{-1}$, $k_a = 3.933 \cdot 10^7$ $\text{s}^{-1}\text{K}^{-1}$ (for cooling), and $k_a = 1.287\,\text{s}^{-1}\text{K}^{-1}$ (for warming) (Piasecka-Belkhayat and Skorupa, 2023; Zhou *et al.*, 2013). Other quantities are also included in the calculations: $L_h = 334 \cdot 10^3\,\text{Jkg}^{-1}$, $\rho_h = 1000\,\text{kg m}^{-3}$, $\rho_w = 2330\,\text{kg m}^{-3}$ (Piasecka-Belkhayat and Skorupa, 2023).

In the numerical example, the parameters related to the time step and the given geometric grid are also specified: $\Delta t = 0.01$ s, $h_1 = 2.0202 \cdot 10^{-6}$ m, $n_{el,1} = 100$ (the number of elements is $l_1 = 99$). In addition, a numerical analysis was also performed for the modified grid, where $h_2 = 1.005 \cdot 10^{-6}$ m, $n_{el,2} = 200$ (the number of elements is $l_2 = 199$). The initial condition determines the values: $T_0 = 22$°C and $\chi_0 = 0$.

The interesting thing is the method of introducing trapezoidal fuzzy numbers into the model. At the time $t = 0$, the deterministic values of the uncertain thermal parameters are determined, and then the triangular and trapezoidal fuzzy numbers are defined according to the relations: $\widetilde{\lambda}_{w\,\text{or}\,s} = (\lambda_{w\,\text{or}\,s} - 0.05\lambda_{w\,\text{or}\,s}, \lambda_{w\,\text{or}\,s}, \lambda_{w\,\text{or}\,s} + 0.05\lambda_{w\,\text{or}\,s})$, $\widetilde{c}_{w\,\text{or}\,s} = (c_{w\,\text{or}\,s} - 0.05c_{w\,\text{or}\,s}, c_{w\,\text{or}\,s}, c_{w\,\text{or}\,s} + 0.05c_{w\,\text{or}\,s})$ and $\widetilde{\lambda}_{w\,\text{or}\,s} = (\lambda_{w\,\text{or}\,s} - 0.025\lambda_{w\,\text{or}\,s}, \lambda_{w\,\text{or}\,s} + 0.025\lambda_{w\,\text{or}\,s}, 0.025\lambda_{w\,\text{or}\,s}, 0.025\lambda_{w\,\text{or}\,s})$, $\widetilde{c}_{w\,\text{or}\,s} = (c_{w\,\text{or}\,s} - 0.025c_{w\,\text{or}\,s}, c + 0.025c_{w\,\text{or}\,s}, 0.025c_{w\,\text{or}\,s}, 0.025c_{w\,\text{or}\,s})$ – compare with the definition of trapezoidal fuzzy numbers provided in (Piasecka-Belkhayat and Korczak, 2020; Skorupa, 2023). As a consequence, the obtained results are triangular and trapezoidal fuzzy numbers as well.

Firstly, the results are reported for a mesh containing 100 nodes ($n_{el,1} = 100$). Figure 3 shows changes in the fuzzy temperature as a function of time for (a) cooling and (b) warming for $\alpha = 0.25$. The results have been prepared for the point in the centre of the sample (point C in Fig. 1, when $z = H_w + H_s$) as a solid line and for the point close to the contact between the sample layer and the chip wall (close to point B in Fig. 1, when $z = 1.01 \cdot 10^{-4}$ m) as a dashed

line. Because the obtained interval width is small, zoomed-in approximations of selected sections of the diagrams have been created.

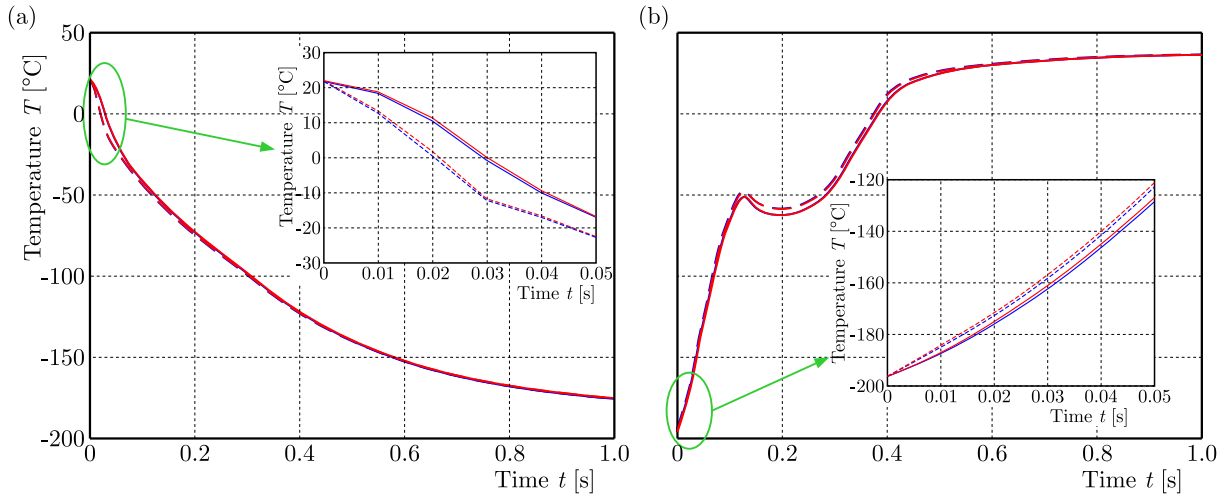


Fig. 3. Fuzzy temperature as a function of time during (a) cooling and (b) warming for $\alpha = 0.25$ with zoomed fragments

As can be observed, the minimum temperature was achieved after 14.1 s, confirming that a high cooling rate was used in the process (it is an average value of the fuzzy number). In addition, it can be concluded that point B responds more rapidly to temperature changes caused by the working fluid. The physiological temperature of the sample was restored after about 7.15 s during warming (it is the average value of the fuzzy number). Furthermore, from these plots and simulation data, it is possible to estimate the time required to pass through the "dangerous temperature region" (DTR), which typically occurs between −20°C and −90°C (Zhou *et al.*, 2013). In our case, it is equal to 0.2 s and 0.26 s for cooling and warming, respectively.

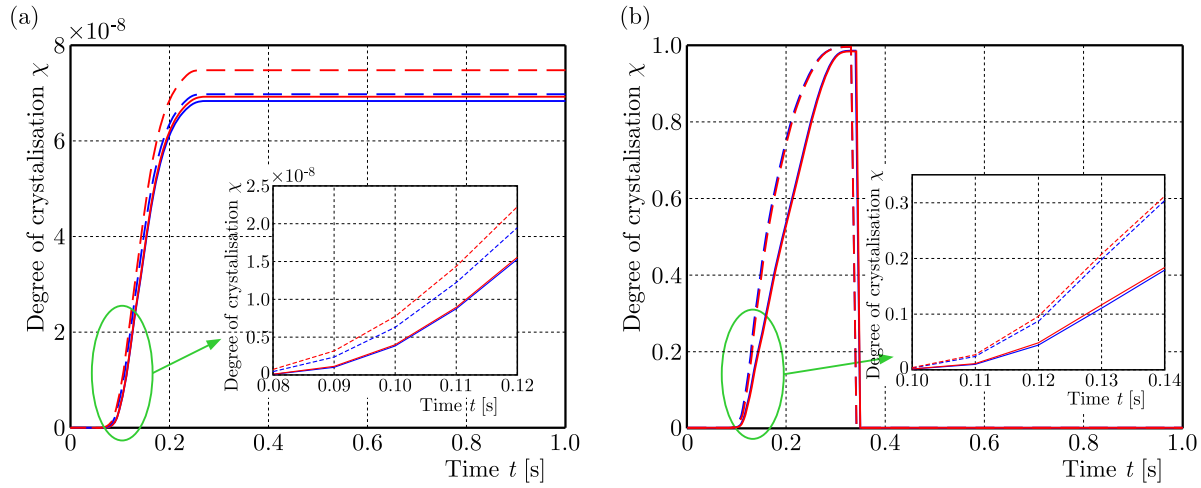


Fig. 4. Fuzzy degree of crystallisation as a function of time during (a) cooling and (b) warming for $\alpha = 0.25$ with zoomed fragments

Figure 4 illustrates the change in the fuzzy degree of crystallisation as a function of time in (a) cooling and (b) warming phases for $\alpha = 0.25$. Once again, functions for points B and C were produced, as well as an approximation of certain fragments of the curve. From Fig. 4, it can be deduced that the chart for cooling stabilises at a certain level. However, for warming, a sudden increase in the degree of crystallisation occurs at the time of passing through the DTR, and then goes to 0. This is caused by the phenomenon of recrystallisation. It should be noted

that the recrystallisation phenomenon is also reflected in Fig. 3b, where at a certain moment the temperature drops despite the continuous warming process.

Based on the analysis of the change in the degree of crystallisation, its maximum values can be indicated (average of the given fuzzy number). For the cooling and warming process, $\chi = 1.075 \cdot 10^{-7}$ and $\chi = 0.999$ were obtained, respectively.

Figure 5 presents a comparison of the trapezoidal fuzzy temperature and the trapezoidal fuzzy degree of crystallisation with the results obtained for triangular fuzzy numbers (compare with (Skorupa, 2023)) for different values of the parameter α. The graphs depict the moment s of simulation when $t = 0.3$ for point C (compare with Fig. 1). It can be noted that the values for $\alpha = 0$ coincide with each other, and that as the value of the parameter α increases, the intervals are narrower. One can see that for triangular fuzzy numbers for $\alpha = 1$ the deterministic values are obtained.

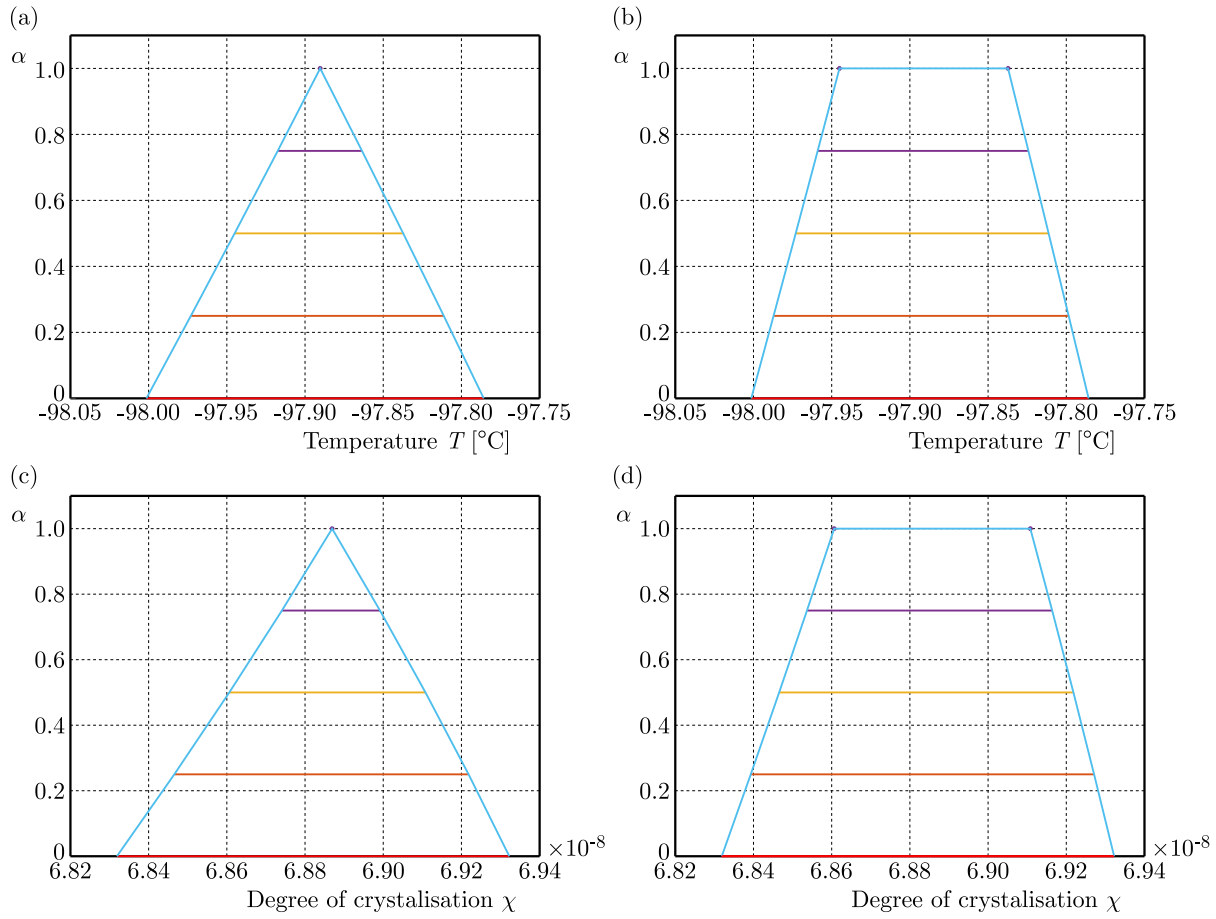


Fig. 5. Fuzzy temperature (a), (b), and fuzzy degree of crystallisation (c), (d) for $t = 0.3$ s during cooling for different values of the parameter α

Next, an analysis was performed for a mesh that includes 200 nodes ($n_{el,2} = 200$). Very similar results were obtained, which in the graphical form would look close to those shown in Figs. 3-5, so it was decided to compare the selected values in Table 1 for $n_{el,1}$ and $n_{el,2}$. The results presented in Table 1 are expressed as trapezoidal fuzzy numbers for $\alpha = 0.25$ at point C ($z = H_w + H_s$, cf. Fig. 1).

Table 1. Results for $\alpha = 0.25$ at point C for $n_{el,1}$ and $n_{el,2}$

Time t [s]	Fuzzy temperature $\widetilde{T}$ [°C]		Fuzzy degree of crystallisation $\widetilde{\chi}\cdot 10^{-8}$	
	$n_{el,1}$	$n_{el,2}$	$n_{el,1}$	$n_{el,2}$
During cooling				
0.0	[22.000; 22.000]	[22.000; 22.000]	[0.000; 0.000]	[0.000; 0.000]
0.1	[−40.985;−40.697]	[−40.817;−40.529]	[0.396; 0.380]	[0.390; 0.372]
0.2	[−72.096;−71.887]	[−71.836;−71.627]	[6.113; 6.178]	[6.222; 6.278]
0.4	[−121.146;−120.984]	[−120.726;−120.564]	[6.839; 6.927]	[6.983; 7.077]
0.6	[−152.109;−152.030]	[−151.796;−151.717]	[6.839; 6.927]	[6.983; 7.077]
0.8	[−166.891;−166.850]	[−166.676;−166.636]	[6.839; 6.927]	[6.983; 7.077]
1.0	[−175.086;−175.061]	[−174.923;−174.899]	[6.839; 6.927]	[6.983; 7.077]
During warming				
0.0	[−196.000;−196.000]	[−196.000;−196.000]	[6.839; 6.927]	[6.983; 7.077]
0.1	[−65.263;−65.940]	[−65.735;−66.411]	$[1.588\cdot 10^5; 1.334\cdot 10^5]$	$[1.410\cdot 10^5; 1.098\cdot 10^5]$
0.2	[−62.255;−62.269]	[−62.344;−62.352]	$[5.348\cdot 10^7; 5.304\cdot 10^7]$	$[5.285\cdot 10^7; 5.224\cdot 10^7]$
0.4	[7.785; 7.449]	[6.828; 6.478]	[0.000; 0.000]	[0.000; 0.000]
0.6	[30.394; 30.366]	[30.278; 30.249]	[0.000; 0.000]	[0.000; 0.000]
0.8	[34.751; 34.740]	[34.690; 34.679]	[0.000; 0.000]	[0.000; 0.000]
1.0	[36.765; 36.759]	[36.725; 36.719]	[0.000; 0.000]	[0.000; 0.000]

5. Discussion

Examining the results obtained, it is worth noting that vitrification was successfully modelled and the temperature given by the working fluid was reached relatively quickly in the sample domain. Furthermore, the study of the degree of crystallisation, which complements the thermal analysis, allows us to estimate the tendency of the solution to vitrify. It can also be treated as a marker of potential sample damage. As reported in the literature, this criterion assumes that $\chi < 10^{-6}$ (Piasecka-Belkhayat and Skorupa, 2023; Skorupa, 2023; Zhou *et al.*, 2013). In our simulation, this condition was fulfilled for freezing, but not for warming. This is due to the recrystallisation that occurs during the DTR transition.

The results can be compared with the data presented by Zhou *et al.* (2013). First, in the article published by Zhou *et al.* (2013), it was indicated that the DTR transition time was 0.042 s for cooling and 0.057 s for warming. It can be said that these values are more satisfactory than those calculated in our study. Similar observations appear in analysis of the maximum degree of crystallisation, which in Zhou *et al.* (2013) was $\chi = 2 \cdot 10^{-11}$ for cooling and $\chi = 2.4 \cdot 10^{-3}$ for warming. Those comparisons suggest that our mathematical and numerical models need some improvements.

On the other hand, it is worth exploring the results in terms of the uncertainties introduced. In this article, uncertain quantities are described by triangular and trapezoidal membership functions, which are non-probabilistic algorithms. Undoubtedly, fuzzy logic is a useful tool for modelling the inaccuracy of biological systems in the cryopreservation process. Comparing the results for triangular and trapezoidal fuzzy numbers, one can deduce that the calculated values are close to each other, where for $\alpha = 0$ the same intervals were achieved. A regularity can also be noticed that the higher the values of parameter α, the narrower intervals are obtained.

When focusing on the fuzzy arithmetics, it is also necessary to refer to the works (Piasecka-Belkhayat and Skorupa, 2023; Skorupa, 2023), in which a similar problem was considered by applying the interval set theory. In this case, the data received coincide with the simulation results

indicated in the literature (Piasecka-Belkhayat and Skorupa, 2023; Skorupa, 2023). Moreover, it can be observed that the results for $\alpha = 0$ are the same as for interval numbers, for which a divergence was defined as 5% of the deterministic value of the thermophysical parameters. It is worth remembering the difference in defining the interval and fuzzy numbers, where for interval numbers the membership function is only equal to 1 or 0. For fuzzy numbers, the partial membership of a given quantity can be formulated (membership function in the range $[0, 1]$), and the modelled phenomenon can be described in more detail. Therefore, this approach allows for a more flexible representation of quantities and their analysis.

The convergence of the finite difference method is related to the mesh step. Reducing the grid step affects the results close to the real solution. In this paper, additional calculations were performed for a mesh step reduced by half. The results obtained were slightly different from the initial value of the grid step (see Table 1). However, it should be emphasized that a mesh step that is too small increases numerical simulation time.

6. Conclusion

The paper presents a model of the vitrification process performed in a microfluidic system. The mathematical and numerical model considers uncertain quantities by applying the fuzzy set theory and α-cuts concept. Comparison of the results with data from the literature confirms that fuzzy numbers are effective in the modelling of uncertainty. Therefore, imprecision of thermophysical parameters is included without drastically increasing the computation time. The differences in the results suggest that the prepared model should be improved. For example, it is possible to change the mesh parameters or reduce the time step. However, after analysing the results shown in Table 1, it can be deduced that by modifying the mesh by half, similar results are attained. Another interesting idea is to calculate the results using different numerical methods, for example, finite volume method (FVM) like Zhou *et al.* (2013) or finite element method (FEM). It is also worth considering extending the model to investigate energy activation and ice nucleation. By exploring the nucleation rate and ice growth (a change in the diameter of the ice crystals), one can further estimate the volume of ice crystals.

An interesting aspect of the paper is the discussion about differences between fuzzy numbers and interval numbers, which was used in the literature to model cryopreservation and crystallisation (Piasecka-Belkhayat and Skorupa, 2023; Skorupa, 2023).

In further research, it would also be worth analysing the nucleation process and ice crystal growth.

Acknowledgment

The research was partially funded from financial resources from the statutory subsidy of Faculty of Mechanical Engineering, Silesian University of Technology.

References

1. Boutron P., Mehl P., 1990, Theoretical prediction of devitrification tendency: Determination of critical warming rates without using finite expansions, *Cryobiology*, **27**, 4, 359-377
2. Caniani D., Lioi D.S., Mancini I.M., Masi S., 2011, Application of fuzzy logic and sensitivity analysis for soil contamination hazard classification, *Waste Management*, **31**, 3, 583-594
3. Desai P.D., 1986, Thermodynamic properties of iron and silicon, *Journal of Physical and Chemical Reference Data*, **15**, 3, 967-983
4. Fourier J.B.J., 1882, *Théorie Analytique de la Chaleur*, Firmin Didot

5. Giachetti R.E., Young R.E., 1997, A parametric representation of fuzzy numbers and their arithmetic operators, *Fuzzy Sets and Systems*, **91**, 2, 185-202

6. Glassbrenner C.J., Slack G.A., 1964, Thermal conductivity of silicon and germanium from 3°K to the melting point, *Physical Review*, **134**, 4A, A1058-AI069

7. Guerra M.L., Stefanini L., 2005, Approximate fuzzy arithmetic operations using monotonic interpolations, *Fuzzy Sets and Systems*, **150**, 1, 5-33

8. Hanss M., 2005, *Applied Fuzzy Arithmetic*, Springer, Berlin

9. Jang T.H., Park S.C., Yang J.H., Kim J.Y., Seok J.H., Park U.S., Choi C.W., Lee S.R., Han J., 2017, Cryopreservation and its clinical applications, *Integrative Medicine Research*, **6**, 1, 12-18

10. Lü H., Shangguan W.-B., Yu D., 2017, Uncertainty quantification of squeal instability under two fuzzy-interval cases, *Fuzzy Sets and Systems*, **328**, 70-82

11. MeGlobalTM, 2008, Ethylene Glycol Product Guide, The MEGlobal Group of Companies, 20–21

12. Mochnacki B., Suchy J., 1993, *Modeling and Simulation of Foundry Solidification* (in Polish), Wydawnictwo Naukowe PWN, Warszawa

13. Moore R.E., 1966, *Interval Analysis*, Printice-Hall, New Jersey

14. Piasecka-Belkhayat A., Korczak A., 2020, Analysis of ultrashort laser pulse irradiation with 2D thin metal films using the fuzzy lattice Boltzmann method, *Journal of Theoretical and Applied Mechanics*, **58**, 1, 209-219

15. Piasecka-Belkhayat A., Skorupa A., 2023, Crystallisation degree analysis during cryopreservation of biological tissue applying interval arithmetic, *Materials*, **16**, 6

16. Shi M., Feng S., Zhang X., Ji C., Xu F., Lu T. J., 2018, Droplet based vitrification for cell aggregates: Numerical analysis, *Journal of the Mechanical Behavior of Biomedical Materials*, **82**, 383-393

17. Skorupa A., 2023, Multi-scale modelling of heat and mass transfer in tissues and cells during cryopreservation including interval methods, Ph.D. Thesis, Silesian University of Technology, Gliwice

18. Song Y.S., Adler D., Xu F., Kayaalp E., Nureddin A., Anchan R.M., Maas R.L., Demirci U., 2010, Vitrification and levitation of a liquid droplet on liquid nitrogen, *Proceedings of the National Academy of Sciences*, **107**, 10, 4596-4600

19. Tan M., Mei J., Xie J., 2021, The formation and control of ice crystal and its impact on the quality of frozen aquatic products: a review, *Crystals*, **11**, 68

20. Tuckerman D.B., Pease R.F.W., 1981, High-performance heat sinking for VLSI, *IEEE Electron Device Letters*, **2**, 5, 126-129

21. Wang C., Matthies H.G., 2021, Coupled fuzzy-interval model and method for structural response analysis with non-probabilistic hybrid uncertainties, *Fuzzy Sets and Systems*, **417**, 171-189

22. Zadeh L.A., 1965, Fuzzy sets, *Information and Control*, **8**, 3, 338-353

23. Zhang Y., Zhao G., Chapal Hossain S.M., He X., 2017, Modeling and experimental studies of enhanced cooling by medical gauze for cell cryopreservation by vitrification, *International Journal of Heat and Mass Transfer*, **114**, 1-7

24. Zhao G., Fu J., 2017, Microfluidics for cryopreservation, *Biotechnology Advances*, **35**, 2, 323-336

25. Zhou X., Liu Z., Liang X.M., Shu Z., Du P., Gao D., 2013, Theoretical investigations of a novel microfluidic cooling/warming system for cell vitrification cryopreservation, *International Journal of Heat and Mass Transfer*, **65**, 381-388

Manuscript received October 24, 2023; accepted for print November 30, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 219-230, Warsaw 2024
https://doi.org/10.15632/jtam-pl/184626

TUNING OF THE EQUILIBRATED RESIDUAL METHOD FOR APPLICATIONS IN GENERAL, DIRECT AND INVERSE PIEZOELECTRICITY[1]

Grzegorz Zboiński

Institute of Fluid Flow Machinery, Polish Academy of Sciences, Gdańsk, Poland, and
University of Warmia and Mazury, Faculty of Technical Sciences, Olsztyn, Poland
e-mail: zboi@imp.gda.pl

This paper presents application and tuning of the equilibrated residual method (ERM) of a posteriori error estimation for coupled electromechanical problems of direct, inverse and general piezoelectricity. In these three cases, either electric potential is induced by strains or strains appear due to the applied electric potential or both phenomena occur simultaneously. The mentioned ERM is assigned for the assessment of modeling and approximation errors of the numerical finite element solution. Such error values usually serve as indication for adaptive hierarchical modeling and adaptive mesh changes within thin and/or solid piezoelectric members so as to obtain the solution of assumed accuracy.

Keywords: coupled problems, piezoelectricity, finite element method, a posteriori error estimation, equilibrated residual method

1. Introduction

The origins and development of the residual equilibrated approach to a posteriori error estimation can be attributed to Ladeveze and Leguillon (1983), Kelly (1984), Bank and Weisser (1985), and Ainsworth and Oden (1993c). Implementation of the equilibrated residual method (ERM) to error estimation of finite element solutions was presented by Ainsworth and Oden (1992). The method was applied to elliptic problems by the same authors in (1993a,b). In 1994, they used the method for analysis of elasticity problems. Application of the method to thin- or thick-walled elastic structures was performed by Oden and Cho (1996), and Zboiński (2013) as well. The recent works on the method concern: stability analysis (Ainsworth *et al.*, 2007), generalizations to singularly perturbed reaction-diffusion problems (Ainsworth and Babuska, 1999), and application to conforming, non-conforming and discontinuous Galerkin finite element methods (Ainsworth, 2005). Application of the method to dielectricity (elliptic) and piezoelectricity (coupled) problems was suggested in (Zboiński, 2018). Recently (Zboiński, 2020), tuning of the method was performed in the case of thin elastic structures and suggested for dielectric and piezoelectric domains. It results from the above works that effective application of ERM may need its tuning by a modified definition of ERM local problems. Particularly, thin-walled elastic and piezoelectric domains need such procedures. In this context, the novelty and scope of the paper include: presentation of ERM for coupled problems as exemplified by piezoelectricity, and introduction of the tuning procedure to thick- or thin-walled piezoelectric domains.

[1]Paper presented during PCM-CMM 2023

2. Model problems

Let us start with the functional of electromechanical potential energy defined within volume V which may represent any bounded three-dimensional piezoelectric domain. In this paper, we limit it to symmetric-thickness, thin- or thick-walled domains defined in a standard way (Zboiński, 2010, 2019) with the use of mid-surface and thickness concepts as this is a typical geometry of piezoelectric transducers (actuators or sensors). After taking the first variation of this functional, one can obtain

$$\begin{aligned}&\int\limits_V (-D^{ijkl}\varepsilon_{kl}\delta\varepsilon_{ij} + C^{ijk}E_k\delta\varepsilon_{ij} + f^i\delta u_i)\,dV + \int\limits_P p^i\delta u_i\,dS\\ &\quad + \int\limits_V (\gamma^{ij}E_j\delta E_i + C^{ikl}\varepsilon_{kl}\delta E_i)\,dV - \int\limits_Q c\delta\phi\,dS = 0\end{aligned} \tag{2.1}$$

With the strain and electric field definitions, weak formulation (2.1) becomes a functional of $\mathbf{u} = \{u_i\}$, $i = 1, 2, 3$ and ϕ, where $\mathbf{u}$ and $\delta\mathbf{u}$ denote the solution and virtual (or admissible) displacements, while ϕ and $\delta\phi$ denote the solution and virtual (or admissible) electric potential. Note that $\delta\mathbf{u} \in \mathbf{w} + U$, with $\mathbf{w}$ being the lift (given displacements) of Dirichlet data (Demkowicz, 2007), and $U = \{\delta\mathbf{u} \in (H^1(V))^3 : \ \delta\mathbf{u} = \mathbf{0} \text{ on } W\}$ representing the space of kinematically admissible displacements within the domain V. Also, $\delta\phi \in \chi + \Phi$, with χ being the lift (given potential) of Dirichlet data, and $\Phi = \{\delta\phi \in H^1(V) : \ \delta\phi = \mathbf{0} \text{ on } F\}$ representing the space of electrically admissible potentials in V. The searched coupled solution belongs to $(\mathbf{u}, \phi) \in U \times \Phi$, i.e.

$$\begin{aligned}&B(\delta\mathbf{u}, \mathbf{u}) - C(\delta\mathbf{u}, \phi) = L(\delta\mathbf{u})\\ &- C(\delta\phi, \mathbf{u}) - b(\delta\phi, \phi) = -l(\delta\phi)\end{aligned} \tag{2.2}$$

where the bilinear, coupling and linear forms are

$$\begin{aligned}B(\delta\mathbf{u}, \mathbf{u}) &= \int\limits_V D^{ijkl}\varepsilon_{kl}\delta\varepsilon_{ij}\,dV = \int\limits_V D^{ijkl}u_{k,l}\delta u_{i,j}\,dV\\ C(\delta\mathbf{u}, \phi) &= \int\limits_V C^{ijk}E_k\delta\varepsilon_{ij}\,dV = \int\limits_V C^{ijk}E_k\delta u_{i,j}\,dV\\ L(\delta\mathbf{u}) &= \int\limits_V f^i\delta u_i dV + \int\limits_P p^i\delta u_i\,dS\\ b(\delta\phi, \phi) &= \int\limits_V \gamma^{ij}E_j\delta E_i\,dV = \int\limits_V \gamma^{ij}\phi_{,j}\delta\phi_{,i}\,dV\\ C(\delta\phi, \mathbf{u}) &= \int\limits_V C^{ikl}\varepsilon_{kl}\delta E_i\,dV = \int\limits_V C^{ikl}\varepsilon_{kl}\delta\phi_{,i}\,dV\\ l(\delta\phi) &= \int\limits_Q c\delta\phi\,dS = 0\end{aligned} \tag{2.3}$$

and represent the first variations of (or virtual) strain, electric field and coupling energies, respectively, B, b and C, and the first variations of (or virtual) works, L and l, of the external forces and charges, respectively. With the reciprocity theorem (Ieşan, 1990), one can convert the above functional into the corresponding local (strong) formulation (Zboiński, 2016). Existence and uniqueness of the solution to problem (2.2) and (2.3) is based on the Lax-Milgram theorem (cf. Cimatti, 2004).

In Eqs. (2.3), ε_{kl} and D^{ijkl}, $i,j,k,l=1,2,3$ stand for the strain and elastic constant tensors, while f^i represent components of the mass load vector $\mathbf{f}$. Additionally, d^i, E_j, $i,j=1,2,3$ are the electric displacement and electric field vectors. The tensors γ^{ij}, C^{kij} and C^{ikl}, $i,j,k,l=1,2,3$ represent dielectric and piezoelectric constants under constant strain. Additionally, p^i are components of the surface load $\mathbf{p}$, while $-c$ stands for the surface charge. Finally, P, W, Q and F denote the loaded, supported, charged and grounded parts of the boundary $S\equiv\partial V$ of the body domain. We assume $S\equiv P\cup W=Q\cup F$.

After finite element hpq- and $h\pi\rho$-approximation (Zboiński, 2016, 2018), where h, p, q and π, ρ are the element size, and longitudinal and transverse approximation orders within the displacement and electric potential fields, respectively, above relations (2.2) and (2.3) can be defined with the aproximation of the solution quantities and spaces, i.e. $(\mathbf{u}^{hpq},\phi^{h\pi\rho})\in U^{hpq}\times\Phi^{h\pi\rho}$.

In the case of direct piezoelectricity (sensing), the external, mechanical (volume and/or surface) loadings produce strains which in turn induce electric potential within the piezoelectric member. No external electric charges are present in this case. As a result, the assumption of $c=0$ has to be substituted into the second line of (2.1). In the case of inverse piezoelectricity (actuation), the mechanical loadings are not present – $\mathbf{f}=\{f^i\}=\mathbf{0}$ and $\mathbf{p}=\{p^i\}=\mathbf{0}$ in the first line of (2.1).

With the use of decoupling assumptions of $C^{ijk}=0$, $C^{ikl}=0$, $i,j,k=1,2,3$, functional (2.1) is replaced with two independent mechanical and electric potential energy functionals.

3. ERM a posteriori error estimation

In this Section, we present our original results of implementetion of the equilibrated residual method (ERM) of a posteriori error estimation to the coupled problems of piezoelectricity. The global η and element $\overset{e}{\eta}$ error estimators are defined by us as

$$\begin{aligned}\eta=\sum_e\overset{e}{\eta}=\sum_e\Bigg[&-\overset{e}{\Pi}(\mathbf{u},\phi)-\int\limits_{S_e\setminus S}\mathbf{u}^{\mathrm{T}}\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle\,d\overset{e}{S}+\int\limits_{S_e\setminus S}\phi\langle\overset{e}{h}(\phi^{h\pi\rho})\rangle\,d\overset{e}{S}\\&-\frac{1}{2}\overset{e}{B}(\mathbf{u}^{hpq},\mathbf{u}^{hpq})+\frac{1}{2}\overset{e}{C}(\mathbf{u}^{hpq},\phi^{h\pi\rho})+\frac{1}{2}\overset{e}{C}(\phi^{h\pi\rho},\mathbf{u}^{hpq})+\frac{1}{2}\overset{e}{b}(\phi^{h\pi\rho},\phi^{h\pi\rho})\Bigg]\end{aligned}\tag{3.1}$$

Note that in (3.1), the sum (over elements e) of terms of the first line represents electromechanical potential energy $\mathit{\Pi}(\mathbf{u},\phi)$ of the exact solution $(\mathbf{u},\phi)$, while the sum of terms of the second line the analogous energy $\mathit{\Pi}(\mathbf{u}^{hpq},\phi^{h\pi\rho})$ of the numerical solution $(\mathbf{u}^{hpq},\phi^{h\pi\rho})$, the error of which is a posteriori estimated. In the first line, the potential energy is defined as a sum of the strain, electric field and coupling energies deminished by the work of external forces and charges, i.e. $\mathit{\Pi}=0.5B-C-0.5b-L+l$. After the partitioning of the domain V into finite elements of volumes $\overset{e}{V}$, $\partial\overset{e}{V}=\overset{e}{S}$, this energy has to be completed by the work of the internal (interelement) forces and charges represented by the last two components of the first line. The terms $\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle$ and $\langle\overset{e}{h}(\phi^{h\pi\rho})\rangle$ represent the equilibrated interelement stress reaction vectors and the equilibrated interelement equivalent electric charge, respectively, typical for the equilibrated residual method.

In the second line of (3.1), the equivalent definition, $\mathit{\Pi}=-0.5B+C+0.5b$, of the electromechanical potential energy is applied. It needs introduction of stationarity results (2.2), with $\delta\mathbf{u}$ and $\delta\phi$ replaced by $\mathbf{u}$ and ϕ, i.e. $L=B-C$ and $-l=-b-C$ into the former definition so as to eliminate the external work from the modified potential energy definition before the partitioning.

As the exact solution $(\mathbf{u},\phi)$ of (3.1) can hardly be found, we approximate this functional with the finite element method and search for the approximation $(\mathbf{u}^{HPQ},\phi^{H\Pi P})$ of the exact solution. Here, H, P, Q and Π, P stand for the element size, and longitudinal and transverse approximation orders within the displacement and electric potential fields, respectively. After taking the first variation of approximated functional (3.1) with respect to $\delta\mathbf{u}^{HPQ}$ and $\delta\phi^{H\Pi P}$ and equating it to zero, one obtains the following global stationarity condition

$$0=\sum_e\left[-\delta\overset{e}{\Pi}(\mathbf{u}^{HPQ},\phi^{H\Pi P})-\int\limits_{S_e\setminus S}(\delta\mathbf{u}^{HPQ})^{\mathrm{T}}\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle\,d\overset{e}{S}+\int\limits_{S_e\setminus S}\delta\phi^{H\Pi P}\langle\overset{e}{h}(\phi^{h\pi\rho})\rangle\,d\overset{e}{S}\right] \tag{3.2}$$

equivalent to the following element (or local) sets of two coupled conditions (cf. Zboiński, 2016)

$$\begin{aligned}&\overset{e}{B}(\delta\mathbf{u}^{HPQ},\mathbf{u}^{HPQ})-\overset{e}{C}(\delta\mathbf{u}^{HPQ},\phi^{H\Pi P})=\overset{e}{L}(\delta\mathbf{u}^{HPQ})+\int\limits_{\overset{e}{S}\setminus S}(\delta\mathbf{u}^{HPQ})^{\mathrm{T}}\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle\,d\overset{e}{S}\\&\overset{e}{C}(\delta\phi^{H\Pi P},\mathbf{u}^{HPQ})+\overset{e}{b}(\delta\phi^{H\Pi P},\phi^{H\Pi P})=\overset{e}{l}(\delta\phi^{H\Pi P})+\int\limits_{\overset{e}{S}\setminus S}\delta\phi^{H\Pi P}\langle\overset{e}{h}(\phi^{h\pi\rho})\rangle\,d\overset{e}{S}\end{aligned} \tag{3.3}$$

where the first definition of the potential energy was utilized on the element level. The searched coupled solution belongs to $(\mathbf{u}^{HPQ},\phi^{H\Pi P})\in U^{HPQ}\times\Phi^{H\Pi P}$, where $\delta\mathbf{u}^{HPQ}\in\mathbf{w}^{HPQ}+U^{HPQ}$, with $\mathbf{w}^{HPQ}$ being the approximated lift of Dirichlet data, and $U^{HPQ}=\{\delta\mathbf{u}^{HPQ}\in(H^1(\overset{e}{V}))^3:\ \delta\mathbf{u}^{HPQ}=\mathbf{0}\text{ on }W\cap d\overset{e}{S}\}$ representing the local (element) space of kinematically admissible displacements within the domain $\overset{e}{V}$. Also, $\delta\phi^{H\Pi P}\in\chi^{H\Pi P}+\Phi^{H\Pi P}$, with $\chi^{H\Pi P}$ being the lift of Dirichlet data, and $\Phi^{H\Pi P}=\{\delta\phi^{H\Pi P}\in H^1(\overset{e}{V}):\ \delta\phi^{H\Pi P}=\mathbf{0}\text{ on }F\cap d\overset{e}{S}\}$ representing the element space of electrically admissible potentials. The above set (3.3) can also be written in the finite element language (Zboiński, 2016, 2018).

The above set (3.3) corresponds to the general piezoelectricity case. The cases of direct or inverse piezoelectricity need neglecting the works of external forces $\overset{e}{l}$ or $\overset{e}{L}$, respectively. For the decoupled problems of elasticity and dielectricity, one needs to neglect coupling energies $\overset{e}{C}$ in both equations (3.3) so as to obtain two independent equations for these cases.

4. ERM local problems determination

It can be demonstrated that above coupled local problems (3.3) can be either Dirichlet ($\overset{e}{S}\cap W\neq\emptyset$ and $\overset{e}{S}\cap F\neq\emptyset$) or Neumann ($\overset{e}{S}\cap W=\emptyset$ and $\overset{e}{S}\cap F=\emptyset$ or mixed (Dirichlet--Neumann of two-types: $\overset{e}{S}\cap W=\emptyset$ and $\overset{e}{S}\cap F\neq\emptyset$ or $\overset{e}{S}\cap W\neq\emptyset$ and $\overset{e}{S}\cap F=\emptyset$). The Dirichlet local problems are well-posed (they are solvable by their definition).

In the case of the Neumann displacement boundary conditions, the local piezoelectric problems are solvable provided that the external and internal load compatibility condition is valid (Ainsworth and Oden, 1993c)

$$-\overset{e}{B}(\mathbf{u}^{hpq},\mathbf{1})+\overset{e}{C}(\phi^{hpq},\mathbf{1})+\overset{e}{L}(\mathbf{1})+\int\limits_{\overset{e}{S}\setminus S}\mathbf{1}^{\mathrm{T}}\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle\,d\overset{e}{S}=0 \tag{4.1}$$

where $\mathbf{1} = (1,1,1)^{\mathrm{T}}$. In the case of the Neumann electric potential boundary condition, we suggest that the external and equivalent charge compatibility condition holds, i.e.

$$\overset{e}{b}(\phi^{h\pi\rho},1)+\overset{e}{C}(\mathbf{u}^{hpq},1)-\overset{e}{l}(1)+\int\limits_{\overset{e}{S}\backslash S}1\langle\overset{e}{h}(\phi^{h\pi\rho})\rangle\,d\overset{e}{S}=0 \tag{4.2}$$

5. Linear and higher-order equilibration

5.1. Equilibrated interelement stress reactions and equivalent charges

Here the linear-equilibration method of Ainsworth and Oden (1993a) and Ainsworth *et al.* (1994) for elliptic (elasticity) problems is utilized. We extend it to the coupled problems (piezoelectricity). In the method, the unknown vectors of the equilibrated interelement stress reactions $\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle$ are defined (Ainsworth and Oden, 1993b; Ainsworth *et al.*, 1994), with the displacements $\mathbf{u}^{hpq}$ from the global problem, i.e.

$$\langle\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})\rangle=\overset{fee}{\boldsymbol{\alpha}}\mathbf{r}(\mathbf{u}^{hpq})+\overset{eff}{\boldsymbol{\alpha}}\mathbf{r}(\mathbf{u}^{hpq}) \tag{5.1}$$

and

$$\overset{e}{\mathbf{r}}(\mathbf{u}^{hpq})=\mathbf{H}(\overset{e}{\boldsymbol{\nu}})\,\overset{e}{\boldsymbol{\sigma}}(\mathbf{u}^{hpq})\qquad\qquad\overset{f}{\mathbf{r}}(\mathbf{u}^{hpq})=\mathbf{H}(\overset{e}{\boldsymbol{\nu}})\,\overset{f}{\boldsymbol{\sigma}}(\mathbf{u}^{hpq}) \tag{5.2}$$

with

$$\mathbf{H}(\overset{e}{\boldsymbol{\nu}})=\begin{bmatrix}\nu_1 & 0 & 0 & \nu_2 & 0 & \nu_3\\ 0 & \nu_2 & 0 & \nu_1 & \nu_3 & 0\\ 0 & 0 & \nu_3 & 0 & \nu_2 & \nu_1\end{bmatrix} \tag{5.3}$$

The vector $\overset{e}{\boldsymbol{\nu}}=[\nu_1,\nu_2,\nu_3]^{\mathrm{T}}$ denotes the normal unit vector, outward to S_e. The terms $\overset{e}{\boldsymbol{\sigma}}$ and $\overset{f}{\boldsymbol{\sigma}}$ represent six-component element stress vectors of the element e and its any neighbour f. The splitting functions are defined with their directional components, i.e. $\overset{fe}{\boldsymbol{\alpha}}=\operatorname{diag}[\alpha_1,\alpha_2,\alpha_3]$, with $\overset{fe}{\boldsymbol{\alpha}}=\mathbf{1}-\overset{ef}{\boldsymbol{\alpha}}$ and $\mathbf{1}=\operatorname{diag}[1,1,1]$. In the case of the first-order equilibration performed within the parametric elements, it is sufficient to define the splitting functions $\overset{fe}{\boldsymbol{\alpha}}$ as linear ones, with the use of the vertex nodes splitting factors $\overset{fe}{\boldsymbol{\alpha}}_k$, $k=1,2,\ldots,K$ of the applied parametrized prismatic ($K=6$) element

$$\overset{fe}{\boldsymbol{\alpha}}=\sum_k\overset{fe}{\boldsymbol{\alpha}}_k\,\lambda_k \tag{5.4}$$

where λ_k represents the vertex node shape functions of the element.

In the electric field, the unknown scalar equilibrated interelement equivalent charge $\langle\overset{e}{h}(u^{h\pi\rho})\rangle$ is proposed by us to be determined by the scalar electric potential $\phi^{h\pi\rho}$ taken from the global problem

$$\langle\overset{e}{h}(\phi^{h\pi\rho})\rangle=\overset{fee}{\beta}h(\phi^{h\pi\rho})+\overset{eff}{\beta}h(\phi^{h\pi\rho}) \tag{5.5}$$

Above

$$\overset{e}{h}(\phi^{h\pi\rho})=\overset{e}{\boldsymbol{\nu}}\mathbf{d}^{e}(\phi^{h\pi\rho})\qquad\qquad\overset{f}{h}(\phi^{h\pi\rho})=\overset{e}{\boldsymbol{\nu}}\mathbf{d}^{f}(\phi^{h\pi\rho}) \tag{5.6}$$

The terms $\overset{e}{\mathbf{d}}$ and $\overset{f}{\mathbf{d}}$ are three-component electric displacement vectors of the element e and its any neighbour f. The quantity $\overset{fe}{\beta}$ is the scalar splitting function, while $\overset{fe}{\beta} = 1 - \overset{ef}{\beta}$. For the first-order equilibration, we suggest to define the splitting function $\overset{fe}{\beta}$ as a linear one, by means of the vertex nodes splitting factors $\overset{fe}{\beta}_k$, i.e.

$$\overset{fe}{\beta} = \sum_k \overset{fe}{\beta}_k \, \lambda_k \tag{5.7}$$

where $k = 1, 2, \ldots, K$ and $K = 6$ again.

5.2. Determination of the splitting factors

The procedure starts with the standard version of the first-order equilibration condition for elasticity (see Ainsworth and Oden, 1993a; Ainsworth *et al.*, 1994) extended by us to the case of piezoelectricity

$$\begin{aligned} &- \overset{e}{B}(\mathbf{u}^{hpq}, \boldsymbol{\lambda}_k) + \overset{e}{C}(\phi^{hpq}, \boldsymbol{\lambda}_k) + \overset{e}{L}(\boldsymbol{\lambda}_k) + \int\limits_{\overset{e}{S}\setminus S} \boldsymbol{\lambda}_k^{\mathrm{T}} \langle \overset{e}{\mathbf{r}}(\mathbf{u}^{hpq}) \rangle \, d\overset{e}{S} = 0 \\ &\overset{e}{b}(\phi^{h\pi\rho}, \lambda_k) + \overset{e}{C}(\mathbf{u}^{hpq}, \lambda_k) - \overset{e}{l}(\lambda_k) + \int\limits_{\overset{e}{S}\setminus S} \lambda_k \langle \overset{e}{h}(\phi^{h\pi\rho}) \rangle \, d\overset{e}{S} = 0 \end{aligned} \tag{5.8}$$

by taking into consideration the coupling form $\overset{e}{C}$ in mechanical condition $(5.8)_1$ and adding electrical condition $(5.8)_2$, where $\boldsymbol{\lambda}_k = \operatorname{diag}[\lambda_k, \lambda_k, \lambda_k]$ due to vectorial character of the displacement field. It is worth noticing that $\sum_{k=1}^{6} \lambda_k = 1$, i.e. the sums of the first and second equation (5.8) gives (4.1) and (4.2), and the load and/or charge compatibility conditions are fulfilled for the elements in the Neumann or mixed (Dirichlet-Neumann) local problems.

Taking advantage of (5.1) and (5.4), and (5.5) and (5.7) as well, substituted into (5.8), we get

$$\begin{aligned} 0 = &- \overset{e}{B}(\mathbf{u}^{hpq}, \boldsymbol{\lambda}_k) + \overset{e}{C}(\phi^{h\pi\rho}, \boldsymbol{\lambda}_k) + \overset{e}{L}(\boldsymbol{\lambda}_k) \\ &+ \sum_f \left[\overset{fe}{\boldsymbol{\alpha}}_k \int\limits_{\overset{ef}{S}} \boldsymbol{\lambda}_k \, \overset{e}{\mathbf{r}}(\mathbf{u}^{hpq}) \, d\overset{ef}{S} + \overset{ef}{\boldsymbol{\alpha}}_k \int\limits_{\overset{ef}{S}} \boldsymbol{\lambda}_k \, \overset{f}{\mathbf{r}}(\mathbf{u}^{hpq}) \, d\overset{ef}{S} \right] \\ 0 = &\; \overset{e}{b}(\phi^{h\pi\rho}, \lambda_k) + \overset{e}{C}(\mathbf{u}^{hpq}, \lambda_k) - \overset{e}{l}(\lambda_k) \\ &+ \sum_f \left[\overset{fe}{\beta}_k \int\limits_{\overset{ef}{S}} \lambda_k \, \overset{e}{h}(\phi^{h\pi\rho}) \, d\overset{ef}{S} + \overset{ef}{\beta}_k \int\limits_{\overset{ef}{S}} \lambda_k \, \overset{f}{h}(\phi^{h\pi\rho}) \, d\overset{ef}{S} \right] \end{aligned} \tag{5.9}$$

where $\overset{fe}{\boldsymbol{\alpha}}_k$ includes three directional stress splitting factors at node k of the element e, while $\overset{fe}{\beta}_k$ denotes scalar charge splitting factor for node k of the element e. Above, the integration over the internal part of element boundary $\overset{e}{S}\setminus S$ from (5.8) was replaced with the integrations over the common sides $\overset{ef}{S}$ of the element e and any of its neighbours f. It is worth noticing that three displacement and one potential equations (5.9) are independent. The procedure for calculation of the four splitting factors may be proposed to take advantage of the sets of equations (5.9) written for the element patches composed of elements surrounding any node of the domain V, at which the element vertex nodes meet (cf. Ainsworth and Oden, 1993b, Ainsworth *et al.*, 1994).

In the case of higher-order equilibration, relations (5.8) and (5.9) have to be modified by introduction of the shape functions $\lambda_{l,m}$ corresponding to any higher-order nodal dof (l,m) at the element edges and sides, where l stands for the edge or side number, and m defines the dof number at this edge or side, instead of the linear vertex node shape functions λ_k. The searched splitting factors $\overset{fe}{\boldsymbol{\alpha}}_{l,m}$ and $\overset{fe}{\beta}_{l,m}$ can be obtained from the sets of modified equations (5.8) and (5.9) written for the element patches composed of elements surrounding any edge or side node (l,m) of the domain V. The method is presented in (Zboiński, 2020) for the elasticity case. Its application to dielectricity is analogous.

6. Numerical experiments

In this Section, we will check the effectivity of ERM error estimation applied to coupled problems of piezoelectricity. We will show that such effectivity is different in the cases of direct, inverse and general piezoelectricity. These results will be compared to the analogous effectivity for the reference problems of uncoupled elasticty (elastostatics) and uncoupled dielectricity (electrostatics). In these tests, the global effectivity indices θ for the modeling, approximation and total errors of the model piezoelectric plate problem will be presented. Such indices are defined as a ratio of the estimated error, expressed by the ERM estimator η and the exact value of the potential energy error e:

$$\theta = \frac{\eta}{\mathrm{e}} \tag{6.1}$$

Three components e_M, e_C, e_E of the potential energy error e will be introduced by us, i.e. related to the mechanical, coupling and electric parts of this energy

$$\begin{aligned} &\Pi(\mathbf{u},\phi) - \Pi(\mathbf{u}^{hpq},\phi^{h\pi\rho}) \equiv \Pi(\mathbf{u}-\mathbf{u}^{hpq},\phi-\phi^{h\pi\rho}) = \frac{1}{2}B(\mathbf{u}-\mathbf{u}^{hpq},\mathbf{u}-\mathbf{u}^{hpq}) \\ &\quad - C(\mathbf{u}-\mathbf{u}^{hpq},\phi-\phi^{h\pi\rho}) - \frac{1}{2}b(\phi-\phi^{h\pi\rho},\phi-\phi^{h\pi\rho}) = \mathrm{e}_M - \mathrm{e}_C - \mathrm{e}_E = \mathrm{e} \end{aligned} \tag{6.2}$$

Above, the energy errors are defined as differences between potential energies. However, they and their components are proposed to be equivalently expressed by energies defined on differences $\mathbf{u}-\mathbf{u}^{hpq}$ and $\phi-\phi^{h\pi\rho}$ of the exact and numerical solutions. Derivation of the above equivalent formula required utilization of the potential energy definitions obtained from (2.1) and (2.3) by means of replacement of $\delta\mathbf{u}$ and $\delta\phi$ by $\mathbf{u}$ and ϕ and addition of coefficient 0.5 before the forms B and b. Then, elimination of the work of external forces L and charges l with use of stationarity conditions (2.2) was performed (with $\delta\mathbf{u}$ and $\delta\phi$ replaced by $\mathbf{u}$ and ϕ again). Finally, taking advantage of the mathematical properties of the quadratic forms B and b and mixed forms C was necessary.

For the thin- or thick-walled piezoelectric members considered in the paper, the hierarchical modelling is proposed by us (cf. Zboiński, 2010, 2016, 2018, 2019) where the mechanical, electric and electromechanical cases are considered. Such modelling implies division of the total energy error $\mathrm{e} \equiv \mathrm{e}^t$ into its modeling e^m and approximation e^a parts in accordance with the following relation describing the total, modeling and approximation errors of the solution displacements and electric potential

$$\begin{aligned} \mathbf{e}^t &= \mathbf{u}-\mathbf{u}^{hpq} = (\mathbf{u}-\mathbf{u}^q) + (\mathbf{u}^q-\mathbf{u}^{hpq}) = \mathbf{e}^m + \mathbf{e}^a \\ e^t &= \phi-\phi^{h\pi\rho} = (\phi-\phi^\rho) + (\phi^\rho-\phi^{h\pi\rho}) = e^m + e^a \end{aligned} \tag{6.3}$$

where $(\mathbf{u}^q,\phi^\rho)$ represents the exact solution to the hierarchical electromechanical (piezoelectric) model of the order (q,ρ) (Zboiński, 2016), with q and ρ denoting the mechanical and electric field transverse orders.

Substitution of (6.3) into (6.2) leads to nine global error components: e_M^t, e_M^m, e_M^a, e_C^t, e_C^m, e_C^a, e_E^t, e_E^m, e_E^a contributing to the energy error e. Also, the analogous division of the global error estimator η introduced in (3.1) is applied in the paper. Thus, nine resulting component effectivity indices can be defined. The analogous component quantities are defined for finite elements.

In the case of the total e^t and approximation e^a error calculations, the unknown values of the exact solutions $(\mathbf{u}, \phi)$ and $(\mathbf{u}^q, \phi^\rho)$ are replaced with their best numerical approximations $(\mathbf{u}^{ref}, \phi^{ref})$ and $(\mathbf{u}^{mod}, \phi^{mod})$, respectively, obtained from the *hpq*- and $h\pi\rho$-approximated version of (2.2). For the total η^t (or approximation η^a) error estimator, relations (6.2) and (6.3) hold, with $\mathbf{u}$ and ϕ (or $\mathbf{u}^q$ and ϕ^ρ) replaced by their proper ERM approximations $\mathbf{u}^{HPQ}$ and $\phi^{H\Pi P}$. Global values of the modeling error and estimator are obtained from $\mathrm{e}^m = \mathrm{e}^t - \mathrm{e}^a$ and $\eta^m = \eta^t - \eta^a$.

6.1. Model problem

The applied model problem concerns a uniformly loaded, hardly clamped, square piezoelectric (piezoceramic) plate. The plate is charged on its top surface, and grounded around its lateral sides. The length of the plate is equal to $l = 3.1415 \cdot 10^{-2}$ m. The plate thickness is $t = 0.15 \cdot 10^{-2}$ m. Young's modulus of the piezoelectric is $E = 0.5 \cdot 10^{11}$ N/m^2. Poisson's ratio equals 0.294. The dielectric permitivity (isotropic dielectric constant) under constant stress is $\delta = 0.1593 \cdot 10^{-7}$ F/m. The non-zero anisotropic piezoelectricity constants under constant stress are equal to: $c_{13} = c_{23} = -0.15 \cdot 10^{-9}$ C/N, $c_{33} = 0.3 \cdot 10^{-9}$ C/N, and $c_{52} = c_{61} = 0.5 \cdot 10^{-9}$ C/N. The way the measurable dielectric and piezoelectric constants under constant stress can be converted into the corresponding constants under constant strain, present in (2.1)-(2.3), can be found in (Preumont, 2006; Zboiński, 2020). The vertical pressure load is equal to $p = 4.0 \cdot 10^6$ N/m^2. The uniform charges applied to the top surface are equal to $c = 0.2 \cdot 10^{-1}$ C/m^2. Due to symmetry of the geometry, load, charge and boundary conditions, only a quarter of dimensions $l/2 \times l/2 \times t$ of the plate is analysed.

Due to space limitation, the applied electromechanical model is limited to one hierarchical model of orders $q = 2$ and $\rho = 2$. This is possible as the solution results for all models $q \geqslant 2$ and $\rho \geqslant 2$ is very close (qualitatively almost identical). Because of the same reason (qualitative similarity observed), only one examplary mesh $3 \times 3 \times 2$ of prismatic elements, is applied for three (direct, inverse and general) piezoelectricity cases and two uncoupled cases of elasticity and dielectrciity.

Our effectivity calculations are performed for changing values ($p = \pi = 2, 3, \ldots, p_{max} = \pi_{max}$, $p_{max} = \pi_{max} = 7$ or 8) of the longitudinal, displacement and electric potential, orders of approximation, as the error estimation is most sensitive to changes in these discretization parameters. The approximations $(\mathbf{u}^{ref}, \phi^{ref})$ and $(\mathbf{u}^{mod}, \phi^{mod})$ of the exact solutions are obtained from (3.1) with $m = 9$, $p = \pi = 9$ and $q = \rho = 6$ and $m = 9$, $p = \pi = 9$ and $q = \rho = 2$, respectively, where $m = l/2h$ and h is the characteristic length of the applied prismatic elements.

6.2. Results

In Table 1, the reference values of effectivity indices for two decoupled problems are presented. The ERM local problems results were obtained with initial tunning within the mechanical field due to thin-walled character of the plate domain. 18 vertex degrees of freedom within each element were constrained instead of constraining 6 such dofs and linear equilibration (cf. Zboiński, 2020). In the initial tuning, the longitudinal and transverse orders of approximation in ERM local problems were increased by 1 with respect to global problems for both the fields. For $p \geqslant 3$ or $\rho \geqslant 3$, all effectivities are close to the desired value of 1.0, i.e. the estimated errors are very close to the true errors.

Table 1. Global effectivities for elasticity (E) and dielectricity (D) cases – global problem parameters: (E) $q = 2$, $p = \text{var}$, $m = 3$, (D) $\rho = 2$, $\pi = \text{var}$, $m = 3$; local problems characterization: (E) 18 dofs constrained within the mechanical field, $H = h$, $P = p + 1$, $Q = q + 1$; (D) 1 dof constrained and linear equilibration within the electric field, $H = h$, $\Pi = \pi + 1$, $P = \rho + 1$

¢Problem type	Estimator component	Component part	Effectivity symbol and values for varying p or π orders								
				1	2	3	4	5	6	7	8
uncoupled elasticity case	mechanical	total	θ_M^t	0.52	1.70	1.17	0.96	0.99	1.00	1.01	1.00
		approx.	θ_M^a	0.52	1.68	1.21	0.92	0.98	1.05	1.15	1.45
		modeling	θ_M^m	0.59	1.90	1.07	0.99	1.00	0.99	0.99	0.98
uncoupled dielectricity case	electric	total	θ_E^t	1.49	2.74	1.07	1.05	0.95	0.92	0.90	0.91
		approx.	θ_E^a	1.49	2.78	1.11	1.15	1.09	1.10	1.04	1.19
		modeling	θ_E^m	1.17	0.85	0.86	0.86	0.87	0.88	0.89	0.90

Table 2. Global effectivities for general (G), direct (D) and inverse (I) piezoelectricity cases – global problem parameters (G, D, I): $q = \rho = 2$, $p = \pi = \text{var}$, $m = 3$; local problems characterization (G, D, I)): 18 dofs constrained within the mechanical field, 1 dof constrained and linear equilibration within the electric field, $H = h$, $P = p + 1$, $Q = q + 1$, $\Pi = \pi + 1$, $P = \rho + 1$

¢Problem type	Estimator component	Component part	Effectivity symbol and values for following p or π							
				1	2	3	4	5	6	7
¢general piezoelectricity case	mechanical	total	θ_M^t	0.68	2.01	1.26	0.99	1.00	1.01	1.02
		approx.	θ_M^a	0.68	2.01	1.50	0.94	0.89	0.96	1.17
		modeling	θ_M^m	0.74	1.96	1.05	1.01	1.02	1.01	1.02
	coupling	total	θ_C^t	1.44	1.76	**0.81**	**0.72**	**0.76**	**0.73**	**0.70**
		approx.	θ_C^a	1.44	1.75	**0.79**	**0.73**	**0.77**	**0.85**	**1.06**
		modeling	θ_C^m	0.28	1.38	**0.37**	**0.32**	**0.39**	**0.30**	**0.35**
	electric	total	θ_E^t	1.13	1.58	**0.84**	**0.78**	**0.78**	**0.71**	**0.62**
		approx.	θ_E^a	1.13	1.63	**0.85**	**0.81**	**0.84**	**0.90**	**1.09**
		modeling	θ_E^m	0.67	0.83	**0.48**	**0.45**	**0.44**	**0.42**	**0.43**
¢direct piezoelectricity case	mechanical	total	θ_M^t	0.59	1.89	1.24	0.97	1.00	1.01	1.02
		approx.	θ_M^a	0.59	1.89	1.44	0.84	0.86	0.95	1.18
		modeling	θ_M^m	0.63	1.90	1.05	1.01	1.02	1.01	1.02
	coupling	total	θ_C^t	0.01	0.66	**0.80**	**0.74**	**0.77**	**0.74**	**0.73**
		approx.	θ_C^a	0.04	0.64	**0.79**	**0.74**	**0.77**	**0.85**	**1.06**
		modeling	θ_C^m	0.36	1.32	**0.49**	**0.45**	**0.48**	**0.40**	**0.44**
	electric	total	θ_E^t	0.73	1.28	**0.83**	**0.74**	**0.74**	**0.68**	**0.58**
		approx.	θ_E^a	0.73	1.33	**0.83**	**0.77**	**0.82**	**0.89**	**1.09**
		modeling	θ_E^m	0.21	0.79	**0.29**	**0.27**	**0.28**	**0.24**	**0.26**
¢inverse piezoelectricity case	mechanical	total	θ_M^t	1.45	3.09	**2.15**	**2.25**	**1.85**	**1.57**	**1.16**
		approx.	θ_M^a	1.44	3.10	**2.26**	**2.54**	**2.43**	**2.72**	**2.92**
		modeling	θ_M^m	8.82	3.66	1.27	1.17	0.94	0.89	0.83
	coupling	total	θ_C^t	1.43	2.82	1.10	1.14	1.06	1.05	1.02
		approx.	θ_C^a	1.43	2.83	1.15	1.21	1.17	1.22	1.32
		modeling	θ_C^m	1.14	0.99	1.08	1.09	1.02	1.00	1.00
	electric	total	θ_E^t	1.57	2.62	1.16	1.18	1.05	0.98	0.94
		approx.	θ_E^a	1.58	2.68	1.17	1.30	1.27	1.36	1.35
		modeling	θ_E^m	1.58	1.07	1.01	0.94	0.93	0.92	0.92

In Table 2, the effectivity results are presented for three piezoelectric problems. The initial tunning of ERM was applied within the mechanical and electric fields as for the referrence elasticity and dielectricity problems, i.e. 18 mechanical vertex dofs were constrained and the ERM local approximation orders were higher by 1 in comparison to global problems. No other (additional) tuning was applied. In the case of general and direct piezoelectricity problems, considerable underestimation (effectivities lower than 1.0) of the modelling, approximation and total errors can be observed for the coupling and electric parts of the error. On the countrary, in the case of the inverse piezoelectricity, substantial overestimation (effectivities higher than 1.0) of the approximation and total errors can be seen for the mechanical part of the error. All unsatisfactory values are shown in bold in the table, for $p = \pi \geqslant 3$.

Table 3. Global effectivities for general (G), direct (D) and inverse (I) piezoelectricity cases – global problem parameters (G, D, I): $q = \rho = 2$, $p = \pi =$ var, $m = 3$; local problems characterization: (G) 18 dofs constrained within the mechanical field, 1 dof constrained and linear equilibration within the electric field, $H = h$, $P = p+1$, $Q = q+1$, $\Pi = \pi+2$, $P = \rho+2$, (D) 18 dofs constrained within the mechanical field, 1 dof constrained and linear equilibration within the electric field, $H = h$, $P = p+2$, $Q = q+2$, $\Pi = \pi+2$, $P = \rho+2$, (I) 18 dofs constrained and higher-order equilibration within the mechanical field, 1 dof constrained and linear equilibration within the electric field, $H = h$, $P = p+1$, $Q = q+1$, $\Pi = \pi+1$, $P = \rho+1$

¢Problem type	Estimator component	Component part	Effectivity symbol and values for the following p or π							
				1	2	3	4	5	6	7
¢general piezoelectricity case	mechanical	total	θ_M^t	0.66	2.03	1.24	0.97	0.98	0.99	1.00
		approx.	θ_M^a	0.66	2.02	1.49	0.93	0.89	0.90	1.17
		modeling	θ_M^m	1.13	2.22	1.02	0.97	1.00	1.00	1.00
	coupling	total	θ_C^t	1.18	1.92	**0.85**	0.83	**0.87**	**0.92**	**0.97**
		approx.	θ_C^a	1.14	1.80	**0.79**	**0.73**	**0.77**	**0.85**	**1.06**
		modeling	θ_C^m	1.89	5.38	**0.91**	**0.93**	**0.88**	**0.89**	**0.87**
	electric	total	θ_E^t	1.44	2.42	**1.01**	**0.93**	**0.92**	**0.96**	**0.99**
		approx.	θ_E^a	1.45	2.54	**0.99**	**0.88**	**0.89**	**0.94**	**1.12**
		modeling	θ_E^m	1.90	4.41	**0.98**	**0.90**	**0.88**	**0.92**	**0.94**
¢direct piezoelectricity case	mechanical	total	θ_M^t	0.73	2.28	1.34	1.01	1.03	1.03	1.04
		approx.	θ_M^a	0.71	2.28	1.55	0.96	1.00	1.10	1.35
		modeling	θ_M^m	1.40	2.28	1.14	1.02	1.03	1.02	1.03
	coupling	total	θ_C^t	0.35	1.03	**0.89**	**0.93**	**0.96**	**1.09**	**1.24**
		approx.	θ_C^a	0.35	1.07	**0.87**	**0.93**	**0.94**	**1.04**	**1.27**
		modeling	θ_C^m	1.53	1.68	**0.89**	**0.86**	**1.04**	**1.19**	**1.28**
	electric	total	θ_E^t	0.90	1.58	**1.00**	**0.93**	**0.96**	**1.06**	**1.17**
		approx.	θ_E^a	0.89	1.53	**0.99**	**0.93**	**0.96**	**1.04**	**1.26**
		modeling	θ_E^m	1.82	2.74	**0.93**	**0.86**	**0.96**	**1.09**	**1.16**
¢inverse piezoelectricity case	mechanical	total	θ_M^t	1.44	2.74	**1.48**	**1.22**	**1.20**	**0.97**	**0.91**
		approx.	θ_M^a	1.44	2.74	**1.54**	**1.30**	**1.45**	**1.27**	**1.63**
		modeling	θ_M^m	8.71	3.14	0.92	0.90	0.85	0.82	0.81
	coupling	total	θ_C^t	1.43	2.86	1.05	1.09	1.01	1.01	1.00
		approx.	θ_C^a	1.43	2.87	1.12	1.20	1.12	1.17	1.20
		modeling	θ_C^m	0.47	1.07	1.09	1.12	1.02	1.01	1.00
	electric	total	θ_E^t	1.57	2.57	1.16	1.18	1.06	0.98	0.94
		approx.	θ_E^a	1.58	2.63	1.17	1.27	1.26	1.32	1.35
		modeling	θ_E^m	1.67	1.20	1.07	0.99	0.94	0.92	0.92

So as to remove the mentioned under- or overestimation, an additional tuning of ERM was performed. The corresponding results are given in Table 3. In the case of the general piezoelectric problem, the additional tuning consisted in increasing the local approximation orders of the electric field by 2 with respect to their global counterparts. For the direct piezoelectricity, such increasing was performed within both mechanical and electric fields. In the inverse piezoelectricity case, higher-order equilibration was applied to the mechanical field. The corrected (previously unsatisfactory in Table 2) values of effectivity indices are marked in bold. They are closer to the desired value of 1.0 than in the cases without additional tuning.

The improvement of the effectivities for piezoelectricity cases of $p = \pi = 1, 2$ needs an individual approach as in the case of uncoupled elasticity (cf. Zboiński, 2020). For the general and direct piezoelectricity with $p = \pi = 1$ or $p = \pi = 2$, the additional tuning should be performed or skipped, repectively. In the case of the inverse piezoelectricity, for $p = \pi = 1, 2$, the local orders of approximation P and Q of the mechanical field should be increased by 3 or 4 with respect to the global values of p and q (cf. Zboiński, 2013).

Control of the tuning method for the piezoelectricity cases can be easily implemented into adaptive finite element analysis of complex electro-mechanical domains in the block-wise manner, by introduction of the control parameter for each geometrical or functional block of the analysed domain, analogusly to the case of complex mechanical systems (cf. Zboiński, 2010).

7. Conclusions

It was shown how to adapt the equilibrated residual method (ERM) of a posteriori error estimation, invented for elliptic problems, i.e. elasticity and dielectricity, to coupled problems including piezoelectricity ones.

Effective application of ERM to piezoelectric problems needs tuning of the error estimator. Different tuning procedures are necessary for general, direct and inverse piezoelectricity. Three such procedures were proposed, and their effectiveness was numerically demonstrated.

Acknowledgements

Partial support of the National Science Centre, Poland (formerly Polish Scientific Research Committee) under research grant No. N N504 5153 040 is thankfully acknowledged.

References

1. Ainsworth M., 2005, A synthesis of a posteriori error estimation techniques for conforming, nonconforming and discontinuous Galerkin finite element methods. [In:] *Recent Advances in Adaptive Computation. Contemporary Mathematics*, Vol. 383, Z.-C. Shi *et al.* (Edit.), AMS, Providence, 1-14

2. Ainsworth M., Babuska I., 1999, Reliable and robust a posteriori error estimation for singularly perturbed reaction-diffusion problems, *SIAM Journal of Numerical Analysis*, **36**, 331-353

3. Ainsworth M., Demkowicz L., Kim C.W., 2007, Analysis of the equilibrated residual method for a posteriori error estimation on meshes with hanging nodes, *Computer Methods in Applied Mechanics and Engineering*, **196**, 3493-3507

4. Ainsworth M., Oden J.T., 1992, A procedure for a posteriori error estimation for *h-p* finite element methods, *Computer Methods in Applied Mechanics and Engineering*, **101**, 73-96

5. Ainsworth M., Oden J.T., 1993a, A posteriori error estimators for second order elliptic systems: Part 1. Theoretical foundations and a posteriori error analysis, *Computers and Mathematics with Applications*, **25**, 101-113

6. Ainsworth M., Oden J.T., 1993b, A posteriori error estimators for second order elliptic systems: Part 2. An optimal order process for calculating self-equilibrating fluxes, *Computers and Mathematics with Applications*, **26**, 75-87

7. Ainsworth M., Oden J.T., 1993c, A unified approach to a posteriori error estimation using element residual methods, *Numerische Mathematik*, **65**, 23-50

8. Ainsworth M., Oden J.T., Wu W., 1994, A posteriori error estimation for *h-p* approximation in elastostatics, *Applied Numerical Mathematics*, **14**, 23-55

9. Bank R.E., Weiser A., 1985, Some a posteriori error estimators for elliptic partial differential equations, *Mathematics of Computation*, **44**, 283-301

10. Cimatti G., 2004, The piezoelectric continuum, *Annali di Matematica Pura ed Applicata*, **183**, 495-514

11. Demkowicz L., 2007, *Computing with hp-Adaptive Finite Elements. Vol. 1. One- and Two-Dimensional Elliptic and Maxwell Problems*, Chapman & Hall/CRC, Boca Raton, FL

12. Ieşan D., 1990, Reciprocity, uniqueness and minimum principles in the linear theory of piezoelectricity, *International Journal of Engineering Science*, **28**, 1139-1149

13. Kelly D.W., 1984, The self-equilibration of residuals and complementary a posteriori error estimates in the finite element method, *International Journal for Numerical Methods in Engineering*, **20**, 1491-1506

14. Ladeveze P., Leguillon L., 1983, Error estimate procedure in the finite element method and applications, *SIAM Journal on Numerical Analysis*, **20**, 485-509

15. Oden J.T., Cho J.R., 1996, Adaptive *hpq*-finite element methods of hierarchical models for plate- and shell-like structures, *Computer Methods in Applied Mechanics and Engineering*, **136**, 317-345

16. Preumont A., 2006, *Mechatronics. Dynamics of Electromechanical and Piezoelectric Systems*, Springer, Dordrecht

17. Zboiński G., 2010, Adaptive *hpq* finite element methods for the analysis of 3D-based models of complex structures. Part 1. Hierarchical modeling and approximation, *Computer Methods in Applied Mechanics and Engineering*, **199**, 2913-2940

18. Zboiński G, 2013, Adaptive *hpq* finite element methods for the analysis of 3D-based models of complex structures. Part 2. A posteriori error estimation, *Computer Methods in Applied Mechanics and Engineering*, **267**, 531-565

19. Zboiński G., 2016, Problems of hierarchical modeling and *hp*-adaptive finite element analysis in elasticity, dielectricity and piezoelectricity, [In:] *Perusal of the Finite Element Method*. R. Petrova (Ed.), InTech, Rijeka (Croatia), 1-29

20. Zboiński G., 1018, Adaptive modeling and simulation of elastic, dielectric and piezoelectric problems, [In:] *Finite Element Method. Simulation, Numerical Analysis and Solution Techniques*, R. Păcurar (Ed.), InTech, Rijeka (Croatia), 157-192

21. Zboiński G., 2019, 3D-based hierarchical models and *hpq*-approximations for adaptive finite element method of Laplace problems as exemplified by linear dielectricity, *Computers and Mathematics with Applications*, **78**, 2468-2511

22. Zboiński G., 2020, Tuning of the equilibrated residual method for applications in elasticity, dielectricity and piezoelectricity, [In:] *AIP Conference Proceedings*, **2239**, W. Cecot *et al.*, (Edit.), AIP Publishing, 020050-1-18

Manuscript received October 29, 2023; accepted for print December 11, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 231-240, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185163

ANALYSIS OF THE INFLUENCE OF GEOMETRICAL IMPERFECTIONS ON THE EQUIVALENT LOAD STABILIZING ROOF TRUSS WITH LATERAL BRACING SYSTEM[1]

Marcin Krajewski
Gdansk University of Technology, Faculty of Civil and Environmental Engineering, Gdansk, Poland
e-mail: markraje@pg.edu.pl

The paper is focused on the numerical analysis of the stability and load bearing capacity of a flat steel truss. The structure is supported by elastic lateral braces. The translational and rotational brace stiffness are taken into account. The linear buckling analysis is performed for the beam and shell model of the truss. The nonlinear static analysis is conducted for the structure initial geometric imperfections. As a result, the buckling and limit load depending on brace stiffness has been obtained. The reactions in elastic braces are compared to stabilizing forces calculated on the basis of actual code requirements.

Keywords: truss, elastic braces, imperfection, stabilizing load

1. Introduction

Flat trusses are often designed as roof girders in large span steel halls. The characteristic feature for that type of constructions is high stiffness and load bearing capacity but only in their plane. In order to stabilize lattice structures in the roof plane, a bracing system is required (e.g. roof cross and longitudinal braces). In many cases, the braces are designed to carry out the horizontal wind loading but also to stabilize the whole roof.

In many cases, during the design process, it is assumed that the braces are rigid. On this basis, the buckling length of the compressed truss chord (out of plane) is calculated as a distance between lateral supports. The stability of trusses braced by elastic supports was considered by Iwicki (2007, 2010). The experimental tests subjected to gravity loading performed for a steel truss stiffened by elastic side supports were carried out by Jankowska-Sandberg and Kołodziej (2013) and also by Krajewski (2021). The influence of brace rotational stiffness (located at the top chord) on the truss bearing capacity in the case of wind loading was studied by Biegus (2015). In this case, on the basis of analytical solutions, a significant impact of rotational stiffness of side supports on the truss bearing capacity was confirmed. The influence of initial geometric imperfections on the truss bearing capacity subjected to wind loading was also investigated by Krajewski and Iwicki (2016).

Rotational stiffness results from bending stiffness of the brace (e.g. purlin) and connection stiffness between the brace and the truss top chord

$$\frac{1}{k_{rot}} = \frac{1}{K_{roof}} + \frac{1}{K_{con}} \tag{1.1}$$

where: k_{rot} – rotational stiffness of the elastic side support, K_{roof} – bending stiffness of the brace, K_{con} – connection stiffness between the brace and the truss top chord.

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

The vertical load subjected to a truss with an initial imperfection (e.g. arch curvature of the top chord due to Eurocode 3) generates horizontal forces in side braces. Based on the actual code requirements (PN-EN 1993-1-1, 2005), the stabilizing load for roof trusses can be determined. However, it is worth noting that the solutions presented in codes are referred to elements with pinned supports and subjected to constant compression along the length. Based on solutions presented by Czepiżak (2013), the iteration process related to determination of the equivalent stabilizing load (due to code) can be omitted. In this case, the stiffness of the roof cross bracing system has to be calculated. Further analytical modifications of this solution were presented by Krajewski (2021). In that case, horizontal stabilizing forces can be calculated if the brace translational stiffness k [kN/m] is known Eq. (1.2). Analytical solutions focused on the determination of those forces were also presented by Biegus and Czepiżak (2018, 2019) and Czepiżak and Biegus (2016). In that case, the structure with an initial imperfection in the form of arch curvature was considered. Furthermore, the influence of the normal force variation along the compressed truss chord was taken into account

$$q_d = \frac{8\left(e_o + \frac{L_{st}}{k} q_z\right) \sum_{i=1}^{m} N_{ed}}{L^2 - 8\frac{L_{st}}{k} \sum_{i=1} m N_{ed}} \qquad R_d = q_d L_{st} \tag{1.2}$$

where: q_d – equivalent stabilizing load, L_{st} – distance between side braces, e_o – imperfection amplitude, k – elastic brace, q_z – horizontal external loading, N_{ed} – normal force at the truss top chord, L – structure length, m – number of braced elements, R_d – stabilizing force.

The present study is devoted to stability and load bearing capacity analysis of a braced truss. A structure with elastic braces is considered. A constant or parabolic brace stiffness distribution along the truss length is taken into account. The influence of intermediate supports stiffness (translational k [kN/m] and rotational k_{rot} [kNm/deg]) on the buckling load is studied. Also, nonlinear static analysis is performed for the structure with initial imperfections. Moreover, stabilizing forces determined on the basis of code requirements are compared to numerical results.

2. Description of the truss

The analysis has been performed for a truss made of steel (S275). The structure length was equal to 18.0 m and hight was 0.7 m. The top chord consisted of RHS $120 \times 100 \times 4$ profiles and the bottom chord of RHS $100 \times 100 \times 4$. The V type of a diagonal system (RHS $50 \times 50 \times 3$) was considered (Fig. 1). The separate members were joined by means of welded connections. The similar structure stiffened by a trapezoidal sheet and subjected to the wind loading was previously studied by Biegus (2015). The structure was pinned at both ends. The possibility of rotation (twisting out of the truss plane) was blocked at marginal supports. The distance between elastic side supports was equal to 3.0 m (brace numbers from 1 to 5, Fig. 1). In this case, the translational and rotational (out of plane) support stiffness was considered.

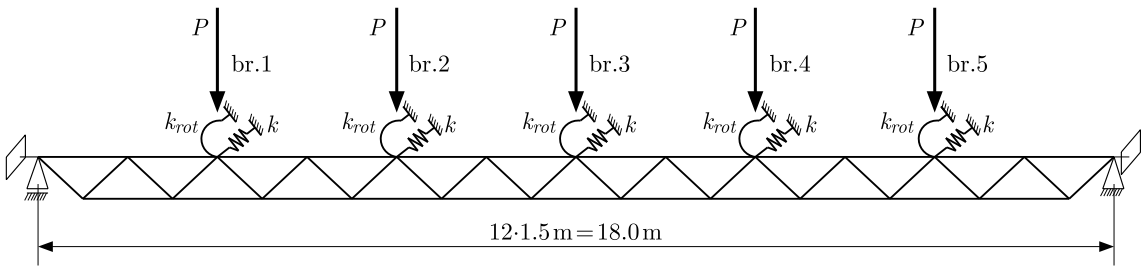


Fig. 1. The static schema

The finite element method was used to conduct analysis for the spatial beam and shell models of the truss. The numerical analysis were performed by means of Femap with NX Nastran (2009). About 42 000 4-node and 3-node shell elements (QUAD4, with six degrees of freedom at node) were used (Fig. 2a). The elements dimensions were up to 2.5 cm×2.5 cm for the top and bottom chord of the truss and about 1.5 cm×1.5 cm for the diagonals. The rigid elements were modeled for welded connections between the diagonals and chords. In this model, structural eccentricities taken from the design project were preserved. About 480 spatial frame elements (with six degrees of freedom at the node) were used to build the beam model of the structure. In this case, the separated members of the truss (diagonals, chords) were joined axially at the nodes. Each element between the nodes was divided to ten parts (Fig. 2b).

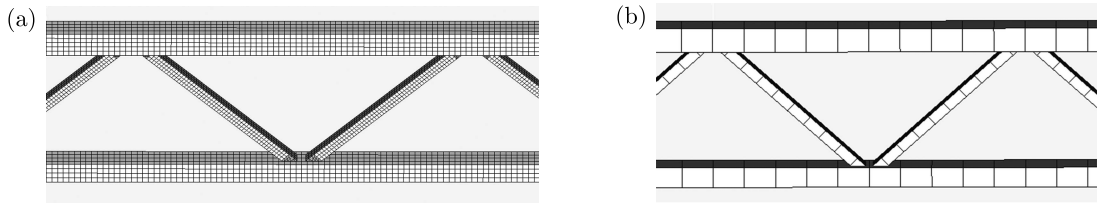


Fig. 2. The numerical models: (a) shell model, (b) beam model

The elastic braces were modeled by means of *Dof-spring* elements with translational and rotational (out of the truss plane) stiffness. In the shell model, these elements were situated at the top shelf of the top chord. In the beam model the elements were connected to the selected nodes located at the center of gravity of the top chord. In both cases (beam and shell models) the truss was loaded by concentrated forces located at the braced joints (gravity loading). The bi-linear elasto-plastic material model was implemented to conduct nonlinear analysis ($E = 210\,\text{GPa}$, $\nu = 0.3$, $f_y = 275\,\text{MPa}$). Numerical research was performed for the structure with a constant stiffness distribution along the top chord length and also for the parabolic stiffness distribution (Fig. 3, Eqs. (2.1) and (2.2))

$$K(x) = k \tag{2.1}$$

where: $K(x)$ – function to describe the distribution of brace stiffness along the truss length, k – brace stiffness (translational), x – distance from the pinned (marginal) support to the elastic support at the top chord (brace)

$$K(x) = \frac{1}{\delta(x)} \tag{2.2}$$

where $\delta(x)$ – function to describe flexibility of the bracing system.

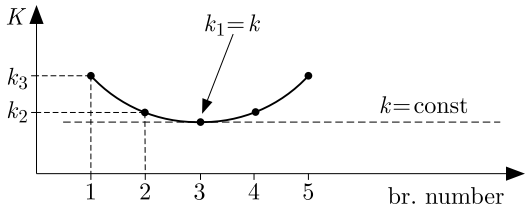


Fig. 3. The distribution of the brace translational stiffness along the truss length

In the second case, the brace stiffness was calculated due to on account of flexibility presented in Eq. (2.3). Based on those calculations the stiffness of braces No. 2 and 4 (k_2) was 12.5% higher, and for braces No. 1 and 5 (k_3) 80% higher in comparison to brace stiffness No. 3 (k_1) located

in the middle of the span (Fig. 1)

$$\delta(x) = \frac{-4\delta_o}{L^2}x^2 + \frac{4\delta_o}{L}x \tag{2.3}$$

where: δ_o – flexibility of the brace system at the middle of the span, L – truss length.

3. Stability and load bearing capacity of the structure

A linear buckling analysis was made for the shell and beam models of the truss. Based on the numerical results one can conclude that the buckling load was dependent on the stiffness of side elastic supports (Fig. 4). In each case, the loading magnitude raised due to an increase in braces translational or rotational stiffness. However, it was observed that there was a threshold (minimum) brace stiffness above which the increase of buckling load was small (less than 10%). For the truss supported by elastic braces of stiffness $k = 650\,\mathrm{kN/m}$ (shell model), the magnitude of buckling load was about 10% lower in comparison to the structure with rigid braces $k = 10^6\,\mathrm{kN/m}$, $k_{rot} = 0\,\mathrm{kNm/deg}$. The same loading magnitude $P_{cr} = 267\,\mathrm{kN}$ was reached for the truss with elastic braces of stiffness equal to $k = 590\,\mathrm{kN/m}$ and $k_{rot} = 5\,\mathrm{kNm/deg}$ or $k = 540\,\mathrm{kN/m}$ and $k_{rot} = 10^6\,\mathrm{kNm/deg}$. In each case (for the beam and shell models of the structure), the above described threshold (translational brace stiffness) decreased as a result of taking into account the rotational stiffness (for the same loading level). A further decrease in the threshold stiffness was noticed if the constant brace stiffness k was replaced by a parabolic stiffness distribution k_1, k_2, k_3.

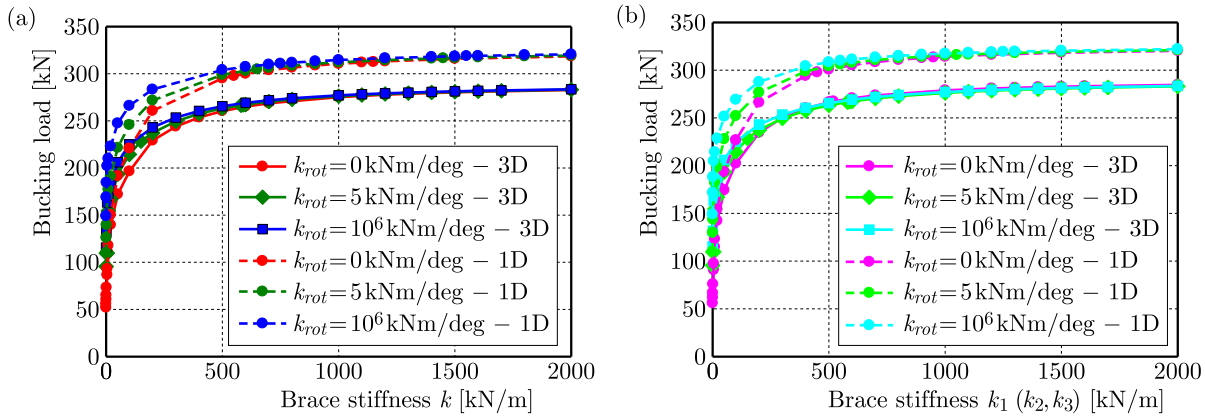


Fig. 4. The relation between the buckling load for the truss (beam and shell models) and brace translational stiffness for: (a) constant brace stiffness k, (b) parabolic brace stiffness distribution k_1, k_2, k_3, with respect to the brace rotational stiffness k_{rot}

In each case, the buckling load raised when the rotational brace stiffness was added. However, this influence decreased due to the rise of brace translational stiffness. The loading magnitudes obtained for the truss supported by braces of constant stiffness k or parabolic stiffness distribution k_1, k_2, k_3 were comparable. In that case, the differences were up to 5%.

The buckling loads obtained for the beam models were up to 15% higher than for the shell models. The reason for these discrepancies were differences in the modeling method described in the previous Section. The buckling modes for selected braced structures are presented on Fig. 5.

A static nonlinear analysis was made for the shell models of the tested structures.

In that case, a bilinear elasto-plastic steel model was implemented. The initial geometric imperfections in the form of arch curvature of the top chord (based on the PN-EN 1993-1-1, 2005 requirements) and an imperfection corresponding to the buckled shape of the truss (with rigid

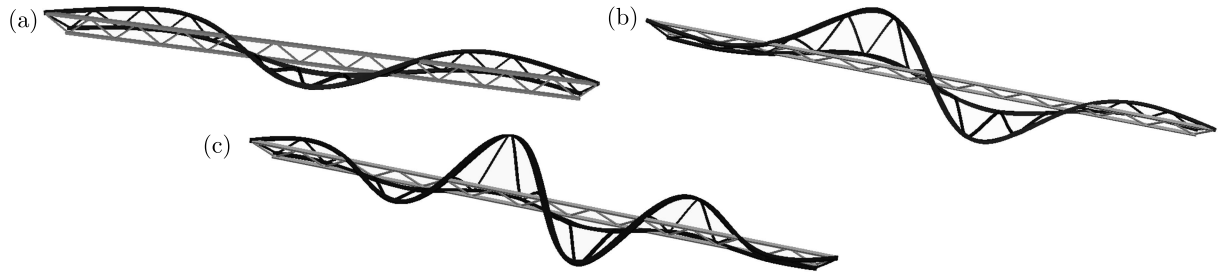


Fig. 5. Buckling modes for the truss (beam and shell models) with braces of stiffness: (a) $k = 100\,\text{kN/m}$ or $k_1 = 100\,\text{kN/m}$, $k_2 = 112.5\,\text{kN/m}$, $k_3 = 180\,\text{kN/m}$ and $k_{rot} = 0\,\text{kNm/deg}$, (b) $k_1 = 100\,\text{kN/m}$, $k_2 = 112.5\,\text{kN/m}$, $k_3 = 180\,\text{kN/m}$ and $k_{rot} = 5\,\text{kNm/deg}$ or $k_{rot} = 10^6\,\text{kNm/deg}$, (c) $k = 10^6\,\text{kN/m}$ or $k = 10^6\,\text{kN/m}$ and $k_{rot} = 10^6\,\text{kNm/deg}$

braces, Fig. 5c) were taken into account. The imperfection amplitude was equal to $L/500$ (the maximum displacements magnitude out of the truss plane). The analysis was carried out for the truss stiffened by elastic (translational) braces with a constant or parabolic stiffness distribution. In each case, the rotational brace stiffness equaled to $k_{rot} = 0\,\text{kNm/deg}$ or $k_{rot} = 5\,\text{kNm/deg}$ or $k_{rot} = 10^6\,\text{kNm/rad}$.

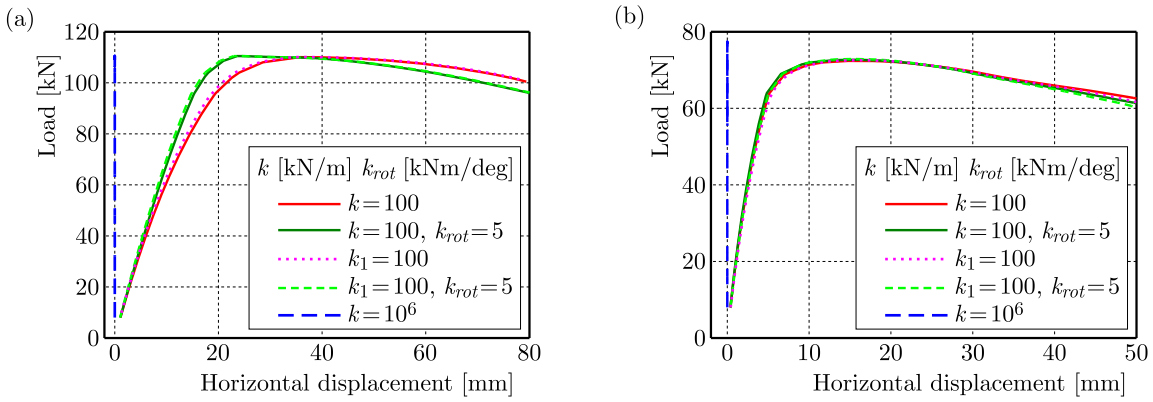


Fig. 6. The relation between truss loading and horizontal displacement of the top chord for the structure with an imperfection in the form of: (a) arch curvature (displacement in the middle of the span), (b) buckling mode (leading displacement at 3.0 m form the middle of the span)

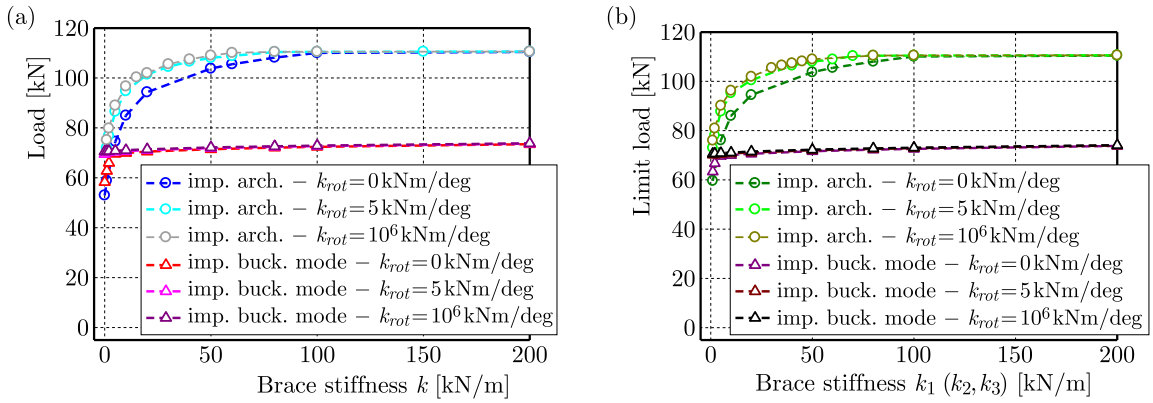


Fig. 7. The relation between the limit load and brace stiffness for the truss with: (a) constant brace stiffness distribution, (b) parabolic brace stiffness distribution, with respect to the shape of the structure initial imperfections

The equilibrium paths obtained from the numerical analysis are presented in Fig. 6. In most cases, the maximum loading magnitude (limit load) depended on the elastic brace stiffness. Similarly to the LBA results, the structure bearing capacity slightly raised (up to 2%) if the parabolic brace stiffness k_1, k_2, k_3 was taken into account in comparison to the truss with constant brace stiffness k. The out of plane truss displacements and limit load significantly depended on the shape of initial geometric imperfections. The loading capacity of the structure with an imperfection in the form of a buckling mode was up to 35% lower in comparison to the truss with the initial arch imperfection. An increase of the brace stiffness above 100 kN/m (k or k_1, k_2, k_3) had no influence on the truss limit load (differences below 1%) for the structure with an initial imperfection shape recommended by code requirements (Fig. 7). However, it is worth noting that for the magnitude of brace stiffness mentioned above, the truss horizontal displacements were much higher (up to 34 mm in the middle of the span, at the limit state) in comparison with the truss with rigid braces (Fig. 7a). The shape of initial geometric imperfection also had a significant impact on the shape of structure deformation at the limit state (Fig. 8). Based on the results, it was observed that the plasticification range (depending on the loading level) occurred at the structural joints and also at the top and bottom chord of the truss.

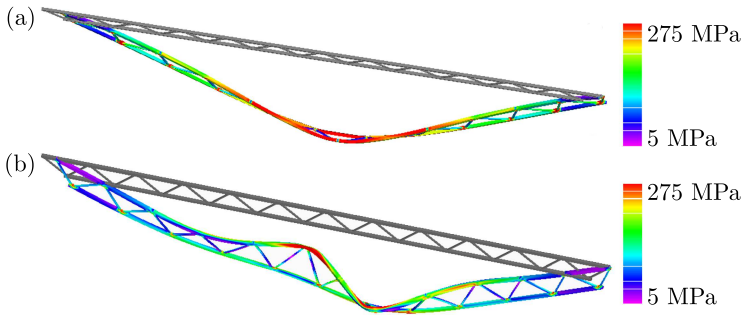


Fig. 8. The deformation and the stress state (HMH hypothesis, top surface of the shell model element) at the limit load for the braced truss $k = 100\,\text{kN/m}$, $k_{rot} = 5\,\text{kNm/deg}$ with an imperfection in the form of: (a) arch curvature, (b) buckling mode

4. Reactions in elastic braces

From the nonlinear analysis, reactions in elastic braces were obtained (Figs. 9-11). For the truss with the arch imperfection, the highest horizontal force was reached for the brace located in the middle of the span. However, the reaction in this brace had the smallest magnitude (absolute value) for the initial imperfection in the form of buckling mode. Based on the analysis results, one can conclude that in many cases at the limit state, the reaction in elastic braces had changeable sign. The reason was not only the brace stiffness but also the shape of the imperfection. In some cases (e.g. $k = 100\,\text{kN/m}$ or $k_1 = 100\,\text{kN/m}$, $k_2 = 112.5\,\text{kN/m}$, $k_3 = 180\,\text{kN/m}$, structure with initial arch imperfection) the braces located near the marginal supports were subjected to compression to a certain loading level. In the next loading step, the global deformation of the structure changed and the considered brace was subjected to tension. The differences in reaction forces in the loaded effort brace with a constant or parabolic stiffness distribution were up to 14%, for the arch imperfection (e.g. $k = 10\,\text{kN/m}$, $k_{rot} = 0\,\text{kNm/deg}$). These differences were up to 13% for the buckling mode imperfection.

The forces in elastic braces obtained from numerical analysis were compared to the stabilizing load calculated on the basis of actual code requirements (PN-EN 1993-1-1, 2005). The equivalent load was determined on the basis of Eq. (1.2). In this case, only the gravity loading was taken into consideration (Fig. 1). The influence of external (horizontal) loading was omitted $q_z = 0\,\text{kN/m}$. The assumed truss loading was equal to 70 kN. That loading level was taken into account on

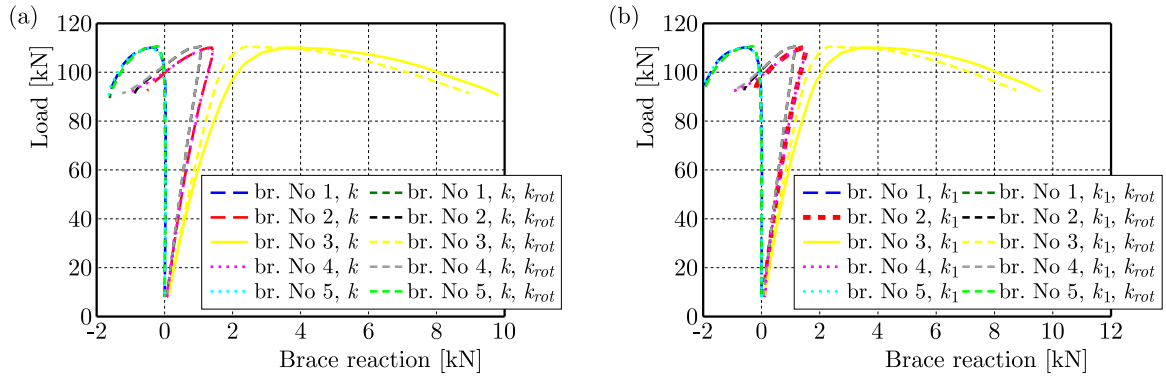


Fig. 9. The relation between the truss loading and reactions in elastic braces (initial imperfection-arc curvature at the top chord): (a) constant brace stiffness distribution $k = 100\,\text{kN/m}$ ($k_{rot} = 0\,\text{kNm/deg}$ or $k_{rot} = 5.0\,\text{kNm/deg}$), (b) parabolic brace stiffness distribution $k_1 = 100\,\text{kN/m}$, $k_2 = 112.5\,\text{kN/m}$, $k_3 = 180\,\text{kN/m}$ ($k_{rot} = 0\,\text{kNm/deg}$ or $k_{rot} = 5.0\,\text{kNm/deg}$)

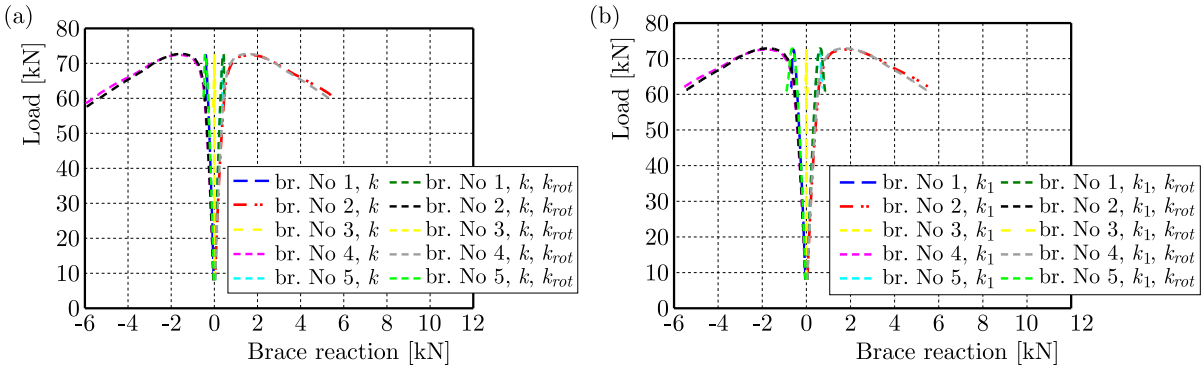


Fig. 10. The relation between the truss loading and reactions in elastic braces (initial imperfection-buckling mode): (a) constant brace stiffness distribution $k = 100\,\text{kN/m}$ ($k_{rot} = 0\,\text{kNm/deg}$ or $k_{rot} = 5.0\,\text{kNm/deg}$), (b) parabolic brace stiffness distribution $k_1 = 100\,\text{kN/m}$, $k_2 = 112.5\,\text{kN/m}$, $k_3 = 180\,\text{kN/m}$ ($k_{rot} = 0\,\text{kNm/deg}$ or $k_{rot} = 5.0\,\text{kNm/deg}$)

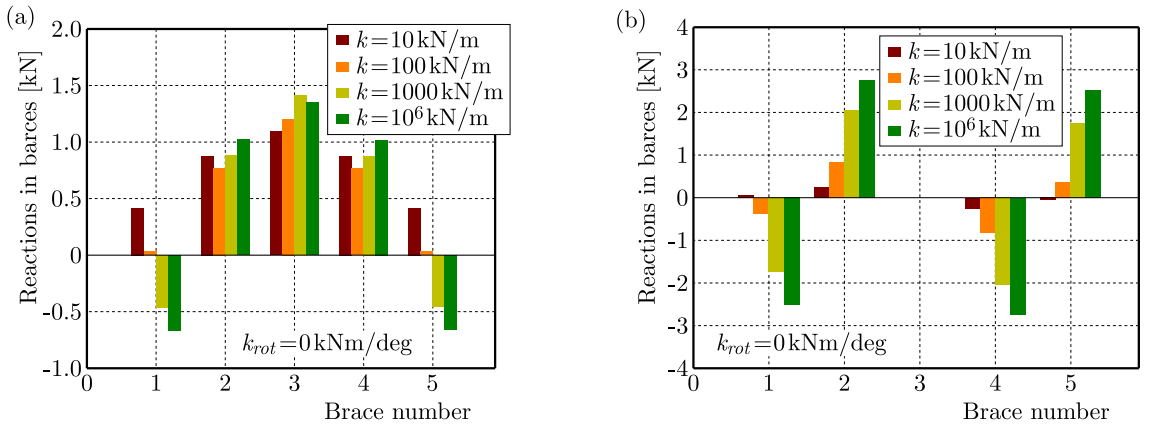


Fig. 11. The reactions in elastic supports versus the brace number for the truss with initial imperfection in the form of: (a) arch curvature of the top chord, (b) buckling mode (for selected stiffness magnitudes)

the basis of nonlinear analysis results (GMNIA) obtained for the structure supported by braces with stiffness $k \geqslant 10\,\text{kN/m}$ or $k_1 \geqslant 10\,\text{kN/m}$ (initial truss imperfection in the form of buckling mode). The maximum magnitude of the normal force N_{ed} in the truss top chord was taken into account (from the middle of the span). The results for the structure with elastic braces with stiffness $k = 100\,\text{kN/m}$ were presented in Fig. 12. It can be observed that the stabilizing force

equal to $R_d = 0.85\,\text{kN}$ (based on Eq. (1.2) and code requirements) was in most cases lower than the reactions in most loaded braces. For the structure with the initial arch imperfection, the differences between R_d and horizontal force in the elastic brace located in the middle of the span were up to 30%. Moreover, it should be point out that the equivalent load determined on the basis of code requirements was constant (linear distribution) in opposite to GMNIA results performed for the structure with two different shapes of initial imperfections. In each case, the reactions in elastic braces (of translational stiffness) decreased due to an increase in rotational brace stiffness.

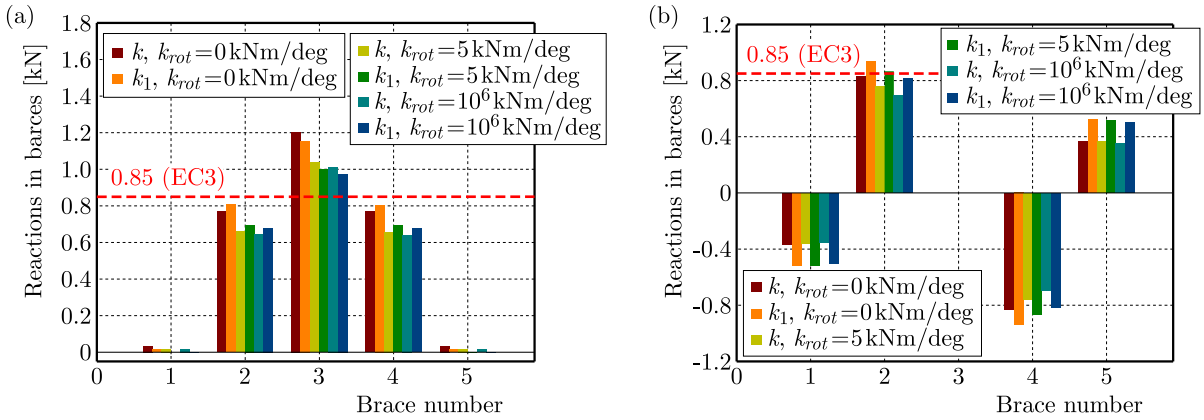


Fig. 12. The reactions in elastic supports versus the brace number for the truss with initial imperfection in the form of: (a) arch curvature of the top chord, (b) buckling mode for $k = 100\,\text{kN/m}$ or $k_1 = 100\,\text{kN/m}$, $k_2 = 112.5\,\text{kN/m}$, $k_3 = 180\,\text{kN/m}$ ($k_{rot} = 0\,\text{kNm/deg}$ or $k_{rot} = 5.0\,\text{kNm/deg}$ or $k_{rot} = 10^6\,\text{kNm/deg}$)

In the present studies, the reactions in braces were obtained from nonlinear analysis made for the imperfect truss model. Based on code recommendations (PN-EN 1993-1-1, 2005), the global imperfections (for the building) and local imperfections (for the element) should be taken into account in the design process. In the research, two different shapes of assumed geometric imperfections were considered. In each case, the maximum imperfection amplitude (horizontal displacements out of the truss plane) was equal to $L/500$ (based on the PN-EN 1993-1-1, 2005, for single element to be braced). However, it is worth noting that the imperfection amplitude can also be assumed on the basis of standards referred to the steel section products. These codes contain detailed information on permissible deviations in terms of shape of the cross-section and element straightness. Moreover, the manufacturing accuracy of the structure production presented in (PN-EN 1090-1, PN-EN 1090-2) can be treated as the guideline for the imperfection modeling. A combination of different shapes of the imperfections which was not considered in the present paper may also influence the structure capacity (Lindgaard *et al.*, 2010). Based on real structure measurements (Šmak and Straka, 2012), the shape and amplitude of the truss initial geometric imperfections had a significant impact on the distribution of loads in the bracing system.

5. Conclusions

Based on the conducted research, the following conclusions can be formulated:

- An increase of elastic brace stiffness above the threshold stiffness did not result in a significant increase of the buckling load or limit load.
- The limit load for the structure with initial imperfection in the form of arch curvature (the shape based on actual code requirements) supported by rigid braces was about 29% higher in comparison to the truss with buckling mode imperfection.

- Based on the GMNIA results, the parabolic brace stiffness distribution (along the truss length) caused that the truss capacity raised only about 2% in comparison to the structure supported by braces with constant stiffness. In that case, the differences between reactions in most loaded elastic supports were up to 14%.
- The stabilizing load determined on the basis of actual code requirements was linear and could be characterized by constant intensity. In that case, the variable normal force distribution along the compressed element length (truss top chord) should be taken into account. In the shell model of the truss, the distribution of reactions in elastic braces was in each case non-uniform. In many cases the reactions had changeable sign due to brace stiffness and shape of the initial geometric imperfection.
- Mostly, the reactions in greatly loaded braces were higher than stabilizing forces given by code requirements. The differences were up to 30%
- An increase of brace rotational stiffness caused a decrease in elastic brace reactions (horizontal forces).

The present research was limited to analysis of a single truss with precisely defined geometry, boundary and loading conditions. However, based on the experimental test results presented in literature (Jankowska-Sandberg and Kołodziej, 2013; Piątkowski, 2021; Krajewski, 2021) performed for different types of flat trusses, the impact of brace stiffness (translational) on the structure capacity was confirmed. In this case, steel trusses characterized by different cross-section shapes (closed or built-up) and geometric dimensions, marginal and intermediate (elastic) supports were taken into consideration. Nevertheless, the real structural connection details between the top chord and the brace (gusset planes, screws etc.) were not taken into account neither in the presented parametric studies nor in the experimental tests mentioned above. The stability analysis of the trusses updated by numerical modeling of connection details and also by experimental tests including real imperfection measurements are the next research step.

Acknowledgement

A part of the research was carried out within the grant 2022/06/X/ST8/00656 National Science Centre, Poland, Miniatura "Influence of the rotational stiffness of the connection between the brace and the chord on the stability of lattice roof trusses subjected to wind loading – experimental tests and numerical analysis".

References

1. Biegus A., 2015, Trapezoidal sheet as a bracing preventing flat trusses from out-of-plane buckling, *Archives of Civil and Mechanical Engineering*, **15**, 3, 735-741
2. Biegus A., Czepiżak D., 2018, Generalized model of imperfection forces for design of transverse roof bracings and purlins, *Archives of Civil and Mechanical Engineering*, **18**, 267-279
3. Biegus A., Czepiżak D., 2019, Equivalent stabilizing force of members parabolically compressed by longitudal variable axial force, *Matec Web of Conferences*, **262**
4. Czepiżak D., 2013, The simplified method for calculation of roof cross braces (in Polish), *Engineering and Construction*, 598-600
5. Czepiżak D., Biegus A., 2016, Refined calculation of lateral bracing system due to global geometrical imperfections, *Journal of Constructional Steel Research*, **119**, 30-38
6. Femap with NX Nastran, Instruction manual, Siemens Product Lifecycle Management Software INC., 2009
7. Iwicki P., 2007, Stability of trusses with linear elastic side-supports, *Thin Walled Structures*, **45**, 849-854

8. Iwicki P., 2010, Sensitivity analysis of critical forces of trusses with side bracing, *Journal of Constructional Steel Research*, **66**, 923-930

9. Jankowska-Sandberg J., Kołodziej J., 2013, Experimental study of steel truss lateral--torsional buckling, *Engineering Structures*, **46**, 165-172

10. Krajewski M., 2021, Stability of trusses with elastic side supports, PhD. thesis, Gdansk University of Technology

11. Krajewski M., Iwicki P., 2016, Stability of an imperfect truss loaded by wind, *Engineering Transactions*, **64**, 4, 509-516

12. Lindgaard E., Lund E., Rasmussen K., 2010, Nonlinear buckling optimization of composite structures considering "worst" shape imperfections, *International Journal of Solids and Structures*, **47**, 3186-3202

13. Piątkowski M., 2021, Experimental research on load of transversal roof bracing due to geometrical imperfections of truss, *Engineering Structures*, **242**, 112558

14. PN-EN 1993-1-1, 2005, Eurocode 3: Design of steel structures – Part 1-1: General rules and rules for buildings

15. PN-EN 1090-1+A1:2012 Execution of steel structures and aluminum structures – Part 1: Requirements for conformity assessment of structural components

16. PN-EN 1090-2, 2018, Construction of steel and aluminum structures. Part 2: Technical requirements regarding steel structures

17. Šmak M., Straka B., 2012, Geometrical and structural imperfections of steel member systems, *Procedia Engineering*, **40**, 434-439

Manuscript received October 30, 2023; accepted for print January 19, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 241-251, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185323

MATERIAL DISCONTINUITY PROBLEMS SOLVED BY A MESHLESS METHOD BASED ON VARIABLY SCALED DISCONTINUOUS RADIAL FUNCTIONS[1]

Artur Krowiak, Jordan Podgórski

Cracow University of Technology, Institute of Computer Science, Cracow, Poland

e-mail: artur.krowiak@pk.edu.pl; jordan.podgorski@pk.edu.pl

The paper presents a meshless method based on global interpolation with radial base functions and shows its application in solving interface problems. Such problems arise when two or more different materials are used to construct the element under consideration. Across the interface between the materials, a discontinuity of material parameters arises. To solve the problem, the radial basis function-based collocation method is applied. To properly reflect the discontinuity, the base functions are modified. In this paper, the method is applied to solve problems described by elliptic equations. Using these examples, the accuracy, stability and convergence of the method are examined.

Keywords: interface problem, meshless method, radial basis functions, discontinuity

1. Introduction

Many problems in science and engineering are defined for elements composed of two or more different materials. Such problems can be encountered, for example, when considering heat transfer in a rod or plate composed of two different materials in contact to form an interface. Across this interface, a discontinuity of material parameters arises, which can lead to non-smooth or discontinuous solutions (LeVeque and Li, 1994).

Low-order approximation methods such as finite element method or finite difference method can generally be used to solve numerically these interface problems because they can easily approximate a discontinuous derivative of an appropriate order across the interface, leading to a non-smooth or discontinuous solution. Moreover, several more sophisticated low order methods have been developed to handle the interface problems. These include the immersed interface method (LeVeque and Li, 1994; Li, 2003) and discontinuous Galerkin immersed finite element method (Yang and Zhang, 2016) to name a few.

Recently, meshless techniques have been strongly developed (Belytschko *et al.*, 1996; Liu, 2003). They, oppositely to the mentioned methods, need only scattered nodes to discretize a problem and therefore they are more versatile than mesh-based methods. Among the meshless techniques, one can also distinguish the methods that can be successfully employed to solve the interface problems (Yoon and Song, 2014; Stevens and Power, 2015; Martin and Fornberg, 2017). They are based on local approximation schemes.

Although all these numerical methods can solve problems with adequate accuracy, they require imposition of many degrees of freedom to ensure this accuracy. On the other hand, there are methods based on global approximation (Fornberg, 1996; Trefethen, 2000; Krowiak, 2008) – high order approximation methods – which are able to solve some problems with high accuracy using only a few nodes. Among the meshless techniques, one can also distinguish such

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

methods. They usually use global approximation with radial basis functions (RBFs) (Ferreira and Fasshauer, 2007; Chen *et al.*, 2014; Krowiak, 2018). It was found that the RBFs have very interesting features: they are very useful in approximation of scattered data and they are very convenient when solving multidimensional problems (Fasshauer, 2007). Unfortunately, the high order approximation methods are not well suited for problems possessing discontinuities. They are generally not able to catch some local features like, for example, discontinuities across interfaces.

Therefore, the goal of the present paper is to modify a meshless method based on a global approximation with the RBFs to accurately solve the interface problems. In this way, we want to preserve all the advantages of this kind of method, extending the possibilities of its use to problems with discontinuities of material parameters across the interfaces. The well-known Kansa (1990) approach is taken under consideration since it possesses all the advantages of the RBFs and is characterized by a simple formulation. To extend the usefulness of the Kansa method for the interface problems, the so-called variably scaled discontinuous kernel interpolation is introduced into the method. This type of interpolation was recently studied by de Marchi *et al.* (2020), showing its high accuracy in the approximation of discontinuous functions.

2. Definition of the interface problem

The examples of one-, two- and three-dimensional physical models composed of different materials that are in contact forming interfaces are presented in Fig. 1. In the figure, Ω^+ and Ω^- denote the areas occupied by materials of different parameters and Γ is the interface.

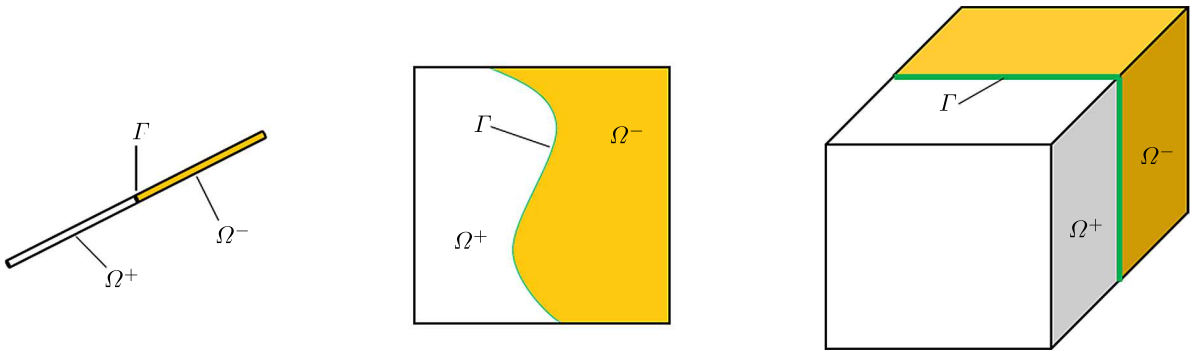


Fig. 1. Examples of physical models with interfaces

Since the method considered in the present paper is used to discretize the spatial variable, we take an elliptic type equation to show an example of mathematical model of the interface problem. The general form of such an equation is as follows

$$\nabla \cdot (\beta(\mathbf{x})\nabla u(\mathbf{x})) + \nabla \cdot (\boldsymbol{\gamma}(\mathbf{x})u(\mathbf{x})) + \kappa(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x}) \tag{2.1}$$

Equation (2.1) with proper boundary conditions can describe boundary or stationary physical problems such as linear elasticity or steady-state heat transfer. In such problems. Eq. (2.1) contains usually the first term called the diffusion one. Then, parameter $\beta(\mathbf{x})$ corresponds to material parameters of the physical model. Therefore, when the interface problem is considered, this parameter has discontinuity across the interface, which can be put as

$$\beta(\mathbf{x}) = \begin{cases} \beta^+(\mathbf{x}) & \text{for} \quad \mathbf{x} \in \Omega^+ \\ \beta^-(\mathbf{x}) & \text{for} \quad \mathbf{x} \in \Omega^- \end{cases} \tag{2.2}$$

Moreover, all the parameters in Eq. (2.1) as well as the right hand side function can be discontinuous across the interface.

To find an unique solution of the interface problem one should enrich the mathematical model given by Eq. (2.1) and proper boundary conditions with other equations that describe the continuity or a jump of the solution as well as the derivative of the solution at the interface. These conditions have the following form for $\mathbf{x} \in \Gamma$

$$u^+(\mathbf{x}) - u^-(\mathbf{x}) = w(\mathbf{x}) \qquad I^+u^+(\mathbf{x}) - I^-u^-(\mathbf{x}) = v(\mathbf{x}) \tag{2.3}$$

In Eq. (2.3) $u^+(\mathbf{x})$, $u^-(\mathbf{x})$ denote the solutions at the interface, when approaching from the Ω^+ or Ω^- subdomain, I^+, I^- are differential operators associated with appropriate subdomains, imposed on the solution at the interface, and $w(\mathbf{x})$, $v(\mathbf{x})$ are functions that define the jump. In mechanical engineering problems, the interface conditions usually have a homogeneous form ($w(\mathbf{x}) = 0$, $v(\mathbf{x}) = 0$) describing the continuity of the solution variable u and the continuity of a flux or stress dependently on the problem considered.

3. Solution method

To solve the problem described in Section 2, a meshless collocation method that uses global interpolation with the RBFs is applied. In the method, the sought solution is interpolated by a series composed of RBFs, which has the following form

$$u_h(\mathbf{x}) = \sum_{j=1}^{N} c_j \phi(\mathbf{x}, \boldsymbol{\xi}_j) \tag{3.1}$$

where N denotes the number of so-called source points $\boldsymbol{\xi}_j$, where the RBFs are centered. In the present paper, the three most often used types of RBFs (Table 1) are considered.

Table 1. Examples of RBFs

RBFs	Function form
Multiquadric	$\phi(\mathbf{x}, \boldsymbol{\xi}_j) = \sqrt{1 + (\varepsilon \lVert \mathbf{x} - \boldsymbol{\xi}_j \rVert_2)^2}$
Invers multiquadric	$\phi(\mathbf{x}, \boldsymbol{\xi}_j) = \dfrac{1}{\sqrt{1 + (\varepsilon \lVert \mathbf{x} - \boldsymbol{\xi}_j \rVert_2)^2}}$
Gaussian	$\phi(\mathbf{x}, \boldsymbol{\xi}_j) = \exp\Big(-(\varepsilon \lVert \mathbf{x} - \boldsymbol{\xi}_j \rVert_2)^2\Big)$

Radial functions presented in Table 1 are infinitely differentiable. They possess a parameter ε that controls their flatness influencing the accuracy and stability of the method. Some interesting issues on approximating as well as solving differential equations using the RBFs are presented by Fasshauer (2007).

Global interpolation with infinitely differentiable RBFs does not allow function (3.1) to accurately reflect the discontinuities across interfaces. Therefore, in the present paper, the RBFs are modified according to the idea presented by de Marchi *et al.* (2020) for kernel functions. Following this idea, the RBFs in the present work are modified by introducing a scaling function $\psi(\mathbf{x})$.

Assuming that the domain of interest belongs to d-dimensional space, i.e. $\mathbf{x} \in \Omega \subseteq \mathbb{R}^d$, then $\Psi : \mathbb{R}^d \to \mathbb{R}$ and the variably scaled RBFs are defined as

$$\phi_\psi(\mathbf{x}, \boldsymbol{\xi}_j) = \phi\Big((\mathbf{x}, \psi(\mathbf{x})), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big) \tag{3.2}$$

where $\phi : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$ and $\phi_\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

In order to introduce a discontinuity to the variably scaled RBFs, the scaling function should be a piecewise continuous one, possessing discontinuity across the interface. It is sufficient to introduce this function as a piecewise constant one, as follows

$$\psi(\mathbf{x}) = \begin{cases} \alpha & \text{for} \quad \mathbf{x} \in \Omega^- \\ \beta & \text{for} \quad \mathbf{x} \in \Omega^+ \end{cases} \tag{3.3}$$

Taking into account the definition of variably scaled RBFs (3.2) and scaling function (3.3), we can conclude that if $\mathbf{x}$ and $\boldsymbol{\xi}_j$ come from the same subdomain one obtains the original RBF, i.e. $\phi(\|\mathbf{x} - \boldsymbol{\xi}_j\|_2)$ but if they are taken from different subdomains, we have $\phi(\|\mathbf{x} - \boldsymbol{\xi}_j\|_2 + (\alpha - \beta)^2)$.

Now, the interpolant written in terms of the modified RBFs approximates the sought solution of the interface problem, what can be put as

$$u_h(\mathbf{x}) = \sum_{j=1}^{N} c_j \phi\left((\mathbf{x}, \psi(\mathbf{x})), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\right) = \sum_{j=1}^{N} c_j \phi_\psi(\mathbf{x}, \boldsymbol{\xi}_j) \tag{3.4}$$

To find the solution, the domain of the problem is discretized with the source points of RBFs $\boldsymbol{\xi}_j$, $j = 1, \ldots, N$ and with the collocation points inside the domain $\mathbf{x}_i^I$, $i = 1, \ldots, N^I$, on the boundaries $\mathbf{x}_i^B$, $i = 1, \ldots, N^B$ and at the interface $\mathbf{x}_i^\Gamma$, $i = 1, \ldots, N^\Gamma$. Then, the function is introduced into the mathematical model of the interface problem. By collocating the equations at appropriate points, the algebraic system of equations is obtained in the following form

$$\begin{aligned}
&\sum_{j=1}^{N} c_j [L\phi_\psi(\mathbf{x}, \boldsymbol{\xi}_j)]_{\mathbf{x}=\mathbf{x}_i^I} = f(\mathbf{x}_i^I) \qquad i = 1, \ldots, N^I \\
&\sum_{j=1}^{N} c_j [B\phi_\psi(\mathbf{x}, \boldsymbol{\xi}_j)]_{\mathbf{x}=\mathbf{x}_i^B} = g(\mathbf{x}_i^B) \qquad i = 1, \ldots, N^B \\
&\sum_{j=1}^{N} c_j \phi_\psi^+(\mathbf{x}_i^\Gamma, \boldsymbol{\xi}_j) - \sum_{j=1}^{N} c_j \phi_\psi^-(\mathbf{x}_i^\Gamma, \boldsymbol{\xi}_j) = w(\mathbf{x}) \qquad i = 1, \ldots, N^\Gamma \\
&\sum_{j=1}^{N} c_j [I^+ \phi_\psi^+(\mathbf{x}, \boldsymbol{\xi}_j)]_{\mathbf{x}=\mathbf{x}_i^\Gamma} - \sum_{j=1}^{N} c_j [I^- \phi_\psi^-(\mathbf{x}, \boldsymbol{\xi}_j)]_{\mathbf{x}=\mathbf{x}_i^\Gamma} = v(\mathbf{x}) \qquad i = 1, \ldots, N^\Gamma
\end{aligned} \tag{3.5}$$

For simplicity of the presentation, the differential operator from the governing equation is denoted by L and the one from boundary conditions is denoted by B. To properly reflect the interface conditions given by the last two equations of system (3.5), the following explanation on how to evaluate function ϕ_ψ^+ and ϕ_ψ^- at $\mathbf{x}_i^\Gamma$ should be made. In this case, taking into account the definition of scaling function (3.3), we have

$$\begin{aligned}
\phi_\psi^+(\mathbf{x}_i^\Gamma, \boldsymbol{\xi}_j) &= \phi\left((\mathbf{x}_i^\Gamma, \alpha), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\right) \\
\phi_\psi^-(\mathbf{x}_i^\Gamma, \boldsymbol{\xi}_j) &= \phi\left((\mathbf{x}_i^\Gamma, \beta), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\right)
\end{aligned} \tag{3.6}$$

Similarly, the derivative of this function at interface nodes is evaluated.

Equations (3.5) can be written in a convenient matrix notation

$$\boldsymbol{\Phi}\mathbf{c} = \mathbf{b} \tag{3.7}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_L, \boldsymbol{\Phi}_B, \boldsymbol{\Phi}_\Gamma^{(1)}, \boldsymbol{\Phi}_\Gamma^{(2)}]^{\mathrm{T}}$, $\mathbf{b} = [\mathbf{f}, \mathbf{g}, \mathbf{w}, \mathbf{v}]^{\mathrm{T}}$, and the vector $\mathbf{c}$ contains all the interpolation coefficients c_j, $j = 1, \ldots, N$.

The elements of the submatrices and subvectors contained in Eq. (3.7) are computed from the formulas for $j = 1, \ldots, N$

$$(\mathbf{\Phi}_L)_{ij} = \Big[L\phi\Big((\mathbf{x}, \psi(\mathbf{x})), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big)\Big]_{\mathbf{x}=\mathbf{x}_i^I} \qquad i = 1, \ldots, N^I$$

$$(\mathbf{\Phi}_B)_{ij} = \Big[B\phi\Big((\mathbf{x}, \psi(\mathbf{x})), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big)\Big]_{\mathbf{x}=\mathbf{x}_i^B} \qquad i = 1, \ldots, N^B$$

$$(\mathbf{\Phi}_\Gamma^{(1)})_{ij} = \phi\Big((\mathbf{x}_i^\Gamma, \alpha), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big) - \phi\Big((\mathbf{x}_i^\Gamma, \beta), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big) \qquad i = 1, \ldots, N^\Gamma$$

$$(\mathbf{\Phi}_\Gamma^{(2)})_{ij} = \Big[I^-\phi\Big((\mathbf{x}, \alpha), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big) - I^+\phi\Big((\mathbf{x}, \beta), (\boldsymbol{\xi}_j, \psi(\boldsymbol{\xi}_j))\Big)\Big]_{\mathbf{x}=\mathbf{x}_i^\Gamma} \qquad i = 1, \ldots, N^\Gamma$$

and from the formulas

$$\mathbf{f}_i = f(\mathbf{x}_i^I) \qquad i = 1, \ldots, N^I$$

$$\mathbf{g}_i = g(\mathbf{x}_i^B) \qquad i = 1, \ldots, N^B$$

$$\mathbf{w}_i = w(\mathbf{x}_i^\Gamma) \qquad \mathbf{v}_i = v(\mathbf{x}_i^\Gamma) \qquad i = 1, \ldots, N^\Gamma$$

In this paper, the discretization which leads to the square system of equations is considered. In this case, the numbers of discretization points have to fulfil the following condition: $N^I + N^B + 2N^\Gamma = N$. Solution of Eq. (3.7) gives the values of the interpolation coefficients, i.e.

$$\mathbf{c} = \mathbf{\Phi}^{-1}\mathbf{b} \tag{3.8}$$

Finally the approximate, analytic solution of the problem, in the form of interpolation function (3.4) is obtained.

Unique solution (3.8) is conditioned by invertibility of $\mathbf{\Phi}$ matrix. Some notes on the invertibility of this type of matrix can be found in the work of Hon and Schaback (2001). From this information, it follows that although the invertibility is not guaranteed, the cases where the matrix is singular are very rare and the method has been successfully applied to many problems in science and engineering (Chen *et al.*, 2014). To guarantee the solvability of the problem, the so-called symmetric collocation approach should be used (Wu, 1992).

The matrices resulting from global interpolation using the RBFs are dense and prone to ill-conditioning. Therefore, some approaches to control this numerical phenomenon with a proper value of the shape parameter of RBFs have to be included during the calculation process. Among the algorithms designed for this purpose, it was found in the present work that a kind of cross-validation approach presented by Rippa (1999) yields good results. This algorithm is used in finding the value of the shape parameter of RBFs, when solving examples in the present paper.

4. Numerical examples and discussion

To examine the method, several examples from mechanical engineering and science have been solved. In the present paper, a few of them are presented. The first example presents the interface problem that can describe one-dimensional elasticity. The governing equation and boundary conditions are as follows

$$\big(E(x)u'(x)\big)' = b(x) \qquad u(0) = 0 \qquad u(10) = 1 \tag{4.1}$$

Due to the use of different materials, the Young modulus is discontinuous across the interface at $x = 5$

$$E(x) = \begin{cases} E^+ & \text{for} \quad x \in [0, 5] \\ E^- & \text{for} \quad x \in (5, 10] \end{cases} \tag{4.2}$$

In the calculations the following values are assumed: $E^+ = 10^4$, $E^- = 10^3$ and the right hand side function describing the body forces is as follows $b(x) = 25x - 7.5x^2 + 0.5x^3$. In the considered problem, the interface conditions define the continuity of displacement and the continuity of stress at the interface point. They have the form

$$u^+(5) = u^-(5) \qquad E^+ u'^+(5) = E^- u'^- \tag{4.3}$$

According to Eq. (4.3), we expect a continuous but non-smooth solution and discontinuous derivative of the solution. The scaling function is assumed as follows

$$\psi(x) = \begin{cases} -1 & \text{for} \quad x \in [0,5) \\ 1 & \text{for} \quad x \in [5,10] \end{cases}$$

In Fig. 2, the results of application of the method with conventional RBFs are shown. To obtain the results, the multiquadric RBFs have been used and the domain has been discretized by $N = 40$ equispaced source points.

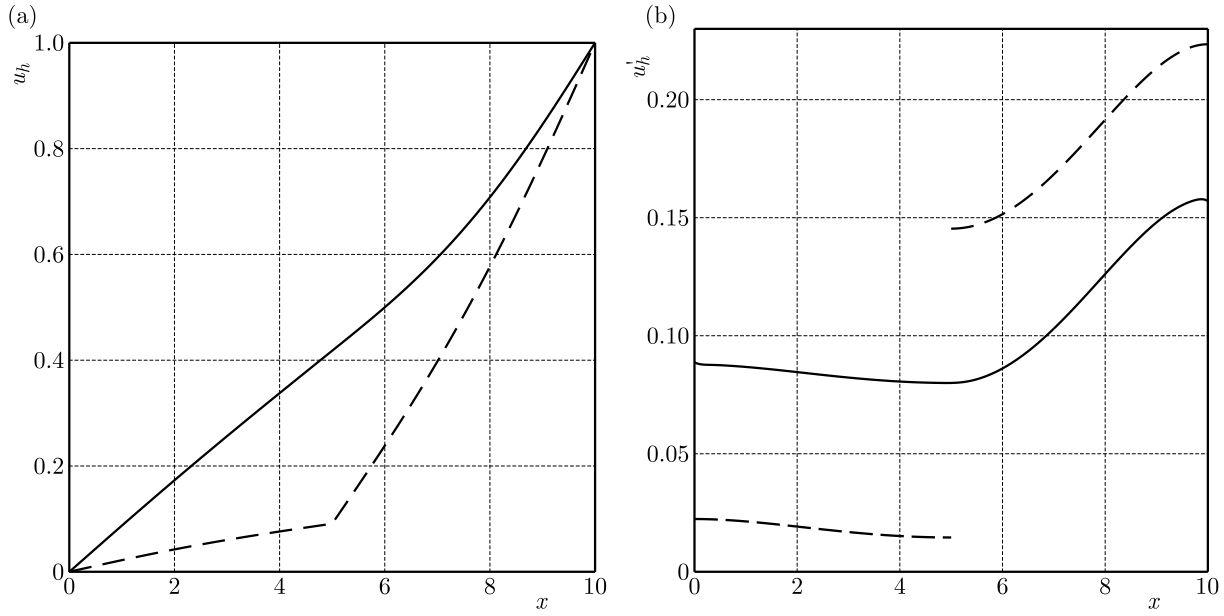


Fig. 2. The solution (a) and derivative of the solution (b) obtained by the classical RBF method; solid line – approximate solution, dash line – exact one

The method that uses classical RBFs does not catch the interface features.

The same problem has been solved by the method with variably scaled discontinuous RBFs using the same discretization parameters. The obtained results are shown in Fig. 3.

The method introduced in Section 3 accurately reflects the non-smooth solution and the discontinuity of its derivative, as can be seen in Fig. 3, where the approximate solution as well as its derivative coincide with the exact one.

In this paper, three types of RBFs listed in Table 1 have been tested for accuracy. For each type of RBFs and each type of point distribution, the shape parameter ε of RBFs has been determined. The results in the form of root-mean-square error between the approximate solution and the exact one are included in Table 2.

The presented results show that the method with variably scaled discontinuous RBFs gives very accurate results using a relatively small number of discretization points.

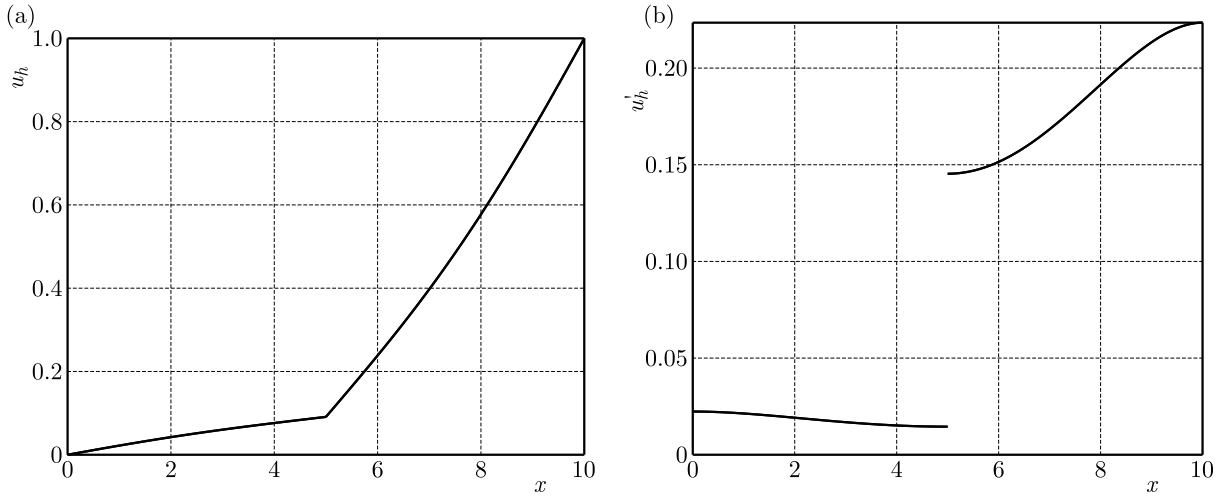


Fig. 3. The solution (a) and derivative of the solution (b) obtained by the modified multiquadric RBFs; solid line – approximate solution, dash line – exact one

Table 2. The accuracy (RMS error) of making use of various types of RBFs

N	RBFs					
	Multiquadric		Invers multiquadric		Gaussian	
	ε	RMS	ε	RMS	ε	RMS
10	0.15	3.0076e-03	0.10	2.3531e-03	0.15	1.5743e-03
20	0.10	7.5026e-06	0.10	1.7451e-05	0.20	3.0356e-06
30	0.35	3.6352e-05	0.25	8.5778e-05	0.35	5.1857e-07
40	0.25	2.5681e-06	0.10	3.3773e-06	0.35	3.5394e-07

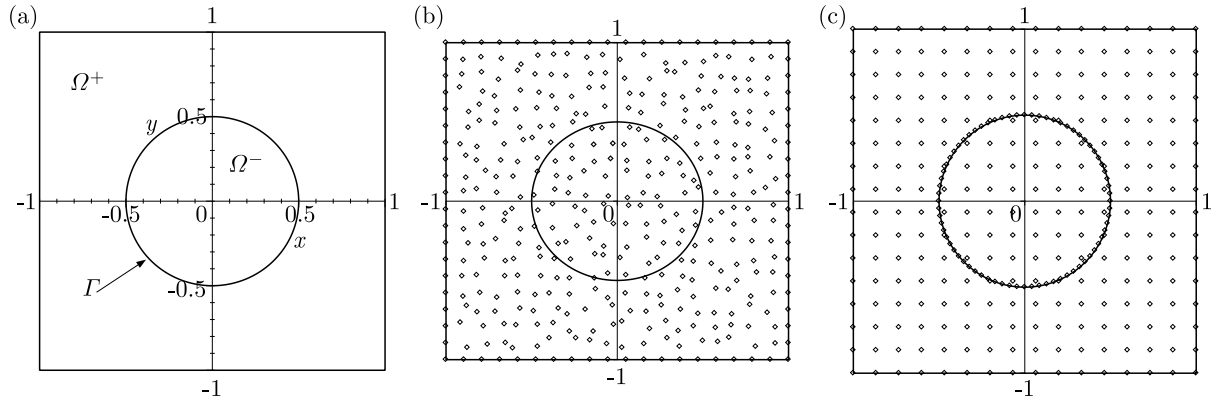


Fig. 4. The domain of the problem (a) with examples of the RBF source point distribution (b) and collocation point distribution (c)

As the second example, a problem modelled by a partial differential equation is considered. It can describe a steady-state heat transfer in a square domain with s circle interface presented in Fig. 4. The mathematical description of the problem is as follows

$$\begin{aligned} &-\nabla\cdot[\beta(\mathbf{x})\,\nabla u(\mathbf{x})] = f(\mathbf{x}) \qquad && \mathbf{x}=(x,y)\in\Omega\subset\mathbb{R}^2 \\ &u(\mathbf{x}) = g(\mathbf{x}) && \mathbf{x}\in\partial\Omega \end{aligned} \tag{4.4}$$

where diffusion parameter $\beta(\mathbf{x})$ has a jump across the interface due to the use of different materials

$$\beta(\mathbf{x}) = \begin{cases} \beta^+ & \text{for} \quad \mathbf{x} \in \Omega^+ \\ \beta^- & \text{for} \quad \mathbf{x} \in \Omega^- \end{cases} \tag{4.5}$$

The interface conditions have the form for $\mathbf{x} \in \Gamma$

$$u^+(\mathbf{x}) = u^-(\mathbf{x}) \qquad \beta^+ \frac{\partial u^+(\mathbf{x})}{\partial \mathbf{n}} = \beta^- \frac{\partial u^-(\mathbf{x})}{\partial \mathbf{n}} \tag{4.6}$$

where $\mathbf{n}$ is the normal outward vector to the interface.

In the calculation, the following values and functions have been assumed in the model

$$\Omega^+ = 10 \qquad \Omega^- = 1 \qquad f(\mathbf{x}) = 0 \qquad g(\mathbf{x}) = \frac{\sqrt{(x^2+y^2)^3}}{\beta^+} - \frac{\beta^+ - \beta^-}{8\beta^+\beta^-}$$

To reflect the discontinuity, the scaling function has been taken as

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{for} \quad \mathbf{x} \in \Omega^+ \\ -1 & \text{for} \quad \mathbf{x} \in \Omega^- \end{cases}$$

The solution obtained using classical RBFs is presented in Fig. 5a along with the exact one (Fig. 5b).

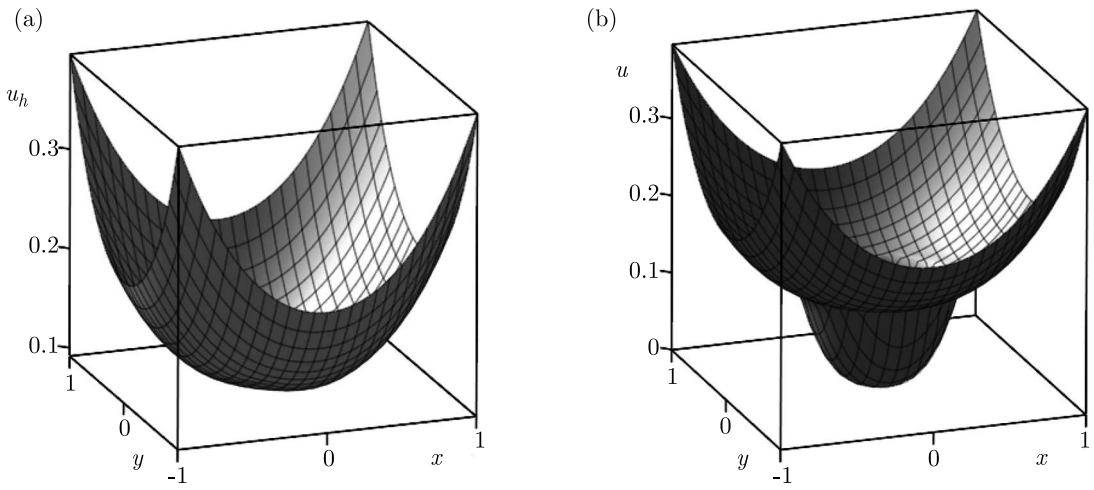


Fig. 5. The solution obtained using classical Gaussian RBFs (a) and the exact one (b)

The solution obtained using variably scaled discontinuous RBFs and the absolute error $\delta = |u_h - u|$ are shown in Fig. 6. To obtain the results, Gaussian RBFs have been applied with following discretization parameters: $N = 400$, $N^I = 196$, $N^B = 60$, $N^\Gamma = 72$. A pseudo-random distribution of the RBF source points (Fig. 4b) and the uniform one for the collocation points (Fig. 4c) have been applied in the calculation.

For a detailed assessment, the profile of the solution $u_h(x, y = 0)$ and the profile of the derivative of the solution $u'_{hx}(x, y = 0)$ along with the exact ones are shown in Fig. 7.

Various types of RBFs have been tested in the present work and several point distributions. In all cases presented in this paper, the RBF source points were distributed pseudorandomly but the collocation points were evenly distributed. The detailed results are shown in Table 3.

The results show that the use of variably scaled discontinuous RBFs allows the method to accurately reflect a non-smooth solution unlike the classical method. All types of RBFs lead to similar accuracy if only the shape parameter is appropriately determined.

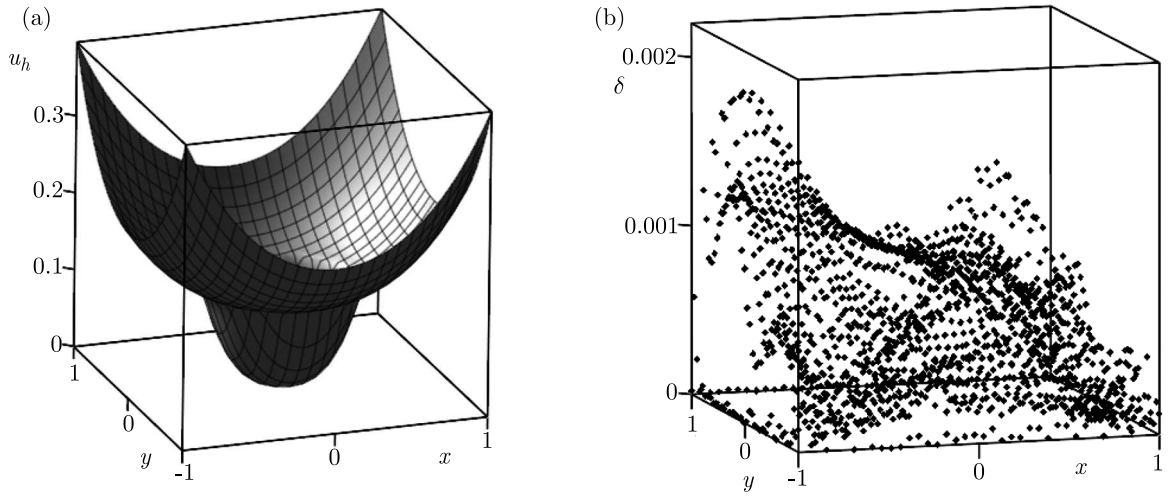


Fig. 6. The solution obtained with the modified Gaussian RBFs (a) and the error of the solution (b)

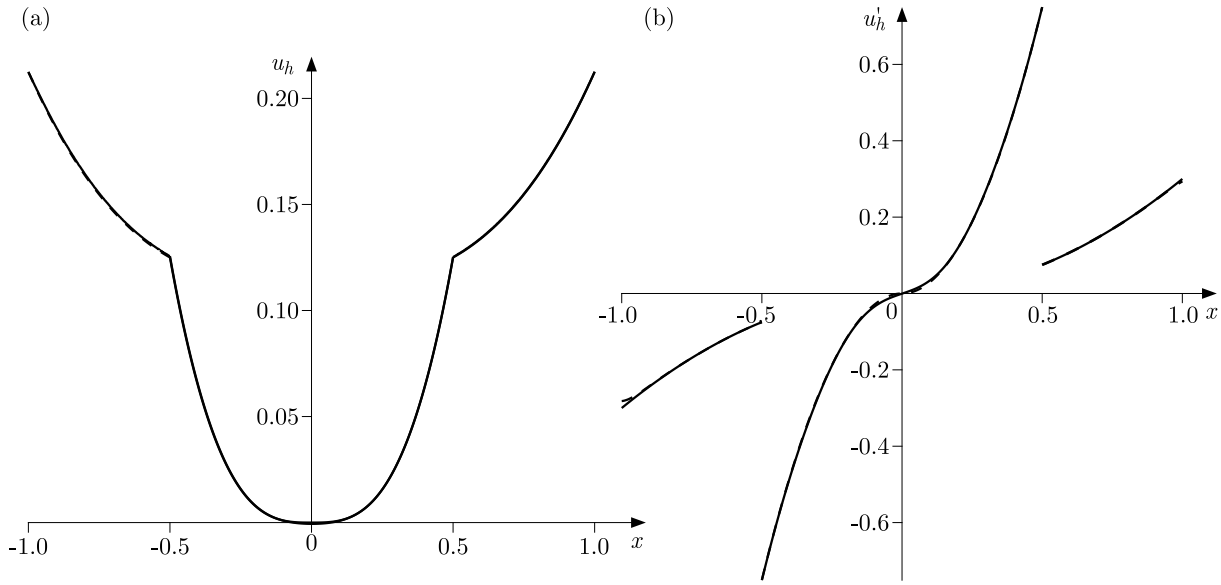


Fig. 7. The profile of the solution (a) and the profile of the derivative (b) obtained with the modified Gaussian RBFs; solid line – approximate solution, dash line – exact one

Table 3. The accuracy (RMS error) of making use of various types of RBFs

N	RBFs					
	Multiquadric		Invers multiquadric		Gaussian	
	ε	RMS	ε	RMS	ε	RMS
289	1.6	3.6022e-02	1.2	2.6867e-01	0.5	3.9298e-01
400	1.2	1.7702e-03	1.0	6.0864e-04	1.8	2.3150e-03
576	1.2	8.5672e-04	1.1	1.8652e-03	2.3	1.9751e-04

5. Conclusion

The conventional RBF collocation method has many advantages: rapid convergence, convenient use in high dimensional problems and simple implementation but it is not able to accurately catch some local features. In the present work, the method is extended to be useful in the interface problems. To this end, the scaling function is introduced into the RBFs. A piecewise

constant scaling function is sufficient to make the RBFs discontinuous across the interface. The use of such RBFs in the conventional collocation procedure of Kansa method allows the method to accurately solve the interface problems while retaining all the other advantages of this method. The results show that various types of RBFs can be used giving similar accuracy, if only appropriate values of the shape parameter are determined.

In the future work, it is planned to find whether the values assumed in the scaling function affect the quality of the results and to find a method of determining these values, which lead to the best accuracy.

References

1. Belytschko T., Krongauz Y., Organ D., Fleming M., Krysl P., 1996, Meshless methods: an overview and recent developments, *Computer Methods in Applied Mechanics and Engineering*, **139**, 3-47
2. Chen W., Fu Z.J., Chen C.S., 2014, *Recent Advances in Radial Basis Function Collocation Methods*, Springer
3. De Marchi S., Erb W., Marchetti F., Perracchione E., Rossini M., 2020, Shape-driven interpolation with discontinuous kernels: error analysis, edge extraction and applications in MPI, *SIAM Journal on Scientific Computing*, **42**, 2, B472-B491
4. Fasshauer G.E., 2007, *Meshfree Approximation Methods with Matlab*, World Scientific, Singapore
5. Ferreira A.J.M, Fasshauer G.E., 2007, Analysis of natural frequencies of composite plates by an RBF-pseudospectral method, *Composite Structures*, **79**, 202-210
6. Fornberg B., 1996, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, Cambridge
7. Hon Y.C., Schaback R., 2001, On nonsymmetric collocation by radial basis functions, *Applied Mathematics and Computation*, **119**, 177-186
8. Kansa E., 1990, Multiquadrics – a scattered data approximation scheme with applications to computational fluid dynamics II: Solutions to parabolic, hyperbolic, and elliptic partial differential equations, *Computers and Mathematics with Applications*, **19**, 147-161
9. Krowiak A., 2008, Methods based on differential quadrature in vibration analysis of plates, *Journal of Theoretical and Applied Mechanics*, **46**, 1, 123-139
10. Krowiak A., 2018, Domain-type RBF collocation methods for biharmonic problems, *International Journal of Computational Method*, **15**, 1850078-1 – 1850078-20
11. LeVeque R.J., Li Z., 1994, The immersed interface method for elliptic equations with discontinuous coefficients and singular sources, *SIAM Journal on Numerical Analysis*, **31**, 4, 1019-1044
12. Li Z., 2003, An overview of the immersed interface method and its applications, *Taiwanese Journal of Mathematics*, **7**, 1, 1-49
13. Liu G.R., 2003, *Mesh-free Methods, Moving Beyond the Finite Element Method*, CRC Press, Boca Raton
14. Martin B., Fornberg B., 2017, Using radial basis function-generated finite differences (RBF-FD) to solve heat transfer equilibrium problems in domains with interfaces, *Engineering Analysis with Boundary Elements*, **79**, 38-48
15. Rippa, S., 1999, An algorithm for selecting a good value for the parameter c in radial basis function interpolation, *Advances in Computational Mathematics*, **11**, 193-210
16. Stevens D., Power H., 2015, The radial basis function finite collocation approach for capturing sharp fronts in time dependent advection problems, *Journal of Computational Physics*, **298**, 423-445

17. Trefethen L.N., 2000, *Spectral Methods in MATLAB*, 3rd. repr. ed., SIAM, Philadelphia

18. Wu Z., 1992, Hermite-Birkhoff interpolation of scattered data by radial basis functions, *Approximation Theory and its Applications*, **8**, 1-10

19. Yang Q., Zhang X., 2016, Discontinuous Galerkin immersed finite element methods for parabolic interface problems, *Journal of Computational and Applied Mathematics*, **299**, 127-139

20. Yoon Y.-C., Song J.-H., 2014, Extended particle difference method for weak and strong discontinuity problems: Part I. Derivation of the extended particle derivative approximation for the representation of weak and strong discontinuities, *Computational Mechanics*, **53**, 1087-1103

Manuscript received October 10, 2023; accepted for print December 14, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 253-267, Warsaw 2024
https://doi.org/10.15632/jtam-pl/183833

PARAMETRIC DYNAMIC ANALYSIS OF TENSEGRITY CABLE-STRUT DOMES[1]

Paulina Obara, Maryna Solovei, Justyna Tomasik
Kielce University of Technology, Faculty of Civil Engineering and Architecture, Kielce, Poland
correspondence author Maryna Solovei, e-mail: msolovei@tu.kielce.pl

The paper contains a parametric dynamic analysis of cable-strut domes. The special structures named tensegrity are considered. Two qualitative different tensegrity domes, i.e., the Geiger dome and the Levy dome are taken into account. The aim of the study is to compare the dynamic behaviour of such structures. The first stage of analysis involves the identification of initial prestress forces (system of internal forces, which holds structural components in stable equilibrium) and infinitesimal mechanisms. The second stage focuses on calculating natural frequencies, while in the last, the impact of time-independent external loads on vibrations is studied. The influence of initial prestress and external load on the dynamic response of the structures is considered. A geometrically non-linear model is used to analysis. Presented considerations are crucial for the next step in the analysis, i.e., dynamic stability analysis of the behaviour of tensegrity structures under periodic loads.

Keywords: tensegrity dome, initial prestress, infinitesimal mechanisms, natural vibrations, free vibrations

1. Introduction

The tensegrity steel domes are special cable-strut trusses. These structures are characterized by a system of internal forces, which holds structural components in stable equilibrium (a self-balanced system of internal forces, self-stress state, initial prestress). Additionally, some of these structures are also characterized by the presence of infinitesimal mechanisms, which are stabilized by a self-balanced system of internal forces. In such cases, a modification of the initial prestress allows for controlling static and dynamic parameters of the structure. Low material demand, lightness of the system, and resistance to various types of loads are the main advantages of these structures. The most common are two tensegrity domes, i.e., the Geiger dome (Geiger, 1988) and the Levy dome (Levy, 1989). In the years 1990-2022, according to Google Scholar, the appearance of the Geiger dome in different articles counts more than 10 000 and more than 18 000 of Levy's dome. Both structures consist of load-bearing systems represented by flat or spatial girders connected with additional longitudinal cables. The approach of tensegrity dome is used for long-span roofs (Levy, 1994; Oribasi *et al.*, 2002) and covers (Geiger *et al.*, 1986; Levy *et al.*, 2013).

Practical application of tensegrity domes requires a thorough examination of static and dynamic properties, as well as overall behaviour of the structure. Most of the research to date focuses on layout design (Rebielak, 2000; Yuan *et al.*, 2007), form-finding methods (Lee *et al.*, 2009), or shape optimization (Kawaguchi *et al.*, 1999; Zhang and Feng, 2017). A smaller number of studies focused on static parameters of tensegrity structures (Shen *et al.*, 2021; Sun *et al.*, 2021; Obara *et al.*, 2023a), and only a few on the dynamic behaviour of dome systems (Obara, 2019; Kim and Sin, 2014; Atig *et al.*, 2017). Due to a non-conventional shape, the complete dynamic

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

analysis of tensegrity domes can be challenging. Some papers contain an initial dynamic analysis of the Geiger dome (Kim and Sin, 2014; Qin *et al.*, 2023).

The analysis of the literature shows that the vast majority of works concerns tensegrity design, the search for stable forms, optimization algorithms, methods of controlling the shape of tensegrity structures under the influence of external loads, and discusses the use of these structures. Against this background, parametric analysis evaluating the influence of the initial prestress on dynamic properties of tensegrity structures is the subject of few studies. In addition, these works relate to specific solutions. The first attempt to fill this gap was the paper (Obara and Solovei, 2023). The influence of initial prestress on the natural frequency for Geiger domes was determined. Two cases of configurations (with a closed and open upper section) were considered. Additionally, two variants of the nature of a dome (regular and modified) were taken into account to compare the dynamic behaviour of domes. In turn, this paper contains a comparison of the dynamic behaviour of the two most popular, qualitatively different, types of tensegrity domes i.e., the Levy and Geiger domes. First, the system of internal forces, which holds structural components in stable equilibrium (initial prestress), and infinitesimal mechanisms are identified. Next, the influence of initial prestress on frequencies is determined. The consideration includes natural and additionally free vibrations. The impact of time-independent external loads on the vibrations is analyzed. The load is treated as an initial disturbance of the equilibrium state, i.e., as imposition of the initial conditions, hence the frequencies are called free. To evaluate this behaviour, a geometrically non-linear model is used. The presented parametric considerations, among others, lead to an answer to the question of how the initial prestress and external load influence the dynamic response of the structure. Additionally, they are crucial for the dynamic stability analysis of the behaviour of tensegrity structures under periodic loads, which will be the subject of the next considerations.

2. Material and methods

Tensegrity domes are described n-element ($e = 1, 2, \ldots, n$) spatial cable-strut trusses with m degrees of freedom $\mathbf{q} \in \mathbb{R}^{m\times 1}$. The elasticity of elements e are described by the elasticity matrix $\mathbf{E} \in \mathbb{R}^{n\times n}$

$$\mathbf{E} = \operatorname{diag}\left[\frac{E_1A_1}{L_1} \quad \frac{E_2A_2}{L_2} \quad \ldots \quad \frac{E_nA_n}{L_n}\right] \tag{2.1}$$

where E_e is Young's modulus, A_e is across-sectional area and L_e is length of the element. In turn, geometry is described by the compatibility matrix $\mathbf{B} \in \mathbb{R}^{n\times m}$, which can be determined using formalism of the finite element method (Zienkiewicz and Taylor, 2000). Additionally, in contrast to traditional steel domes, tensegrity domes are characterized by a self-balanced system of internal forces (self-stress state). The first step of the analysis of tensegrity structures relies on the identification of the self-stress state. The most frequently used methods are the force density method (Zhang and Ohsaki, 2006), dynamic relaxation (Bel Hadj *et al.*, 2010), energy optimization (Li *et al.*, 2011), reduced coordinates method (Arsenault and Gosselin, 2005), iteration method (Ma *et al.*, 2018), genetic algorithm (Obara *et al.*, 2023b), singular value decomposition of the force density and equilibrium matrices (Tran and Lee, 2013) or of the compatibility matrix (Gilewski *et al.*, 2016). Since the self-stress state does not depend on geometrical and mechanical characteristics and on an external load, one of the simplest methods to identify it is spectral analysis of the matrix $\mathbf{BB}^{\mathrm{T}} \in \mathbb{R}^{n\times n}$. The self-stress state is considered as an eigenvector $\mathbf{y}_S \in \mathbb{R}^{n\times 1}$ related to the zero eigenvalue of the matrix $\mathbf{BB}^{\mathrm{T}}$ (Obara, 2019; Obara and Solovei, 2023). The self-equilibrium system of longitudinal forces $\mathbf{S} \in \mathbb{R}^{n\times 1}$ depends on the eigenvector $\mathbf{y}_S$ and on the initial prestress level S

$$\mathbf{S} = \mathbf{y}_S S \tag{2.2}$$

The range of initial prestress level S is a property of the structure and depends on its characteristics and external load. The minimum prestress level S_{min} is related to the appropriate distribution of normal forces in the elements of the structure. The external load can cause a different distribution of normal forces, and it can be corrected by the introduction of a proper initial prestress level. In turn, the maximum prestress level S_{max} is related to the load-bearing capacity of the most stressed elements.

The aim of the paper is to assess the impact of prestress level on the dynamic behaviour of tensegrity domes under time-independent external loads $\mathbf{P} \in \mathbb{R}^{m\times 1}$. The most interesting for all are tensegrity domes characterized by the occurrence of infinitesimal mechanisms. In the absence of the initial prestress forces such systems are unstable, i.e., geometrically variable. The stabilization occurs only after the introduction of initial prestress. It should be noted, the mechanism is an eigenvector $\mathbf{x}_S \in \mathbb{R}^{m\times 1}$ related to the zero eigenvalue of the matrix $\mathbf{B}^{\mathrm{T}}\mathbf{B} \in \mathbb{R}^{m\times m}$. The modification of the initial prestress level S allows for control, among others, dynamic parameters of the structure. In the paper, natural vibrations and free vibrations (taking into account the impact of the load, which is treated as the initial disturbance of the equilibrium state, i.e., as imposition of the initial conditions) are considered. The frequencies of vibrations are determined using the modal analysis

$$[\mathbf{K}_L + \mathbf{K}_G - (2\pi f)^2\mathbf{M}]\widetilde{\mathbf{q}} = \mathbf{0} \tag{2.3}$$

where $\mathbf{K}_L = \mathbf{B}^{\mathrm{T}}\mathbf{E}\mathbf{B} \in \mathbb{R}^{m\times m}$ is a linear stiffness matrix, $\mathbf{M} \in \mathbb{R}^{m\times m}$ is a consequent mass matrix, f is the natural ($f_i(0)$) or free ($f_i(P)$) frequency of vibrations, $\widetilde{\mathbf{q}}$ is an amplitude vector and $\mathbf{K}_G \in \mathbb{R}^{m\times m}$ is a geometry stiffness matrix.

In the case of natural vibrations, the geometry stiffness matrix depends only on the self-equilibrium system of longitudinal forces $\mathbf{S}$ (2.2), consequently $\mathbf{K}_G = \mathbf{K}_G(\mathbf{S})$. For tensegrity domes characterized by infinitesimal mechanisms, the omission of the influence of prestress ($\mathbf{S} = \mathbf{0}$) in Eq. (2.3) leads to zero natural frequencies. The number of them is equal to the number of the infinitesimal mechanisms, and the forms of vibrations correspond to the forms of mechanisms. In the case of free vibrations, the geometry stiffness matrix depends additionally on the longitudinal forces $\mathbf{N} \in \mathbb{R}^{n\times 1}$ caused by the external load

$$\mathbf{K}_G = \mathbf{K}_G(\mathbf{S}) + \mathbf{K}_{GN}(\mathbf{N}) \tag{2.4}$$

Another specific property of tensegrity systems is the size of displacements, which can be large even with small deformations. Due to this, to calculate the axial forces, a geometrically non-linear model is used, assuming the hypothesis of large displacements. The non-linear theory of elasticity in terms of the Total Lagrangian (TL) was adopted as the basis for formulating tensegrity lattice equations

$$[\mathbf{K}_L + \mathbf{K}_G + \mathbf{K}_{NL}(\mathbf{q})]\mathbf{q} = \mathbf{P} \tag{2.5}$$

where $\mathbf{K}_{NL}(\mathbf{q}) \in \mathbb{R}^{m\times m}$ is a non-linear displacement stiffness matrix. The explicit forms of the matrices mentioned above can be found, for example, in (Obara, 2019).

3. Results and discussion

The paper presents dynamic parametric analyzes of two of the most well-known tensegrity domes, i.e., the Levy dome and Geiger dome. The domes consist of uniformly distributed systems of load-bearing girders. Comparing the geometry of both domes, significant differences can be noticed. In the case of the Levy dome, the load-bearing girders are spatial (Figs. 1a,b) whereas in the case of the Geiger dome – flat (Figs. 1c,d). The load-bearing girders consist of tensioned cables

(elements: 1-6) and compressed struts (elements: S_1, S_2, S_3), which are connected by additional circumferential cables (elements: C_1-C_6). The node coordinates of the load-bearing girders are presented in Table 1 – diameter of 12 m and height of 3.25 m of all domes were adopted. The domes are supported in every external node of the lower section.

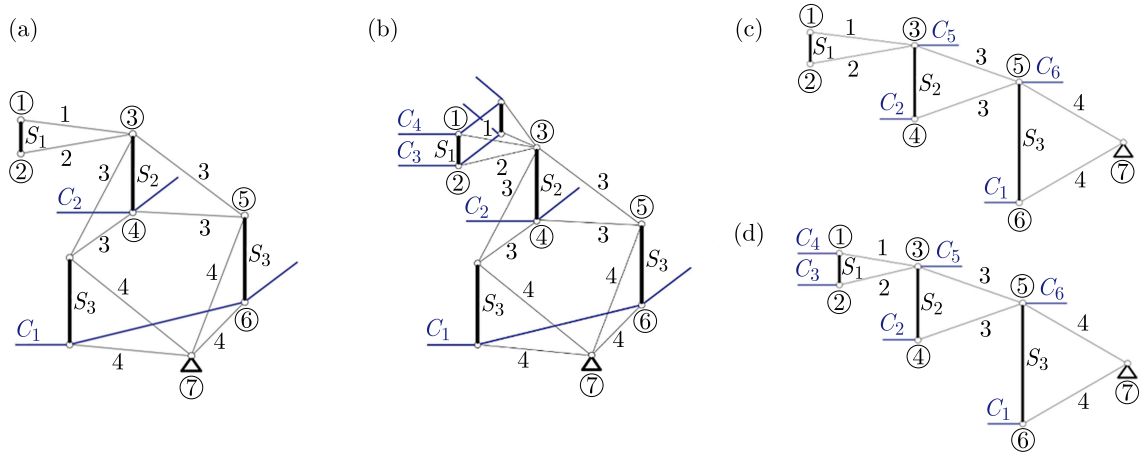


Fig. 1. Load-bearing girders of: (a) Levy type A, (b) Levy type B, (c) Geiger type A, (d) Geiger type B

Table 1. Node coordinates [m] of the load-bearing girders

No. nodes	Type of girder	1	2	3	4	5	6	7
x	A	0.0	0.0	2.0	2.0	4.0	4.0	6.0
	B	0.5	0.5					
z	A and B	2.1	1.5	1.85	0.45	1.15	−1.15	0.0

It should be noted, see the literature, there are two design solutions in the case of Geiger dome – regular (Yuan *et al.*, 2007; Kim and Sin, 2014; Qin *et al.*, 2023) and modified (Atig *et al.*, 2017). The comparison of both of them, in the natural frequency range, was the subject of our previous studies (Obara and Solovei, 2023). In this paper, due to being more similar to the Levy dome, a modified solution was chosen. The considerations contain two cases of configurations, i.e., type A – with a closed upper section (Figs. 1a and 1c) and type B – with an open upper section (Figs. 1b and 1d) and a different number of load-bearing girders i.e., 6 (Figs. 2a, 2c, 3a, 3c), 8, 10 and 12 (Figs. 2b, 2d, 3b, 3d). The names of analyzed domes are acronyms: G – Geiger dome, L – Levy dome, the number – the number of load-bearing girders and letter A or B – girders type e.g., "L 6A" is the Levy dome with 6 girders type A.

In order to compare the behaviour of both domes, the same maximum prestress level $S_{max} = 50\,\text{kN}$ was adopted (due to the maximum effort of the cables of the Geiger dome $W_{max} = 0.93$). In turn, the minimum prestress level S_{min} is an individual characteristic for every dome. Wherein, in the case of natural vibrations, the minimum prestress value is assumed as $S_{min} = 0\,\text{kN}$.

3.1. Identification of self-stress states and infinitesimal mechanisms

The first step in the analysis of the Levy and Geiger domes is the identification of immanent features of tensegrity structures, such as infinitesimal mechanisms and self-stress states. The results of this analysis are shown in Table 2. The domes differ in the number of these features. The Levy dome type A are characterized by zero mechanisms and type B by 1 mechanism,

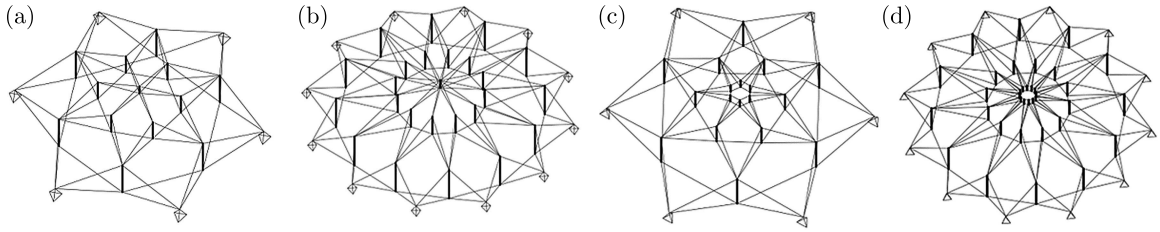


Fig. 2. Levy domes: (a) L 6A, (b) L 12A, (c) L 6B, (d) L 12B

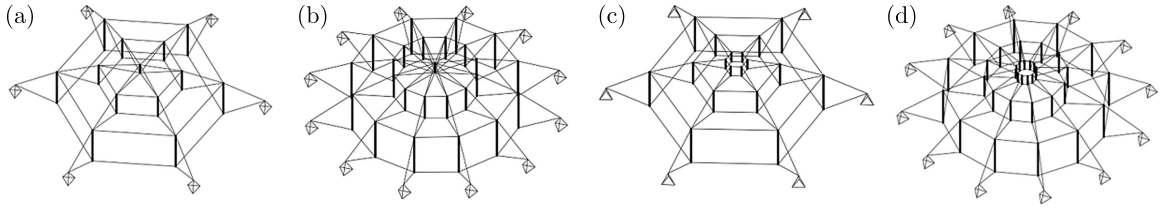


Fig. 3. Geiger domes: (a) G 6A, (b) G 12A, (c) G 6B, (d) G 12B

Table 2. Immanent features of tensegrity domes

Type and No. girders		¢No. nodes	Levy dome			Geiger dome		
			No. elements	No. mechanisms	No. self-stress states	No. elements	No. Mechanisms	No. self-stress states
A	6	32	85	0	7	73	8	3
	8	42	113	0	11	97	8	3
	10	52	141	0	15	121	8	3
	12	62	169	0	19	145	8	3
B	6	42	114	1	7	90	21	3
	8	56	152	1	9	120	27	3
	10	70	190	1	11	150	33	3
	12	84	228	1	13	180	39	3

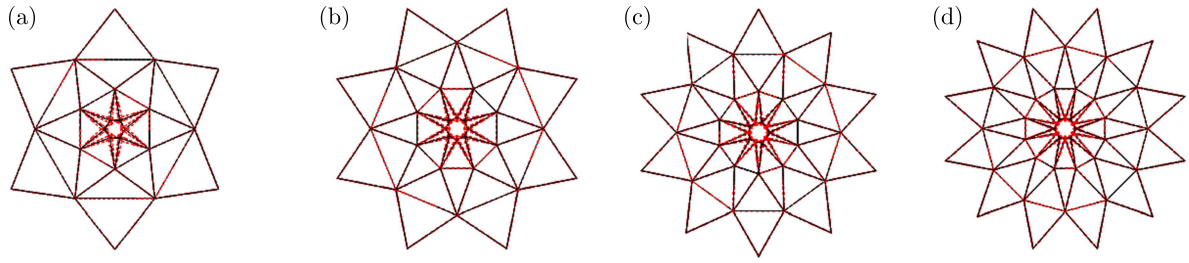


Fig. 4. Infinitesimal mechanism of domes: (a) L 6B, (b) L 8B, (c) L 10B, (d) L 12B

regardless of the number of girders. It should be noted that the mechanism is related only to the upper section (Fig. 4). In turn, in the case of Geiger dome, the number of mechanisms increases with the number of girders. The type of mechanisms differs in the case of the Levy dome – it is related to the entire structure (Fig. 5). In turn, in the case of self-stress states, their number does not depend on the number of girders in the case of Geiger dome (always equals three). In the case of the Levy dome, the number of self-stress states increases with the number of girders. Unfortunately, none of the identified self-stress states identify the type of elements properly. A superposition is needed. The superposed values of self-stress states $\mathbf{y}_S$ for the Levy and Geiger domes are presented in Tables 3 and 4, respectively.

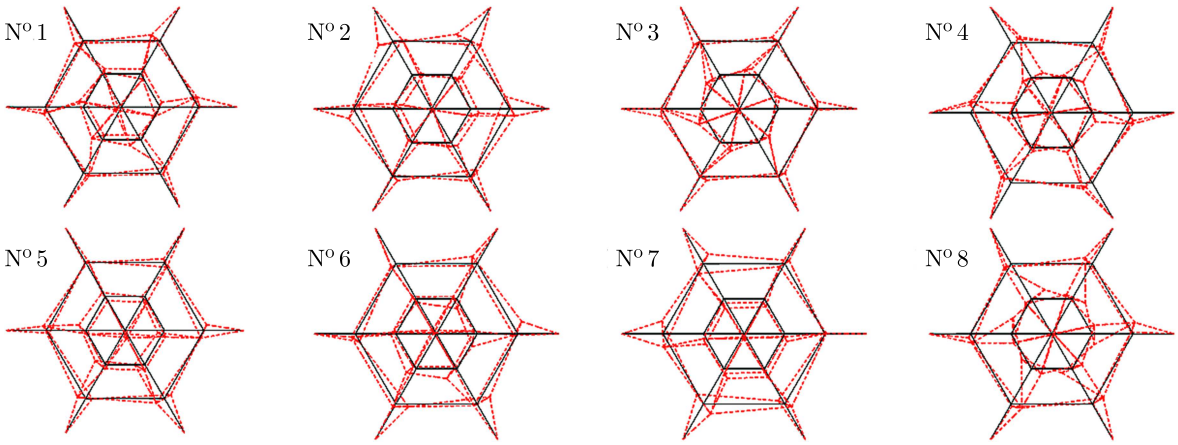


Fig. 5. Infinitesimal mechanisms of dome G 6A

Table 3. Values of self-stress state $\mathbf{y}_S$ of Levy domes

Type A						Type B					
el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$
S_1	¢$-0.147^{(6)}$ $-0.308^{(8)}$ $-0.465^{(10)}$ $-0.616^{(12)}$	1	¢$0.197^{(6)}$ $0.311^{(8)}$ $0.375^{(10)}$ $0.414^{(12)}$	C_1	¢$1.040^{(6)}$ $1.753^{(8)}$ $2.401^{(10)}$ $3.016^{(12)}$	S_1	¢$-0.031^{(6)}$ $-0.050^{(8)}$ $-0.061^{(10)}$ $-0.068^{(12)}$	1	¢$0.100^{(6)}$ $0.157^{(8)}$ $0.189^{(10)}$ $0.209^{(12)}$	C_1	¢$1.040^{(6)}$ $1.753^{(8)}$ $2.401^{(10)}$ $3.016^{(12)}$
S_2	¢$-0.161^{(6)}$ $-0.218^{(8)}$ $-0.248^{(10)}$ $-0.264^{(12)}$	2	¢$0.142^{(6)}$ $0.224^{(8)}$ $0.270^{(10)}$ $0.298^{(12)}$	C_2	¢$0.336^{(6)}$ $0.691^{(8)}$ $1.032^{(10)}$ $1.359^{(12)}$	S_2	¢$-0.161^{(6)}$ $-0.218^{(8)}$ $-0.248^{(10)}$ $-0.264^{(12)}$	2	¢$0.073^{(6)}$ $0.114^{(8)}$ $0.137^{(10)}$ $0.151^{(12)}$	C_2	¢$0.336^{(6)}$ $0.691^{(8)}$ $1.032^{(10)}$ $1.359^{(12)}$
S_3	-1.000	3	¢$0.295^{(6)}$ $0.372^{(8)}$ $0.406^{(10)}$ $0.424^{(12)}$	C_3		S_3	-1.000	3	¢$0.295^{(6)}$ $0.372^{(8)}$ $0.406^{(10)}$ $0.424^{(12)}$	C_3	¢$0.109^{(6)}$ $0.252^{(8)}$ $0.396^{(10)}$ $0.534^{(12)}$
		4	¢$1.491^{(6)}$ $1.303^{(8)}$ $1.204^{(10)}$ $1.147^{(12)}$	C_4				4	¢$1.491^{(6)}$ $1.303^{(8)}$ $1.204^{(10)}$ $1.147^{(12)}$	C_4	¢$0.154^{(6)}$ $0.353^{(8)}$ $0.554^{(10)}$ $0.748^{(12)}$

$^{(6)}$ dome with 6 girders; $^{(8)}$ dome with 8 girders;
$^{(10)}$ dome with 10 girders; $^{(12)}$ dome with 12 girders

3.2. Influence of the number of girders on natural frequencies

After the identification of self-stress states and infinitesimal mechanisms, the influence of initial prestress level S on the natural frequencies is considered. Particularly, the impact of the number of girders on the frequencies is analyzed. It is assumed that the cables are made of steel S460N. Type A cables with Young's modulus 210 GPa (EN 1993-1-11: 2006) are used. The cable diameter and load-bearing capacity are 20 mm and 110.2 kN, respectively. The struts are made of hot-finished circular hollow sections (steel S355J2) with Young's modulus 210 GPa. The diameter and thickness of struts are 76.1 mm and 2.9 mm, respectively. The struts were divided into three groups according to length and load-bearing capacity. Group 1 lengths are 0.6 m and $N_{Rd} = 224.3$ kN, group 2 are 1.4 m and 170.5 kN, and group 3 are 2.3 m and 107.1 kN. The density of steel is equal to $\rho = 7860\,\mathrm{kg/m^3}$. The calculations were made using quasi-linear and non-linear models implemented in the Mathematica environment.

Table 4. Values of self-stress state $\mathbf{y}_S$ of Geiger domes

Type A						Type B					
el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$	el.	$\mathbf{y}_S$
S_1	¢$-0.228^{(6)}$ $-0.304^{(8)}$ $-0.380^{(10)}$ $-0.455^{(12)}$	1	0.306	C_1	¢$1.739^{(6)}$ $2.272^{(8)}$ $2.814^{(10)}$ $3.360^{(12)}$	S_1	-0.051	1	0.308	C_1	¢$1.739^{(6)}$ $2.272^{(8)}$ $2.814^{(10)}$ $3.360^{(12)}$
S_2	-0.265	2	0.220	C_2	¢$0.756^{(6)}$ $0.988^{(8)}$ $1.223^{(10)}$ $1.461^{(12)}$	S_2	-0.265	2	0.223	C_2	¢$0.756^{(6)}$ $0.988^{(8)}$ $1.223^{(10)}$ $1.461^{(12)}$
S_3	-1.000	3	0.801	C_3		S_3	-1.000	3	0.801	C_3	¢$0.217^{(6)}$ $0.283^{(8)}$ $0.351^{(10)}$ $0.419^{(12)}$
		4	2.006	C_4				4	2.006	C_4	¢$0.303^{(6)}$ $0.396^{(8)}$ $0.491^{(10)}$ $0.586^{(12)}$
				C_5	¢$0.236^{(6)}$ $0.308^{(8)}$ $0.381^{(10)}$ $0.455^{(12)}$					C_5	¢$0.236^{(6)}$ $0.308^{(8)}$ $0.381^{(10)}$ $0.455^{(12)}$
				C_6	¢$0.227^{(6)}$ $0.297^{(8)}$ $0.368^{(10)}$ $0.439^{(12)}$					C_6	¢$0.227^{(6)}$ $0.297^{(8)}$ $0.368^{(10)}$ $0.439^{(12)}$

$^{(6)}$ dome with 6 girders; $^{(8)}$ dome with 8 girders;
$^{(10)}$ dome with 10 girders; $^{(12)}$ dome with 12 girders

The first part of the assessment concerns the influence of initial prestress level S on natural frequencies $f_i(0)$. The dynamic behaviour of the dome is highly dependent on the type of load--bearing girder and on the number of identified infinitesimal mechanisms. In Fig. 6, the first and last frequencies corresponding to the infinitesimal mechanisms are presented. The zero level of initial prestress leads to zero natural frequencies, however, they increase with an initial prestress level. The range of changes mainly depends on the kind of dome, which means, on the number of mechanisms. In the case of the dome with one mechanism, i.e., Levy domes type B (Fig. 6a), the first frequency for S_{max} is 17.62 Hz (L 6B), 30.49 Hz (L 8B), 42.28 Hz (L 10B) and 54.2 Hz (L 12B), which means that with the number of girds the frequency increases (comparing with L 6B) by 73%, 140% and 208%, respectively. In turn, for domes with eight mechanisms, i.e., Geiger domes type A (Fig. 6b), the influence of the number of girders on frequencies is significantly smaller. For example, the value of eighth frequency f_8 for S_{max} varies within the range of 12.3 Hz (G 6A) to 13.7 Hz (G 12A), this means an increase of up 11%. In the case of the Geiger domes type B (Fig. 6b) with a different number of infinitesimal mechanisms, the influence of the number of girders depends on the frequency. The first natural frequency for all domes is almost the same – $f_1(S_{max}) = 5.1$ Hz to 5.6 Hz but the last frequency, which corresponds to the mechanism, for S_{max} is 29.79 Hz (G 6B), 35.94 Hz (G 8B), 48.31 Hz (G 10B) and 57.83 Hz (G 12B), which means that with growing number of girds the frequency increases (comparing with G 6B) by 21%, 62% and 94%, respectively. Additionally, as can be seen, for Geiger domes type B, the higher frequencies are more sensitive to a change in prestressing.

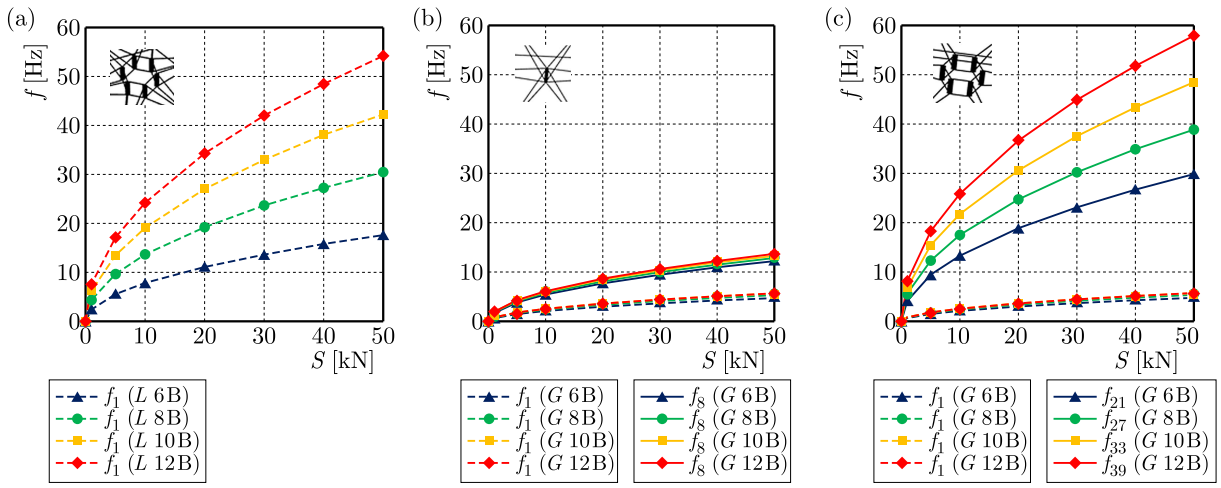


Fig. 6. Natural frequencies corresponding to the infinitesimal mechanisms: (a) Levy domes type B, (b) Geiger domes type A, (c) Geiger domes type B

It is well known that the number of natural frequencies, depending on prestressing, is equal to the number of infinitesimal mechanisms f_{nm}, but in the case of Levy dome type B and Geiger domes of type A, it is different. These structures are characterized by additional natural frequencies f_{add}. The number of them, and the sensitivity to the initial prestress changes, depends on the number of load-bearing girders $No.g$

$$f_{add} = \begin{cases} No.g - 4 & \text{for Levy domes} \\ No.g - 3 & \text{for Geiger domes} \end{cases} \tag{3.1}$$

For comparison, the sensitivity to the prestress changes, in Figs. 7 and 8, the last frequency corresponding to the infinitesimal mechanism, the additional depending on prestress and the first independent of prestress are shown. In the absence of prestress, the additional frequencies, unlike the frequency corresponding to the mechanism, are not zero and the character of the dependence on prestress relies on the types of domes. In the case of Levy domes (Fig. 7), the additional frequencies are more sensitive to a change in prestressing than in the case of Geiger domes (Fig. 8). Additionally, the nature of changes is different. First, for Geiger domes, the additional frequencies are directly proportional to the initial prestress level, whereas in the case of Levy domes, they are not. Secondly, the influence of the number of girders on the frequency is more significant in Levy domes. Thirdly, for Geiger domes, regardless of the number of girders, the value of the frequency independent of prestress is much higher than the frequencies dependent on prestress.

The behaviour of the Levy domes type A is completely different, compared to the Levy dome type B and Geiger domes of type A and B. In these domes, the mechanism was not identified. In Fig. 9, the influence of the initial prestress S on the first, second, and third frequency is shown. The dependencies are linear and almost constant, especially for domes with a small number of girders. In the case of the first frequency, with a growth of the prestress from S_{min} to S_{max}, the frequency increased only by 6.4% (L 6A), 6% (L 8A), 7% (L 10A), and 8.3% (L 12A), whereas in the case of the third frequency – 0.4% (L 6A), 3.5% (L 8A), 5.1% (L 10A), and 6.4% (L 12A). Comparing all results, we can say that due to the lack of mechanisms in the case of Levy dome type A, the natural frequencies are practically not affected by the initial prestress, independent of the number of load-bearing girders.

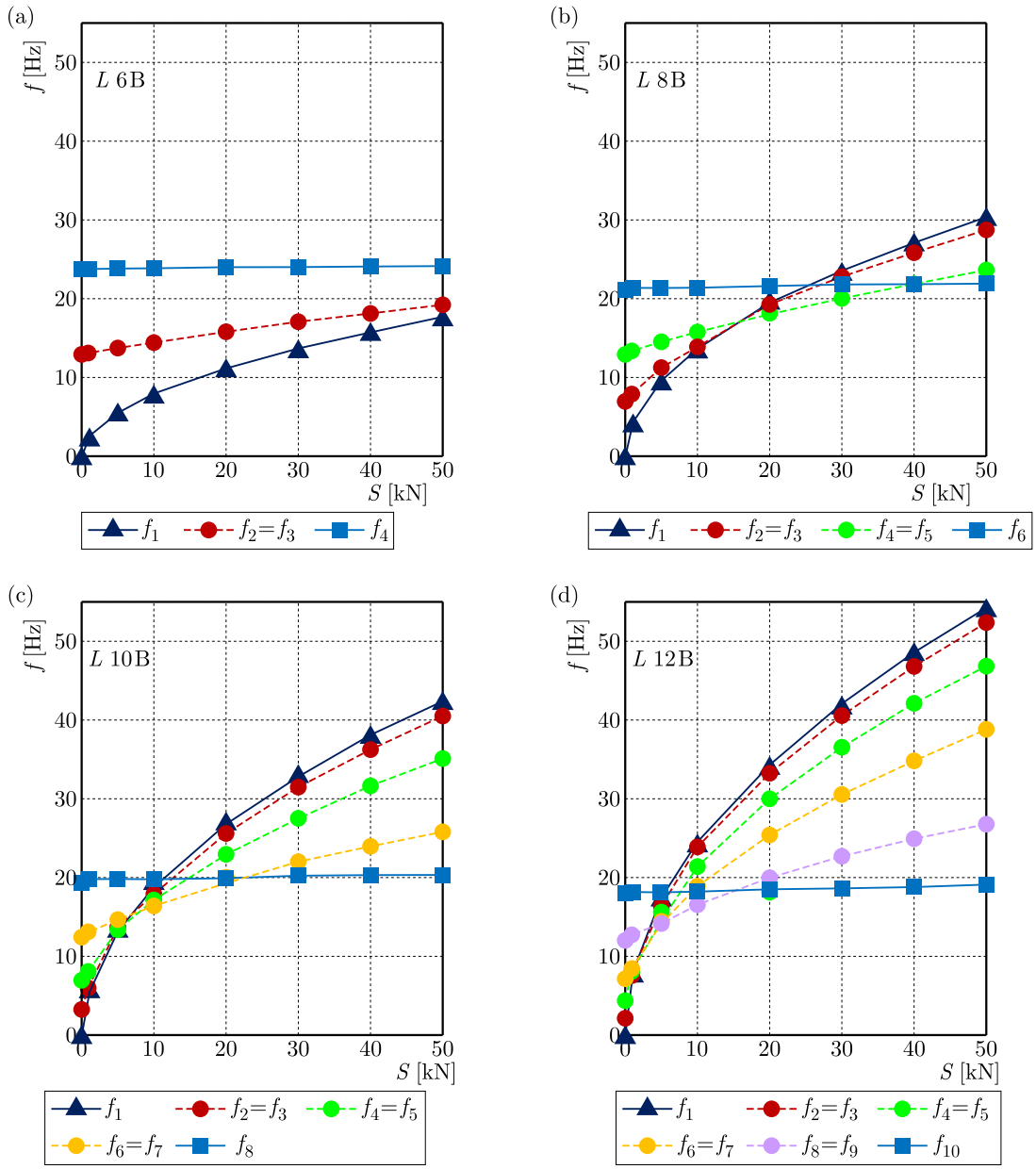


Fig. 7. Influence of the initial prestress S on the natural frequency for Levy domes type B: (a) L 6B, (b) L 8B, (c) L 10B, (d) L 12B

3.3. Influence of the initial prestress level on free frequencies

Next, the influence of initial prestress level S on the free $f_i(P)$ frequencies of the domes is calculated. The time-independent concentrated force applied vertically (gravity load) to one top node is considered. Two variants of loads i.e., $P = 1\,\text{kN}$ and $P = 5\,\text{kN}$ are taken into account. To compare the response to external disturbance, the load is applied in three different positions. The first position is a node of the upper section of the dome, the second position corresponds with a node on the hoop of the second section, and the third one – is with the third section, respectively. It means, according to Fig. 1, that the load is applied to the 1st, 3rd, and 5th node. It should be noted that taking into account the external load, the initial conditions change, and the influence of initial prestress decreases. The load causes additional stress in the system and it is necessary to determine the minimum prestress level S_{min}. S_{min} must ensure the appropriate identification of the element type and provide the positive definite matrix. This is calculated

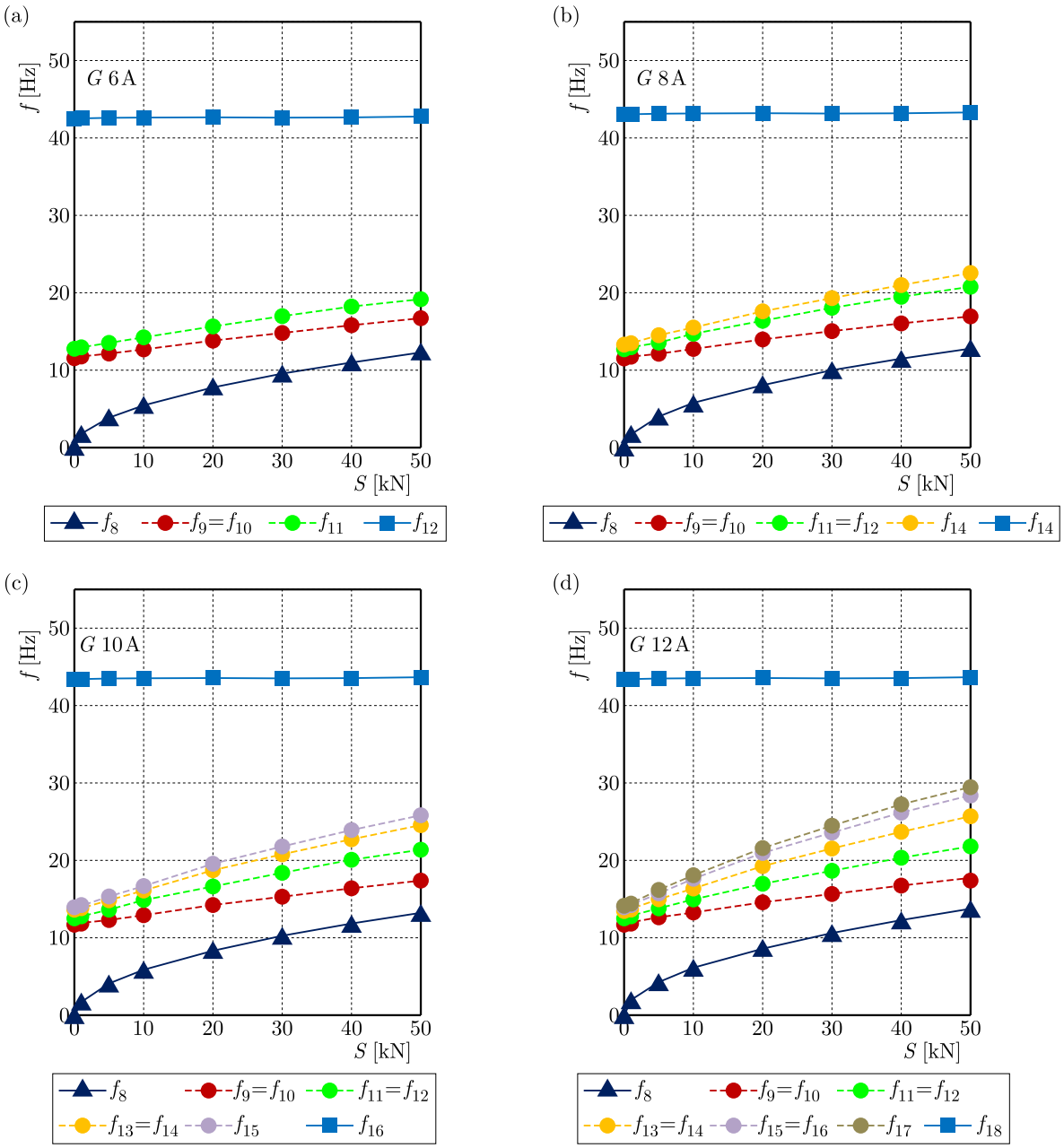


Fig. 8. Influence of the initial prestress S on the natural frequency for Geiger domes type A: (a) G 6A, (b) G 8A, (c) G 10A, (d) G 12A

individually for each dome for each variant of the load. As an example, the results of the analysis of domes with 6 load-bearing girders are shown (Tables 5-8).

In the case of Levy domes type A, the first natural $f_1(0)$ and free $f_1(P)$ frequencies are shown in Table 5. Due to the lack of mechanisms, the free frequencies, like natural ones, are practically not affected by the initial prestress, and are independent of the number of load-bearing girders. The natural frequencies are the same as natural ones. The only thing that changes is the lowest level of initial prestress. It depends not only value of the load but also on the number of loaded nodes. The second position (3rd node) corresponding with a node on the hoop of the second section is the worst, and S_{min} level is only 42 kN. The first position (1st node) corresponds with a node of the upper section of the dome, and S_{min} is equal to 18 kN. The third position (5th node) corresponds with a node of third section, and S_{min} is equal to 5 kN.

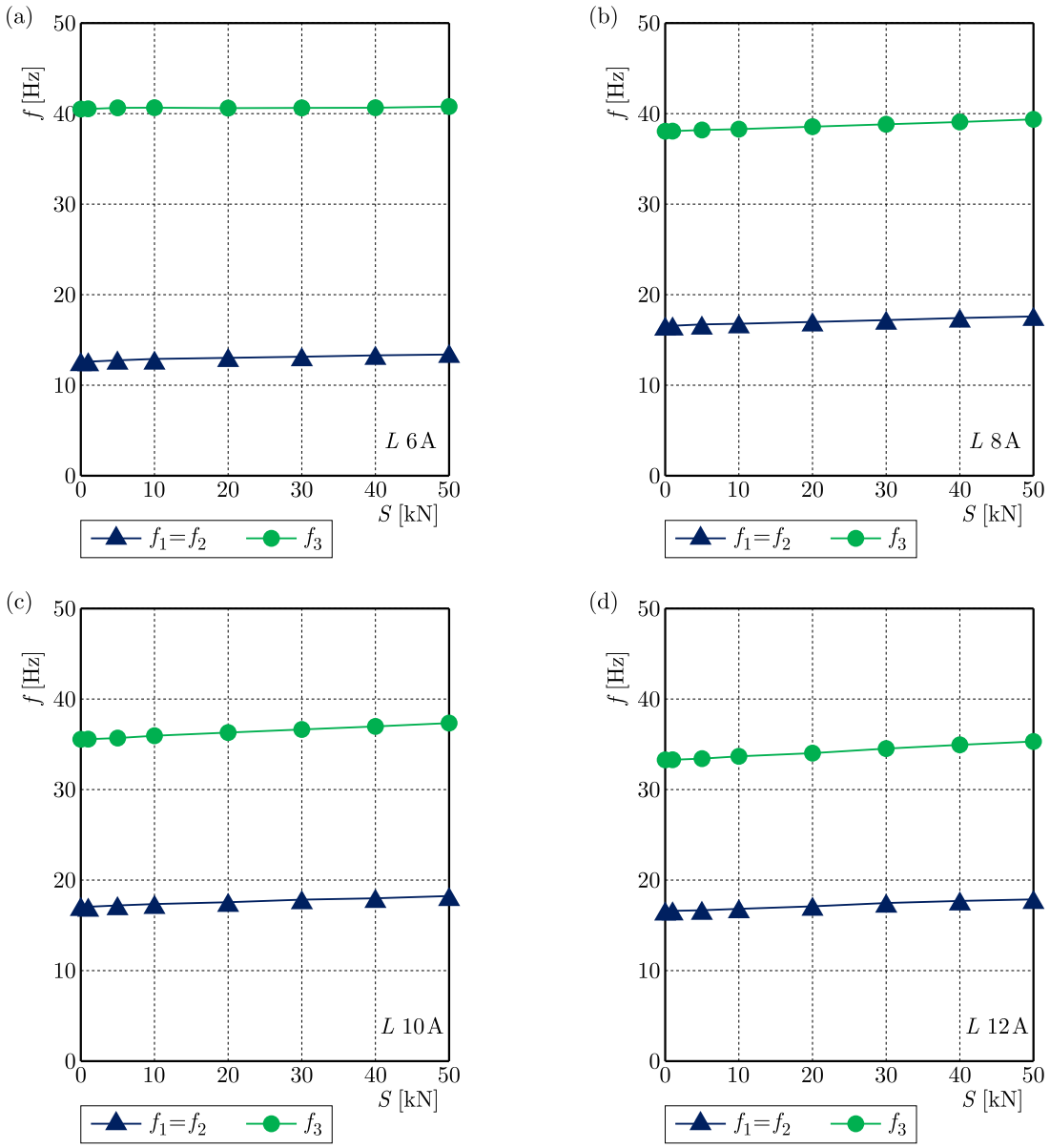


Fig. 9. Influence of the initial prestress S on the natural frequency for Levy domes type A: (a) L 6A, (b) L 8A, (c) L 10A, (d) L 12A

In the case of Levy domes type B, the first natural $f_1(0)$ and free $f_1(P)$ frequencies corresponding to the one identified mechanism are shown in Table 6. The external force placed in the first position (1st node) corresponds with a node of the upper section of the dome causing too much disturbance of the equilibrium state, and it is impossible to obtain the minimal prestress. The second position (3rd node) corresponds with a node on the hoop of the second section, and S_{min} is equal to 50 kN. The third position (5th node) corresponds with a node of the third section, and S_{min} is equal to 12 kN. The free frequencies are changing almost linearly.

In turn, in the case of Geiger domes, the first and last natural $f_i(0)$ and free $f_i(P)$ frequencies corresponding to the mechanisms are presented in Table 7 (for domes type A) and in Table 8 (for domes type B). For the Geiger domes type A, the third position (5th node) corresponding with a node of the third section is the worst, and S_{min} is equal to 36 kN, for both free and natural frequencies. The second position (3rd node) corresponds with a node on the hoop of the second section, and S_{min} is equal to 34 kN. The first position (1st node) corresponds with a node of the

Table 5. First natural $f_1(0)$ and free $f_1(P)$ frequency [Hz] for dome L 6A

ϕS [kN]	$f_1(0)$	$f_1(P)$					
		1st node		3rd node		5th node	
		1 kN	5 kN	1 kN	5 kN	1 kN	5 kN
0	12.84						
1	12.86					12.86	
4	12.91	12.91				12.91	
5	12.92	12.92				12.92	12.92
9	12.99	12.99		12.99		12.99	12.99
10	13.01	13.01		13.00		13.01	13.00
18	13.14	13.14	13.14	13.14		13.14	13.13
20	13.17	13.18	13.18	13.17		13.18	13.17
30	13.34	13.34	13.34	13.34		13.34	13.33
40	13.50	13.50	13.51	13.50		13.50	13.50
42	13.53	13.54	13.53	13.54	13.52	13.53	13.53
50	13.66	13.67	13.67	13.66	13.65	13.66	13.66

Table 6. First natural $f_1(0)$ and free $f_1(P)$ frequency [Hz] for dome L 6B

ϕS [kN]	$f_1(0)$	$f_1(P)$					
		1st node		3rd node		5th node	
		1 kN	5 kN	1 kN	5 kN	1 kN	5 kN
0	0.00						
1	2.49						
3	4.32					4.23	
5	5.57					5.49	
10	7.88			7.51		7.84	
12	8.63			8.29		8.56	8.26
20	11.15			10.88		11.11	10.84
22	11.69	11.54		11.38		11.64	11.35
30	13.65	13.53		13.42		13.62	13.38
40	15.76	15.65		18.04		15.74	15.49
50	17.62	17.53		19.12	15.53	17.59	17.34

upper section of the dome, and S_{min} is equal to 11 kN. However, for the Geiger dome type B, the first position (1st node) corresponding with a node of the upper section is the worst, and S_{min} is equal to 41 kN. The second position (3rd node) corresponds with a node on the hoop of the second section, and S_{min} is equal to 26 kN. The third position (5th node) corresponds with a node of the third section, and S_{min} is equal to 2 kN.

4. Conclusions

In this paper, the dynamic behaviour of tensegrity domes is explored. It is well known that the number of prestress-dependent frequencies is equal to the number of infinitesimal mechanisms. In the absence of prestress, these frequencies are zero, and the corresponding forms of vibrations implement the mechanisms. After introducing the initial prestress, the frequencies increase. If several mechanisms are identified, the higher frequencies are more sensitive to the initial prestress changes. The sensitivity of these natural frequencies to the initial prestress is so great that the

Table 7. Natural $f_i(0)$ and free $f_i(P)$ frequency [Hz] for dome G 6A

S [kN]	$f_1(0)$	$f_1(P)$						$f_8(0)$	$f_8(P)$					
		1st node		3rd node		5th node			1st node		3rd node		5th node	
		1 kN	5 kN	1 kN	5 kN	1 kN	5 kN		1 kN	5 kN	1 kN	5 kN	1 kN	5 kN
0	0.00							0.00						
1	0.73							1.75						
3	1.26	1.19						3.03	3.02					
5	1.62	1.56						3.91	3.90					
8	2.05	1.99		2.08				4.95	4.94		5.27			
10	2.30	2.26		2.31				5.52	5.52		5.73			
11	2.41	2.35	2.22	2.39				5.79	5.79	5.79	5.95			
12	2.51	2.46	2.34	2.49		2.51		6.05	6.05	6.05	6.17		6.06	
20	3.25	3.22	3.12	3.23		3.23		7.81	7.81	7.80	7.85		7.80	
30	3.98	3.96	3.88	3.96		3.96		9.57	9.56	9.55	9.58		9.56	
34	4.23	4.22	4.19	4.20	4.20	4.23		10.19	10.18	10.17	10.22	10.44	10.25	
36	4.36	4.35	4.33	4.32	4.32	4.36	4.31	10.48	10.48	10.45	10.52	10.71	10.56	10.47
40	4.59	4.58	4.51	4.58	4.55	4.58	4.54	11.05	11.05	11.04	11.05	11.22	11.04	11.03
50	5.13	5.12	5.06	5.12	5.08	5.12	5.09	12.35	12.35	12.34	12.35	12.45	12.34	12.32

Table 8. Natural $f_i(0)$ and free $f_i(P)$ frequency [Hz] for dome G 6B

S [kN]	$f_1(0)$	$f_1(P)$						$f_8(0)$	$f_8(P)$					
		1st node		3rd node		5th node			1st node		3rd node		5th node	
		1 kN	5 kN	1 kN	5 kN	1 kN	5 kN		1 kN	5 kN	1 kN	5 kN	1 kN	5 kN
0	0.00							0.00						
1	0.72							4.21						
2	1.02					1.41	2.16	5.96					8.64	13.21
5	1.61					1.76	2.33	9.42					10.56	14.20
10	2.28			2.32		2.29	2.61	13.32			13.56		13.52	15.87
14	2.70	2.69		2.72		2.66	2.92	15.77	18.91		15.71		15.66	17.45
20	3.23	3.22		3.22		3.13	3.28	18.84	20.72		18.77		18.85	19.61
26	3.68	3.67		3.55	3.72	3.69	3.68	21.48	22.67		21.33	21.65	21.47	21.65
30	3.96	3.94		3.94	3.97	3.94	3.94	23.08	23.92		22.97	23.08	23.06	23.33
40	4.57	4.55		4.55	4.54	4.56	4.53	26.65	27.06		26.54	26.42	26.63	26.72
41	4.62	4.61	4.57	4.61	4.58	4.69	4.59	26.98	27.31	32.27	26.97	26.79	27.16	26.91
50	5.11	5.09	5.06	5.09	5.07	5.10	5.06	29.79	30.01	33.81	29.69	29.47	29.77	29.79

change in the level of prestress can be successfully used to control the dynamic properties of the structure. Theoretically, other frequencies should be practically insensitive to self-stress changes. However, in the case of some analyzed domes, i.e., Levy domes type B and Geiger domes type A, it is different. There are additional frequencies that depend on the initial prestress. In the absence of prestress, the additional frequencies, unlike to frequency corresponding to the mechanism, are not zero. The number of them, and the sensitivity to initial prestress changes, depends on the kind of dome and number of girders.

Comparing all the results, we can say that due to the lack of the mechanisms, i.e., Levy dome type A, the natural and free frequencies are practically not affected by the initial prestress, independent of the number of load-bearing girders.

The considerations contained in this paper indicate the unusual behaviour of tensegrity domes. The obtained results are important for dynamic stability analysis of behaviour of tenseg-

rity structures under periodic loads, which will be the subject of feature investigation. The dynamic stability analysis cannot be carried out without the analysis presented in this paper.

References

1. Arsenault M., Gosselin C.M., 2005, Kinematic, static, and dynamic analysis of a planar one--degree-of-freedom tensegrity mechanism, *Journal of Mechanical Design*, **127**, 6, 1152-1160
2. Atig M., El Ouni M.H., Kahla N.B., 2017, Dynamic stability analysis of tensegrity systems, *European Journal of Environmental and Civil Engineering*, **23**, 6, 675-692
3. Bel Hadj Ali N., Rhode-Barbarigos L., Smith I.F.C., 2010, Analysis of clustered tensegrity structures using a modified dynamic relaxation algorithm, *International Journal of Solids and Structures*, **48**, 5, 637-647
4. EN 1993-1-11, 2006, Eurocode 3: Design of steel structures – Part 1-11: Design of structures with tension components
5. Geiger D.H., 1988, *Roof Structure*, U.S. Patent 4 736 553
6. Geiger D.H., Stefaniuk A., Chen D., 1986, The design and construction of two cable domes for the Korean Olympics, *Proceeding of the IASS Symposium on Shells, Membranes and Space Frames*, Osaka, Japan, 265-272
7. Gilewski W., Kłosowska J., Obara P., 2016, Form finding of tensegrity structures via Singular Value Decomposition, *Advances in Mechanics: Theoretical, Computational and Interdisciplinary Issues*, 191-195
8. Kawaguchi M., Tatemichi I., Chen P.S., 1999, Optimum shapes of a cable dome structure, *Engineering Structures*, **21**, 8, 719-725
9. Kim S.-D., Sin I.-A., 2014, A comparative analysis of dynamic instability characteristic of Geiger--typed cable dome structures by load condition, *Journal of the Korean Association for Spatial Structures*, **14**, 1, 85-91
10. Lee J., Tran H.C., Lee K., 2009, Advanced form-finding for cable dome structures, *Proceeding of IASS Symposium*, 2116-2127
11. Levy M.P., 1989, Hypar-tensegrity dome, *Proceeding of International Symposium on Sports Architecture*, Beijing, China, 157-162
12. Levy M.P., 1994, The Georgia dome and beyond: achieving lightweight-longspan structures, *Proceedings of the IASS-ASCE International Symposium*, Atlanta, USA, 560-562
13. Levy M., Jing T.-F., Brzozowski A., Freeman G., 2013, Estadio ciudad de La Plata (La Plata Stadium), *Structural Engineering International*, **23**, 303-310
14. Li Q., Skelton R.E., Yan J., 2011, Energy optimization of deployable tensegrity structure, *Proceeding of the 30th Chinese Control Conference*, Yantai, China, 12383827
15. Ma Q., Ohsaki M., Chen Z., Yan X., 2018, Step-by-step unbalanced force iteration method for cable-strut structure with irregular shape, *Engineering Structures*, **177**, 331-344
16. Obara P., 2019, *Dynamic and Dynamic Stability of Tensegrity Structures* (in Polish), Wydawnictwo Politechniki Świętokrzyskiej, Kielce, Poland
17. Obara P., Solovei M., 2023, Assesment of the impact of the number of girders on the dynamic behaviour of Geiger dome, *Archives of Civil Engineering*, **69**, 3, 597-611
18. Obara P., Solovei M., Tomasik J., 2023a, Genetic algorithm as a tool for the determination of the self-stress states of tensegrity domes, *Applied Sciences*, **13**, 9, 5267
19. Obara P., Solovei M., Tomasik J., 2023b, Qualitative and quantitative analysis of tensegrity steel domes, *Bulletin of Polish Academy of Sciences*, **71**, 1, 1-8

20. Oribasi A., Paronesso A., Dauner H.-G., 2002, The new world cycling center in aigle, *Stahlbau*, **71**, 584-591

21. Qin W., Gao H., Xi Z., Feng P., Li Y., 2023, Shaking table experimental investigations on dynamic characteristics of CFRP cable dome, *Engineering Structures*, **281**, 115748

22. Rebielak J., 2000, Structural system of cable dome shaped by means of simple form of spatial hoops, *Proceeding of Lightweight Structures in Civil Engineering*, 114-115

23. Shen X., Zhang Q., Lee D.S.H., Cai J., Feng J., 2021, Static behavior of a retractable suspen-dome structure, *Symmetry*, **13**, 7, 1105

24. Sun G., Xiao S., Xue S.,, 2021, Test and numerical investigation mechanical behavior of cable dome, *International Journal of Steel Structures*, **21**, 4, 1502-1514, 2021

25. Tran H.C., Lee J., 2013, Form-finding of tensegrity structures using double singular value decomposition, *Engineering with Computers*, **29**, 71-86

26. Yuan X., Chen L., Dong S., 2007, Prestress design of cable domes with new forms, *International Journal of Solids and Structures*, **44**, 2773-2782

27. Zhang P., Feng J., 2017, Initial prestress design and optimization of tensegrity systems based on symmetry and stiffness, *International Journal of Solids and Structures*, **106-107**, 68-90

28. Zhang J.Y., Ohsaki M., 2006, Adaptive force density method for form-finding problem of tensegrity structures, *International Journal of Solids and Structures*, **43**, 18-19, 5658-5673

29. Zienkiewicz O. C., Taylor R.L., 2000, *The Finite Element Method. Vol. 1. The Basis*, Elsevier Butterworth-Heinemann, London, Great Britain

Manuscript received October 30, 2023; accepted for print December 9, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 269-278, Warsaw 2024
https://doi.org/10.15632/jtam-pl/184262

THE PROBABILISTIC MODEL FOR SYSTEM RELIABILITY ANALYSIS OF A STEEL PLANE AND SPATIAL TRUSSES[1]

Katarzyna Kubicka
Kielce University of Technology, Kielce, Poland
e-mail: k.kubicka@tu.kielce.pl

The article focuses on the system reliability analysis of steel trusses (plane and spatial). The computations are realized by the use of a developed by the author C++ code. The following loads are taken into account: self-weight, weight of coverings, wind and snow. The limit state function is defined as a difference between the bearing capacity and the effect of action of an element. The paper presents how effective tool is the system reliability analysis compared with traditional structural design methods. The methods of transforming Gumbel distribution into normal and generating random variables are described.

Keywords: system reliability analysis, plane steel truss, spatial steel truss, cut-sets, normal distribution transformation

1. Introduction

Nowadays, the reliability of structures is very important topic. The methods of reliability and optimization are constantly developed, improved and their meaning in design constantly increases. Nothing unusual, because exactly these methods seem to be the solution of the problem how to design with high safety, but with as costs low as possible.

The reliability methods can be divided into two groups: considering the reliability of single elements and the reliability of the whole structural systems. In the first group one can enumerate approximation methods as: FORM (Breitung, 2015; Ditlevsen, 1987; Keshtegar and Meng,2017), SORM (Cai and Elishakoff, 1994; Hu *et al.*, 2021) and simulation methods like: Monte Carlo (Rausch *et al.*, 2019; Sharma, 2020; Zaeimi and Ghoddosain, 2020), Importance Sampling (Melchers, 1989; Papaioannou *et al.*, 2018), Artificial Neural Networks (Flood, 2008; Potrzeszcz-Sut and Dudzik, 2022).

The presented article uses a system reliability method, whose basics have been well-known for decades. Instead of this fact, as the method is not easy to implement, especially for big structures (with many possible causative elements) its application is limited. Nevertheless, in the last years the system reliability analysis has been successively becomming more popular, especially for steel truss analysis (Mochocki *et al.*, 2018; Park *et al.*, 2004, Zabojszcza and Radoń, 2020).

In the article, the plane and the spatial truss were analysed. For both types of structures the main task was to identify so-called cut-sets, i.e. the possible way of transforming the structure into a mechanism. In other words, the cut-sets are such sets of elements whode failure determines the failure of the whole structure. The method of searching cut-sets for plane trusses was presented in the previous author's papers (Kubicka *et al.*, 2019; Kubicka, 2022; Kubicka and Sokol, 2023). Generally, the method is based on spectral matrix stiffness analysis and is appropriate for both persistent and accidental (fire) design situation. If the cut-sets for the analysed structure are known, it is possible to define the reliability system of a mixed type, that is a parallel-series

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

system, which is a combination of two basic types of systems, which will be described in the following part of the paper.

2. Materials and methods

In the paper, the system reliability method was applied. The following steps of this method are presented in Fig. 1 with comparison with a traditional design. The red frames indicate the steps, during which the random character of variables is taken into account.

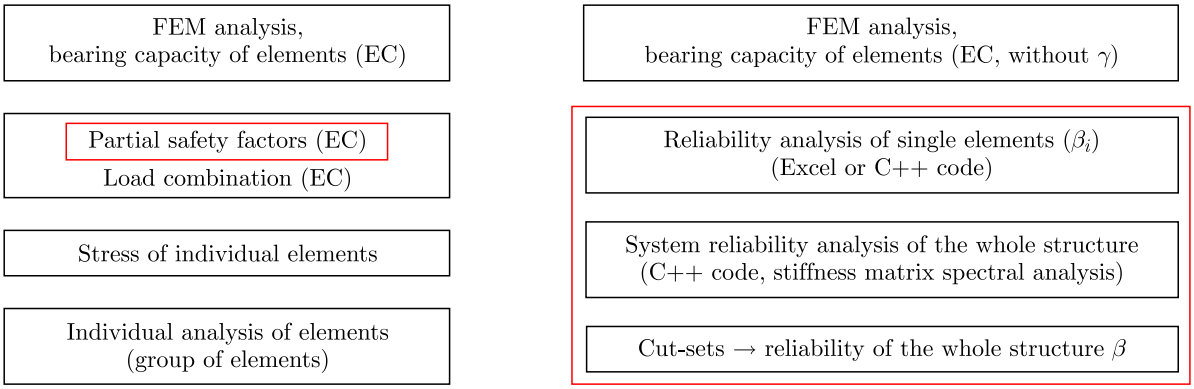


Fig. 1. Comparison of the traditional design method (a) with the proposed probabilistic method (b)

In Fig. 1, γ corresponds to the partial safety coefficient, β is the reliability index, which will be defined in the following part of the article, subscript i informs that the value is defined for i-th element.

The proposed method has a few advantages, including:

- Possibility to decide which values are random and which are deterministic
- Coefficient of variation can be defined individually for each random variable
- Possibility to use different types of distribution

After computing the bearing capacity and conducting FEM analysis, knowing the effect of action in individual bars, it is possible to compute reliability for each individual element according to Table 1. In this method, all variables (N, Eff) must have normal distribution, in other case the transformation to normal distribution is essential.

Table 1. The method of computing reliability for individual elements

¢Number of element	Bearing capacity (N) Effect of action (Eff)	¢Safety margin (M)	Reliability index of element (β_i)	Probability of elements failure (P_{fi}) Reliability of element (R_i)
i	$N_i(\mu_{Ni}, \mu_{Ni})$	$M_i = N_i - Eff_i$	$\beta_i = \frac{\mu_{Mi}}{\sigma_{Mi}}$	$P_{fi} = \Phi(-\beta_i)$
K	$Eff_i(\mu_{Eff_i}, \sigma_{Eff_i})^*$	$\mu_{Mi} = \mu_{Ni} - \mu_{Eff_i}{}^*$		$R_i = 1 - P_{fi}$
		$\sigma_{Mi} = \sqrt{\sigma_{Ni}^2 + \sigma_{Eff_i}^2}{}^*$		Φ – Laplace function

* μ_{Xi} – mean value of X value for i-th element,
σ_{Xi} – standard deviation of X value for i-th element

Knowing the reliability of single elements, it is possible to compute the reliability of the whole structure. To realize this task, it is essential to find so-called cut-sets, i.e. the way the structure

can transform into a mechanism. On this basis, a mixed system is created, whose elements are connected in groups in a parallel way, and then are connected in series. Therefore, the parallel systems have to be computed and than the series system. The series system is reliable as long as all of its elements are reliable. In other words, the failure of a single element determines the failure of the whole structure. The reliability of a series system is computed according to the following formula

$$R = \prod_{i=1}^{n} R_i \tag{2.1}$$

The parallel system is such a system which is reliable as long as at least one of its element is reliable. The reliability of such a type of system is computed according to

$$R = 1 - \prod_{i=1}^{n} (1 - R_i) \tag{2.2}$$

In the paper, plane and spatial truss were analysed according to the algorithm presented in Fig. 1 with a simple and more advanced probabilistic model described in the following part. FEM analysis was conducted in Robot Structural Analysis program, computation connected with reliability was realized in the author C++ code.

2.1. Simplified probabilistic model

Simplified probabilistic model was used in some of the previous author's papers. In this model, it is assumed that the truss element may fail if the bearing capacity of the element exceeds the effect of action. Therefore, the limit state function g can be written as

$$g = N - E_{eff} \tag{2.3}$$

where N is defined as in Eq. $(2.4)_1$ for compressed elements, and in Eq. $(2.4)_2$ for element in tension

$$N_{b,fi} = \chi_{fi} A f_y \qquad N_{c,fi} = A f_y \tag{2.4}$$

In this model, it is assumed that the buckling coefficient χ_{fi} is a deterministic value computed according to Eurocode. In Table 2, characteristics of all variables are presented, all of them are assumed to have normal distribution.

Table 2. The characteristic of variables used in the simplified model

Value	Deterministic or probabilistic	Coefficient of variation ν [%]	Distribution type
Buckling coefficient χ_{fi}	deterministic	–	–
Cross-sectional area A	probabilistic	8	normal
Yield strength f_y	probabilistic	6	normal
Effect of action E_{eff}	probabilistic	6	normal

Because of the formulation of the limit state function (product of random variables) and assumption about normal distribution, the coefficient of variation for bearing capacity can be approximated according to the following formula (Biegus, 1999)

$$\nu_N = \sqrt{\nu_{fy}^2 + \nu_A^2} = \sqrt{0.06^2 + 0.08^2} = 0.1 = 10\% \tag{2.5}$$

2.2. "Full" probabilistic model

In the previous work (Kubicka and Radoń, 2018) it was demonstrated that treating the buckling coefficient as a deterministic value leads to getting significantly different results than in the case of considering it random. What is more, the buckling coefficient is a function of few variables, where some of them (A, f_y) were defined random

$$\chi_{fi} = f(f_y, A, E, I_y, L) \tag{2.6}$$

what also suggest that the value χ_{fi} should be defined probabilistic. So, the "full" probabilistic model was extended by the assumption that the buckling coefficient is a random value. In the function of buckling coefficient only length of element L was assumed to be deterministic. All other variables, i.e. cross-sectional area A, yield strength f_y, Young's modulus E and moment of inertia I_y, were treated probabilistic. The assumption that all random variables have normal distribution was abandoned, especially it was assumed that atmospheric loads (wind, snow) have Gumbell distribution. Therefore, the transformation to normal distribution had to be made.

2.2.1. Transformation from Gumbell to normal distribution

In the article, the transformation from Gumbell to normal distribution was realized according to two methods, namely the method of moments and collocation point method (Murzewski, 1989):

— method of moments

$$\overline{x} = \widetilde{x} + Cu_x \qquad C = 0.57721 \qquad \mu_x = \frac{\pi}{\sqrt{6}} u_x \tag{2.7}$$

— collocation point method

$$\overline{x} = \widetilde{x} + t_N u_x \qquad t_N = 0.3457 \qquad \mu_x = \frac{u_x}{\theta_N} \qquad \theta_N = 0.9762 \tag{2.8}$$

where $\overline{x}$, μ_x are characteristics of normal distribution and symbols $\widetilde{x}$, u_x correspond to Gumbell distribution.

2.2.2. Generation of random samples using the Box-Müller algorithm

In the presented method, the set of random variables was generated for each probabilistic value and, on this base, the mean value and standard deviation were computed. A part of code is presented in Fig. 2. The generator of random variables available in C++ generates variables with uniform distribution. To conduct reliability analysis, it is necessary to have variables with normal distribution. One of the method that enables transforming uniform variables into variables with normal distribution is using the Box--Müller algorithm (https://mathworld.wolfram.com/Box-MullerTransformation.html).

In this method, the variable with normal distribution Y_1 is generated on the base of two random variables with uniform distribution (y_1, y_2), according to the formula

$$Y_1 = \cos(2\pi y_2)\sqrt{-2\ln y_1} \tag{2.9}$$

After generation of variables with normal distribution, the loop that generates random values of variables (A, f_y, E, I_y) is started. It is realized n_s (number of simulation) times, in the presented article $n_s = 100$. In each realization of the loop, the initially assumed standard deviation of a random variable was multiplied by a random value generated previously, and it was added to the mean value of the random variable. Symbol A[i] and Iy[i] means the cross-sectional area and yield strength of the i-th structure element. sA, sfy, sE, sIy define the initial standard deviation.

Having 100 samples of rA, rfy, rE, rIy, the mean value and the standars deviation were computed for each random value, and these characteristics were taken into account in the further reliability analysis.

```
double normalRandom()
{
  double y1=RandomGenerator();
  double y2=RandomGenerator();
  return cos(2*M_PI*y2)*sqrt(-2.*log(y1));
}
   for (int z=1; z<=ns; z++)
   {
       double rA = normalRandom()*sA+A[i];
       double rfy = normalRandom()*sfy+fy;
       double rE = normalRandom()*sE+E;
       double rIy = normalRandom()*sIy+Iy[i];
```

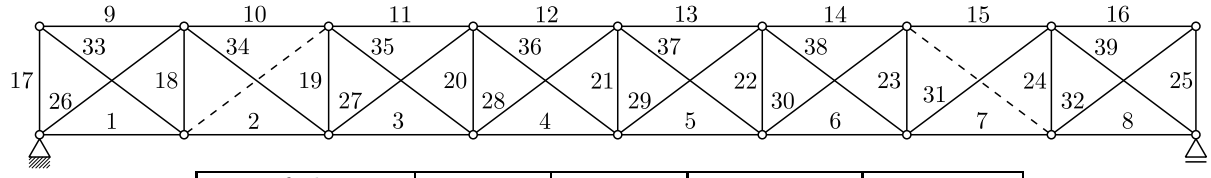
Fig. 2. The part of C++ code generating random samples of variables using the Box-Müller algorithm

3. Results

In the paper, two types of trusses (plane and spatial) were analysed. Both of them are statically indeterminate, because in the case of such types of structure the advantages of system reliability is most visible.

3.1. Plane truss

The plane truss analysed in the paper is presented in Fig. 3. Two of cross-braces are drawn by the dashed line. They were not taken into account during reliability analysis, because FEM analysis indicated that the bearing capacity in this elements is exceeded. Nevertheless, reliability analysis indicated that the structure as a whole is safe despite the fact that these two elements are unreliable.



No. of element	1-8	9-16	17,25,26,39	27-38
Profile	IPE 160	IPE 300	CHS 108x5	CHS 70x5

Fig. 3. The analysed plane truss

It was assumed that the structure was loaded by: self-weight (sw), cover (c), snow-3rd zone(s) and wind-1st zone(w). Each type of load was considered individually during the reliability analysis.

The cut-sets identified for the truss are presented in Table 3. During searching cut-sets, only posts (17-25) and cross-braces (26-39) were taken into account as the most probable elements of failure. As from the system reliability analysis point of view, the most important are initial cut-sets, consisting of few elements. The searching for cut-sets was limited to 4-elements. Usually, in the case of cut-sets consisting of a higher number of elements the reliability is equal to one. The more elements in the cut-set, the more probably some of them will have reliability equal 1.0. According to Eq. (2.2), in such a case, the reliability of a parallel subsystem is equal 1, what does not change the final reliability of the structure.

The reliability of each element under a single type of load: self-weight (sw), cover (c), snow (s) or wind (w) was computed according to the procedure presented in Table 1. Then reliabilities of the whole structure under the individual type of load (R_{sw}, R_c, R_s, R_w) were computed according to the cut-sets presented in Table 3. Elements in the brackets are connected in parallel (for example $\{17, 18\}$, $\{19, 20, 28\}$, $\{19, 20, 21, 29\}$), so the reliabilities of such a set of elements are

computed according to Eq. (2.2). Then all set of elements in the brackets are connected in series, what is computed according to Eq. (2.1). The final reliability R_{fin} is computed according to Eq. (3.1) as a product of reliabilities under individual loads. This corresponds to a series system and is correct for the most unsafe situation, when all types of loads act together

$$R_{fin} = R_{SW} R_C R_S R_W \tag{3.1}$$

Table 3. The cut-sets identified for a plane truss

1-element cut-sets	2-element cut-sets	3-element cut-sets	4-element cut-sets
{31}	{17, 18}	{19, 20, 28}	{19, 20, 21, 29}
{34}	{17, 26}	{19, 20, 36}	{19, 20, 21, 37}
	{17, 33}	{20, 27, 28}	{20, 21, 27, 29}
	{18, 26}	{20, 27, 36}	{20, 21, 27, 37}
	{18, 33}	{20, 28, 35}	{20, 21, 29, 35}
	{19, 27}	{20, 35, 36}	{20, 21, 35, 37}
	{19, 35}	{21, 28, 29}	{21, 22, 23, 28}
	{23, 30}	{21, 28, 37}	{21, 22, 23, 36}
	{23, 38}	{21, 29, 36}	{21, 22, 28, 30}
	{24, 25}	{21, 36, 37}	{21, 22, 28, 38}
	{24, 32}	{22, 23, 29}	{21, 22, 30, 36}
	{24, 39}	{22, 23, 37}	{21, 22, 36, 38}
	{25, 32}	{22, 29, 30}	
	{25, 39}	{22, 29, 38}	
	{26, 33}	{22, 30, 37}	
	{27, 35}	{22, 37, 38}	
	{28, 36}		
	{29, 37}		
	{30, 38}		
	{32, 39}		

In the paper during computing, the final reliability simplified assumption was made, one type of load for snow and wind was taken into account. To be more precise, these reliabilities should be considered as the following functions

$$R_S = R(S_1, S_2) \qquad R_w = R(W_1, \ldots, W_n) \tag{3.2}$$

The reliability of the truss presented in Fig. 3 was estimated in few attempts. The first attempt was completely conducted according to a simplified probabilistic model, described in Section 2.1. In the second attempt, the model was extended to take into account different types of distribution (for wind and snow Gumbell distribution was assumed). So, some transformation method to normal distribution had to be applied (method of moments and collocation point method). In the last attempt, the (3rd) "full" model was applied. Characteristics of random variables are presented in Table 4.

In Table 5, the results of system reliability analysis, conducted according to the previously described method are presented.

It is noticeable that the reliability index β is slightly different in the subsequent attempts. But the structure, according to each attempt, is reliable because $\beta > 3.8$, what is the minimum value recommended by Eurocode (EN-1990, 2002).

Table 4. The variables used in the "full" model for the plane truss

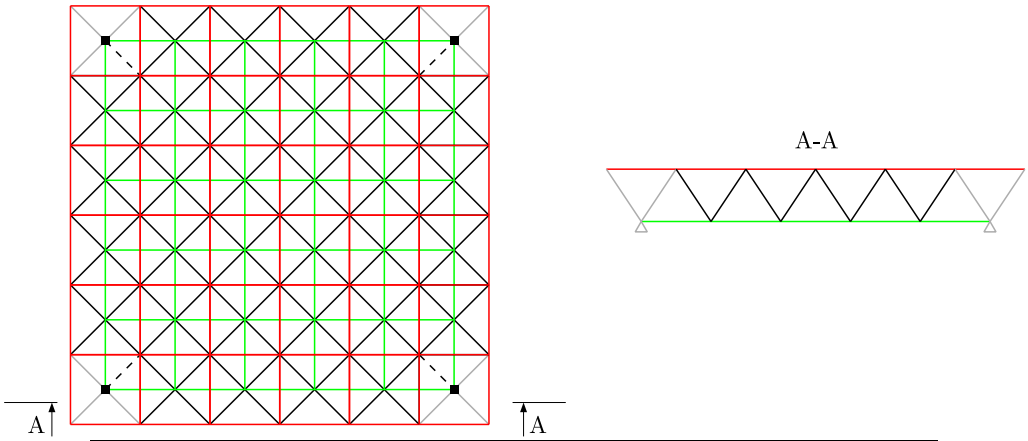
Random variable	Coefficient of variation [%]
Yield strength f_y	6
Cross-sectional area A	8
Modulus of elasticity E	5
Moment of inertia J	8
Effect of action *Eff*	6

Table 5. Reliability analysis results for the plane truss

No. of attempt	1	2		3
		Method of moments	Collocation point method	
R_{sw}	1.0	1.0	1.0	$0.(9)^{12}88$
R_c	$0.(9)^{8}829535$	$0.(9)^{8}829535$	$0.(9)^{8}829535$	$0.(9)^{5}8634554069$
R_s	$0.(9)^{7}84158383$	$0.(9)^{7}71059492$	$0.(9)^{7}83384258$	$0.(9)^{5}8634554069$
R_w	1.0	1.0	1.0	$0.(9)^{10}83968$
R_{fin}	$0.(9)^{7}82387918$	$0.(9)^{7}69289027$	$0.(9)^{7}69289027$	$0.(9)^{5}76909$
β	5.513	5.415	5.506	4.546

3.2. Spatial truss

The next example concerns the spatial truss presented in Fig. 4. Similarly, like in the case of the plane truss, some elements are drawn with dashed lines, what means they were excluded from the system reliability analysis. The reason was analogous as in the previous example, the bearing capacity of this elements was exceeded according to FEM analysis, but the system reliability analysis indicated that the structure is safe as a whole, despite of the fact that some of elements are unreliable as individuals.



¢Element	Bottom chord (green)	Top chord (red)	Cross-braces	
			grey	black
Profile	IPE 100	IPE 200	CHS 70x5	CHS 51x4

Fig. 4. The analysed spatial truss

Identification of cut-sets was realized again with the assumption that the cross-braces are most likely elements to fail. Creating the elements, 1-,2-, 3- and 4-element cut-sets are presented in Fig. 5. In Table 6, the identified cut-sets are presented. The interpretation is the same as in the case of the plane truss analysed in Section 3.1.

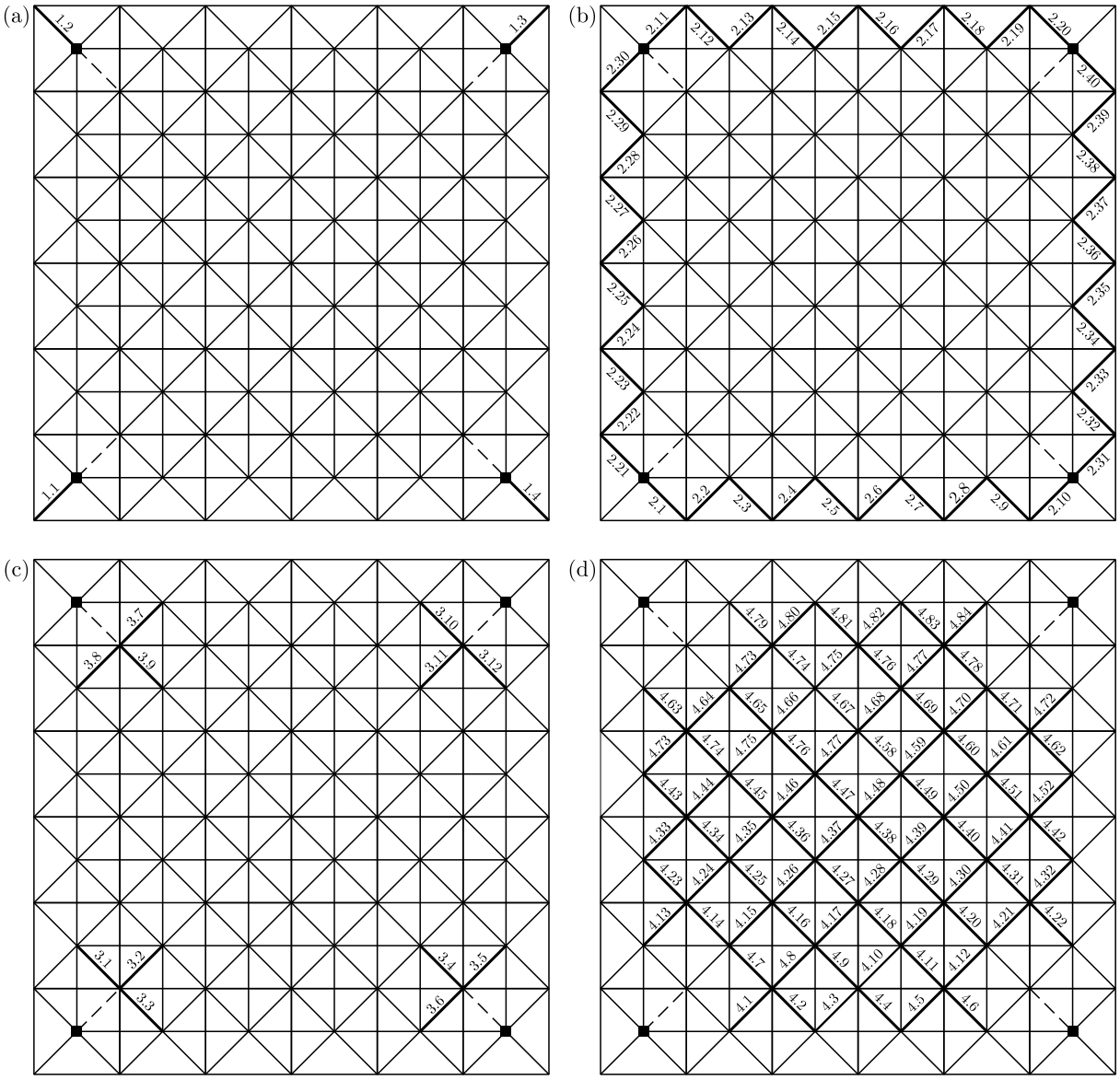


Fig. 5. Creating the elements (a) 1-, (b)2-, (c) 3-, (d) 4-element cut-sets for the analysed spatial truss

For the analysed spatial truss, the final reliability computed analogously as in the case of the plane truss was equal to 1.0, what means that the probability of failure of the structure is practically equal to 0. This is the result of the fact that the analysed structure is highly statistically indeterminate, so it is redundant what means that the failure of a single member merely changes the system behaviour and does not result in the collapse of the whole structure. That is why the system reliability analysis is a useful tool during designing of the structure. It allows the designer to follow possible failure path and decide which element may be designed without satisfying ULS.

Table 6. The cut sets identified for the analysed spatial truss

1-element cut-sets	2-element cut-sets	3-element cut-sets	4-element cut-sets
$\{1.1\}$	$\{2.1, 2.2\}$	$\{3.1, 3.2, 3.3\}$	$\{4.1, 4.2, 4.7, 4.8\}$
$\{1.2\}$	$\{2.3, 2.4\}$	$\{3.4, 3.5, 3.6\}$	$\{4.3, 4.4, 4.9, 4.10\}$
$\{1.3\}$	$\{2.5, 2.6\}$	$\{3.7, 3.8, 3.9\}$	$\{4.5, 4.6, 4.11, 4.12\}$
$\{1.4\}$	$\{2.7, 2.8\}$	$\{3.10, 3.11, 3.12\}$	$\{4.13, 4.14, 4.23, 4.24\}$
	$\{2.9, 2.10\}$		$\{4.15, 4.16, 4.25, 4.26\}$
	$\{2.11, 2.12\}$		$\{4.17, 4.18, 4.27, 4.28\}$
	$\{2.13, 2.14\}$		$\{4.19, 4.20, 4.29, 4.30\}$
	$\{2.15, 2.16\}$		$\{4.21, 4.22, 4.31, 4.32\}$
	$\{2.17, 2.18\}$		$\{4.33, 4.34, 4.43, 4.24\}$
	$\{2.19, 2.20\}$		$\{4.35, 4.36, 4.45, 4.46\}$
	$\{2.21, 2.22\}$		$\{4.37, 4.38, 4.47, 4.48\}$
	$\{2.23, 2.24\}$		$\{4.39, 4.40, 4.49, 4.50\}$
	$\{2.25, 2.26\}$		$\{4.41, 4.42, 4.51, 4.52\}$
	$\{2.27, 2.28\}$		$\{4.53, 4.54, 4.63, 4.64\}$
	$\{2.29, 2.30\}$		$\{4.55, 4.56, 4.65, 4.66\}$
	$\{2.31, 2.32\}$		$\{4.57, 4.58, 4.67, 4.68\}$
	$\{2.33, 2.34\}$		$\{4.59, 4.60, 4.69, 4.70\}$
	$\{2.35, 2.36\}$		$\{4.61, 4.62, 4.71, 4.72\}$
	$\{2.37, 2.38\}$		$\{4.73, 4.74, 4.79, 4.80\}$
	$\{2.39, 2.40\}$		$\{4.75, 4.76, 4.81, 4.82\}$
			$\{4.77, 4.78, 4.83, 4.84\}$

4. Conclusions

The presented analysis undoubtedly indicated that the system reliability analysis is an appropriate tool to estimate the reliability of both plane and spatial trusses. What is more, the proposed method is elastic, so the user can define characteristics of random variables. The method have to be developed especially by taking into account the load combination in the probabilistic point of view. What is more, in the initial research all processes were considered as time-independent. In fact, wind should be considered as time-dependent, what can be realized by using stochastic dynamics (Śniady, 2000). That is what the author is going to do in the nearest future. After this, the proposed probabilistic method can be considered as a complement to the traditional design method based on the partial safety factor. It seems that such an approach could result in limitation of the structure cost. Because of the redundancy of highly statically indeterminate structures it is not always necessary to select profiles for some group of elements according to "the weakest" element, what may lead to the situation that some of the elements are almost not stressed. Thanks to the system reliability analysis, it is possible to choose profiles with smaller dimensions, which reduces the volume of steel used for the structure and directly translates into not only lower costs, but also lower self-weight of the structure.

References

1. Biegus A., 1999, *The Probabilistic Analysis of Steel Structures* (in Polish), PWN, Warszawa-Wrocław
2. Breitung K., 2015, 40 years FORM: Some new aspects? *Probabilistic Engineering Mechanics*, **42**, 71-77

3. Cai G.Q., Elishakoff I., 1994, Refined second-order reliability analysis, *Structural Safety*, **14**, 4, 267-276
4. Ditlevsen O., 1987, On the choice of expansion point in FORM or SORM, *Structural Safety*, **4**, 243-245
5. EN-1990: Basis of structural design. European standard, 2002, European Committee for standardization Annex B
6. Flood I., 2008, Towards the next generation of artificial neural networks for civil engineering, *Advanced Engineering Informatics*, **22**, 4-14
7. https://mathworld.wolfram.com/Box-MullerTransformation.html
8. Hu Z., Mansour R., Olsson M., Du X., 2021, Second-order reliability methods: a review and comparative study, *Structural and Multidisciplinary Optimization*, **64**, 3233-3263
9. Keshtegar B., Meng Z., 2017, A hybrid relaxed first-order reliability method for efficient structural reliability analysis, *Structural Safety*, **66**, 84-93
10. Kubicka K., 2022, The new method of searching cut-sets in the system reliability analysis of plane steel trusses, *Applied Sciences*, **12**, 10, 1-19
11. Kubicka K., Obara P., Radoń U., Szaniec W., 2019, Assessment of steel truss fire safety in terms of the system reliability analysis, *Archives of Civil and Mechanical Engineering*, **19**, 2, 417-427
12. Kubicka K., Radoń U., 2018, Influence of randomness of buckling coefficient on the reliability index's value under fire conditions, *Archives of Civil Engineering*, **64**, 3, 173-179
13. Kubicka K., Sokol M., 2023, Fire safety of plane steel truss according to system reliability analysis combined with FORM method: The probabilistic model and SYSREL computation, *Applied Sciences*, **13**, 4, 2647
14. Mochocki W., Obara P., Radoń U., 2018, System-reliability analysis of steel truss towers, *MATEC Web of Conference*, **02001**, 1-8
15. Murzewski J., 1989, *Reliability of Engineering Structures* (in Polish), Arkady, Warszawa
16. Melchers R.E., 1989, Importance sampling in structural systems, *Structural Safety*, **6**, 1, 3-10
17. Papaioannou J., Breitung K., Straub D., 2018, Reliability sensitivity estimation with sequential importance sampling, *Structural Safety*, **75**, 24-34
18. Park S., Choi S., Sikorsky C., Stubbs N., 2004, Efficient method for calculation of system reliability of a complex structure, *International Journal of Solids and Structures*, **41**, 5035-5050
19. Potrzeszcz-Sut B., Dudzik A., 2022, The application of a hybrid method for the identification of elastic-plastic material parameters, *Materials*, **15**, 12, 4139 1-16
20. Rausch Ch., Nahangi M., Haas C., Liang W., 2019, Monte Carlo simulation for tolerance analysis in prefabrication and offsite construction, *Automation in Construction*, **103**, 300-314
21. Sharma M. K., 2020, Monte Carlo simulation applications for construction project management, *International Journal of Civil Engineering and Technology*, **11**, 88-100
22. Śniady P., 2000, *The Basics of Stochastic Dynamics of Structure* (in Polish), Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław
23. Zabojszcza P., Radoń U., 2020, Stability analysis of the single-layer dome in probabilistic description by the Monte Carlo method, *Journal of Theoretical and Applied Mechanics*, **56**, 2, 425-436
24. Zaeimi M., Ghoddosain A., 2020, System reliability based design optimization of truss structures with interval variables, *Periodica Polytechnica Civil Engineering*, **64**, 1, 42-59

Manuscript received October 30, 2023; accepted for print December 19, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 279-292, Warsaw 2024
https://doi.org/10.15632/jtam-pl/184231

DESIGN AND SIMULATION OF A MOBILE PLATFORM WITH A SEMI-ACTIVE SUSPENSION FOR UNEVEN TERRAIN[1]

Michał Olinski, Kacper Cholewa
Wroclaw University of Science and Technology, Faculty of Mechanical Engineering, Department of Fundamentals of Machine Design and Mechatronic Systems, Wroclaw, Poland
corresponding author Michał Olinski, e-mail: michal.olinski@pwr.edu.pl

The paper is focused on the design of a mobile wheeled platform able to move in uneven/unstructured terrains and particularly intended for supporting the work in agriculture. An independent double wishbone suspension is chosen to obtain a light and compact structure. Furthermore, a semi-active suspension with magnetorheological dampers and the ability to change the track of wheels is proposed to minimize uncontrolled vertical movements of the platform. A dynamic model is formulated to carry out simulations including various obstacles and cases with constant/controlled damping coefficients. As a final result, a conceptual CAD model is built with selected motors and standardized parts.

Keywords: robotics, agriculture, wheeled robot, wishbone suspension, magnetorheological damper

1. Introduction

The topic of vehicles and robots motion in uneven/harsh terrain has been studied by many researchers resulting in designing devices with various locomotion systems (Bruzzone and Quaglia, 2012). These are not only mobile platforms/robots equipped with tracks, legs (Garimella and Revzen, 2021; Raibert *et al.*, 2008) or wheels (Husti, 2019; Shamshiri *et al.*, 2018), but also more complex solutions of wheeled robots with high mobility (Shah *et al.*, 2012), including hybrid like wheel-legged devices (Niu *et al.*, 2018; Olinski and Ziemba, 2014; Sperzyński *et al.*, 2018). Mobile platforms need to move in unstructured environments (off-road terrains), hazardous surroundings, catastrophe sights (buildings, rubble), due to their various applications in transport, rescue (Niu *et al.*, 2018), exploration including for instance cultural heritage (Ceccarelli *et al.*, 2017) or space (Harrington and Voorhees, 2004), military, as well as in agriculture (Ackerman, 2015; Roldán *et al.*, 2018). Nowadays, the automation of agriculture by using mobile platforms and robots becomes crucial, since it not only helps enhancing productivity, but also improving safety by assisting human labor with heavy machinery, pesticides, etc. Therefore, the aim of this paper is to design a platform characterised by high mobility and maneuverability in uneven (off-road) terrains and particularly suitable for application/supporting work in agriculture.

The first step is to decide the type of the locomotion system. Walking systems are usually very energy consuming, complicated in construction/control, and their advantages including the ability to move in uneven terrains, to adapt the walking mode to the current terrain or regulate unit pressures on the ground do not seem to be significant for a designed light off-road mobile platform. Considered was also the possibility of using a tracked drive studied for instance for its advantageous low soil compaction in (Raper, 2004), or in (Chołodowski, 2023), where the movement resistance in tracked off-road vehicles was modeled. However, taking into account the fact of better maneuverability of wheeled vehicles and their lower costs, as well as lesser internal

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

resistances which translate into the possibility of selecting smaller motors, it was decided to use a wheeled chassis. It can also move on any terrain, and wheels with tires in combination with an appropriate suspension and control systems, can provide good stability and maneuverability. Thus, the focus in this paper is placed on wheel vehicles and suspension systems.

In the process of designing a wheeled vehicle that can move efficiently in unstructured terrains (off-road), the key stage is the appropriate suspension design. It provides the possibilities of, among others, continuous generation of the traction force by minimizing the time of slip and separation of wheels from the ground. As a consequence, the device is able to overcome obstacles, reduce vibrations that lead to wear of its equipment (sensors, cameras, etc.), as well as to maintain correct height of the body, minimizing its uncontrolled movements. The suspension consists of a set of movable, rigid and elastic parts connecting the wheels with the chassis. The suspension systems can be generally divided into dependent (vertical movement of one wheel has an impact on other wheels) and independent (each wheel moves on its own); as well as into passive, active and semi-active (due to the level of controlling suspension parameters).

The most commonly applied suspension systems are passive, where the structure consists of deformable and damping elements with constant/unchanging characteristics. An active suspension is a system that uses motors and electronics to control suspension damping and stiffness in real time. Alternatively, a semi-active suspension cannot change the stiffness of its elements, but is able to modify the damping force in real time. This solution is cheaper, simpler to build, less prone to failures and consumes much less energy than fully active suspension systems (Fischer and Isermann, 2003).

Furthermore, within years, many configurations of suspension systems were developed. One of them is a multi-rocker suspension system shown in Fig. 1a and applied in Scout 2.0 robot. Others include, for instance, a swing-arm (trailing arm) as in Scout Mini from AgileX company or a rocker-bogie suspension system widely used in space vehicles like Mars rovers (Harrington and Voorhees, 2004). However, many existing wheeled mobile robots have limited capabilities of maintaining a stable position/orientation of the platform, since instead of using a suspension system they depend only on deformability of wheels/tires. Examples are BoniRob (Ackerman, 2015) or ecoRobotix (Fig. 1b), a prototype of platform monitoring agricultural fields, recognizing and spraying weeds (Ecorobotix, 2019).

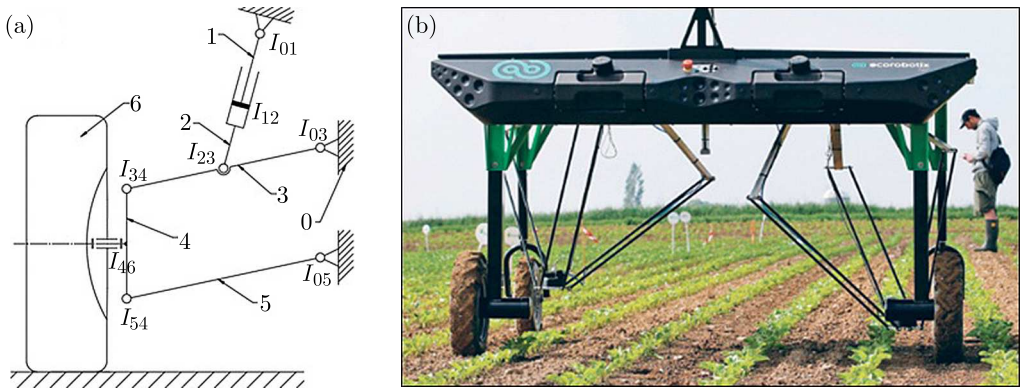


Fig. 1. Views of existing: (a) independent multi-rocker suspension with real mobility equal to 1, (b) ecoRobotix wheeled mobile platform for agricultural applications (Ecorobotix, 2019)

For the above presented reasons, it is concluded that there is still a need for lightweight mobile platforms that could move in uneven terrains and perform tasks to support agriculture. A particular aim is designing a suspension system that would allow one to minimize the uncontrolled vertical movement and vibrations, so the platform can be easily applied in various conditions to a wide range of tasks. This way it will also be ready for installation of different types of additional equipment and sensors, especially manipulators, cameras, lidars, etc. Thanks

to these solutions, a complete device could achieve independence and perform planned tasks autonomously. Therefore, in the paper, the kinematic form of the chassis with the suspension and selected dimensions will be determined. This will allow building a numerical model (including the estimated masses of elements) and conduct experiments. One of the aims is also to simulate a semi-active suspension modeled with magnetorheological (MR) dampers. Basing on the outcomes of trials, elements such as drives will be selected, and the final result of work will be a 3D CAD model of the mobile platform (Cholewa, 2023).

2. Conceptual design of a mobile platform

2.1. Assumptions and designed kinematics

Developing a solution for the design of a mobile platform requires first to formulate working conditions and assumed functionalities including flexibility of application and movement in unstructured terrains (focusing on cultivated farmlands), minimizing the uncontrolled vertical movements and vibrations, supporting work in agriculture like harvesting/monitoring of crops and gathering samples.

After checking the state of the art and considerations of applications, the consecutive structural assumptions for the mobile platform and its suspension have been formulated:

- Mobile platform as a 4-wheeled vehicle,
- Independent double wishbone suspension system with a motor for each wheel,
- Semi-active suspension based on a shock absorber as a spring with an MR damper,
- Changing the dimension of track of wheels by at least 30%,
- Compact design, platform size fitting in 1 m^3 to simplify its transport,
- Lightweight design – less than 65 kg, platform mass with assumed additional equipment.

Taking into account all the requirements and assumptions, the design of the suspension system as a kinematic scheme (Fig. 2) has been created. An independent double wishbone suspension (type of earlier presented multi-rocker suspension) is chosen to obtain a light and compact structure allowing one to apply wide tires and to lower the platform center of mass to increase stability. The turning of the vehicle is going to be realized by varying the rotational speed of motors installed in each wheel.

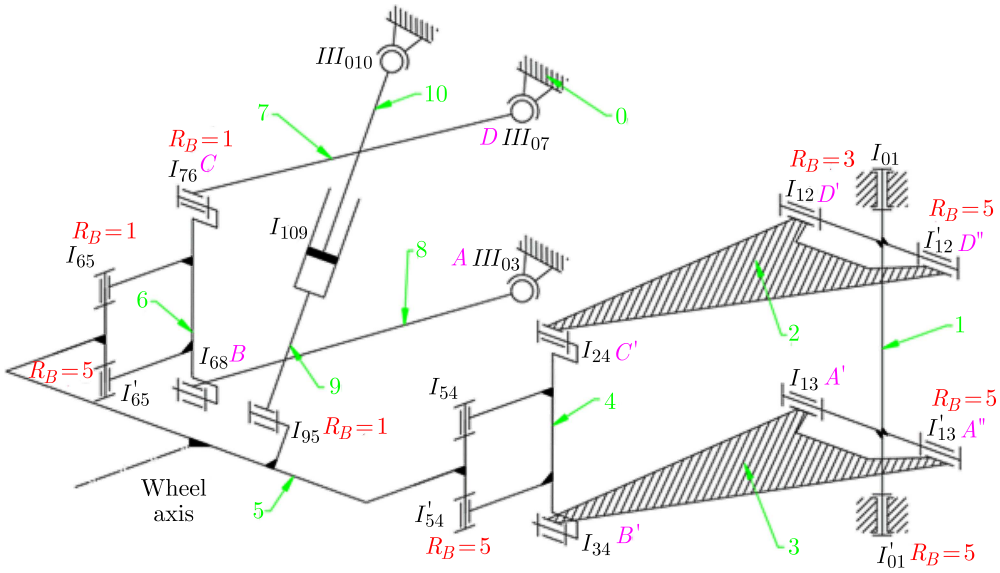


Fig. 2. Kinematic scheme of the proposed suspension mechanism (view for one wheel)

The theoretical mobility has been calculated according to the well-known structural equation as

$$W_T = 6(n-1) - 5p_1 - 4p_2 - 3p_3 = 6 \cdot 10 - 5 \cdot 16 - 3 \cdot 3 = -29 \tag{2.1}$$

and the real mobility as

$$W_R = W_T - R_B = -29 + (5 \cdot 5) + 3 + 1 + 1 + 1 = 2 \tag{2.2}$$

where: n – number of mechanism elements, p_i – number of kinematic pairs of particular i class, R_B – number of passive (excessive) constraints. The passive constraints are in this case connected with multiple parallel first class kinematic pairs (5 of these) and building plane 4-bar mechanisms in space.

The shock absorber consists of a spring and damper that are modeled together as elements 9, 10 in Fig. 2. Moreover, it is assumed that the platform is going to be equipped with a manipulator, camera and set of sensors, so in an attempt to minimize uncontrolled vibrations, vertical movements of the main body and consequently errors in its positioning, a semi-active suspension with magnetorheological dampers is proposed. Their application advantageously provides fast responsiveness, relatively low weight, as well as simple construction, since viscosity of the damper fluid and, consequently, the damping coefficient and force can be controlled in real time (Sarami, 2009). To realize this, an MR damper LORD RD-1005-3 has been selected in accordance with the planned dimensions and predicted movements. Its stroke is 35 mm and the damping coefficient ranges from 1.3 Ns/mm to 2.9 Ns/mm (Mohd Yamin *et al.*, 2022).

In addition, due to various types of terrains, the desire to avoid obstacles and to adjust the device to driving between irregular rows of plants, it was decided to include in the vehicle the ability to change its track of wheels. Therefore, the proposed suspension system is a mechanism with the mobility of two. The first one is the vertical wheel movement which enables overcoming obstacles and minimizing vertical chassis movements, whereas the second mobility is an additional rotation of the suspension around the vertical axis (element 1 in Fig. 2), resulting in altering the track of wheels with a simultaneous change of the wheelbase. This mechanism is intended to increase the flexibility of vehicle applications.

Overall, the proposed double-wishbone suspension is characterized by connecting the chassis and the wheel with two wishbones, thus the entire system takes the form of a 4-bar mechanism (Fig. 2). The wishbones are elements No. 2, 3 and the same function is fulfilled by elements No. 7, 8. The duplication of parts is required for proper operation of the system for changing the track of wheels.

2.2. Determined dimensions of the device

Considering the developed kinematics, including the suspension and system for changing the track of wheels, as well as the defined assumptions/requirements, the dimensions are determined for the platform (Fig. 3a) and suspension (Fig. 3b). A couple of additional assumptions formulated particularly for the suspension dimensions include wheel maximum vertical lifting 55 mm and maximum lowering 45 mm, wheel tilt its max/min wheel vertical position: $-4°/5°$. Rotational pairs A and D (Fig. 2 – mounting points of the lower and upper wishbones) of the 4-bar mechanism ABCD have common vertical axis of rotation to simplify the design for changing the track of wheels.

The determined dimensions are proposed also in view of the desired change of track of wheels by at least 30%. The principle of work for this mechanism is presented in Fig. 4. It proves that a 70° rotation of the suspension results in changing the track of wheels by more than 230 mm, which for the maximum width of 690 mm is a change by over 33%. During the process, the wheel all the time stays parallel to the chassis which enables a continuous change during the ride and

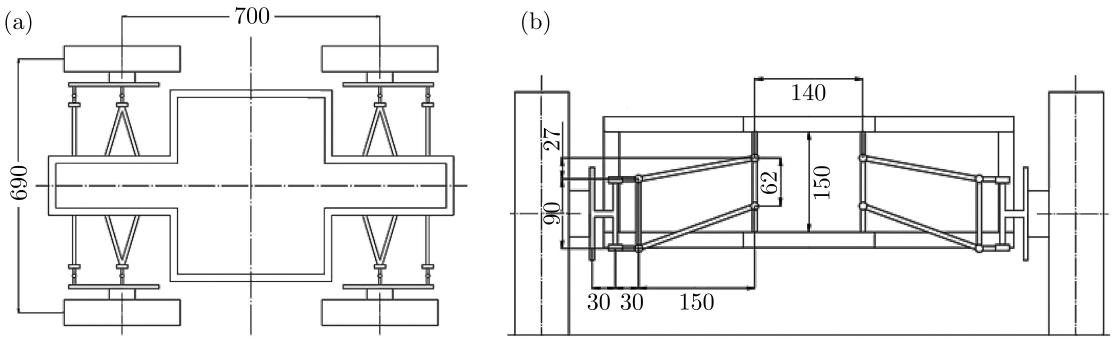


Fig. 3. Designed dimensions (in mm) indicated in: (a) an overall scheme of the mobile platform, (b) a scheme with details of the suspension system

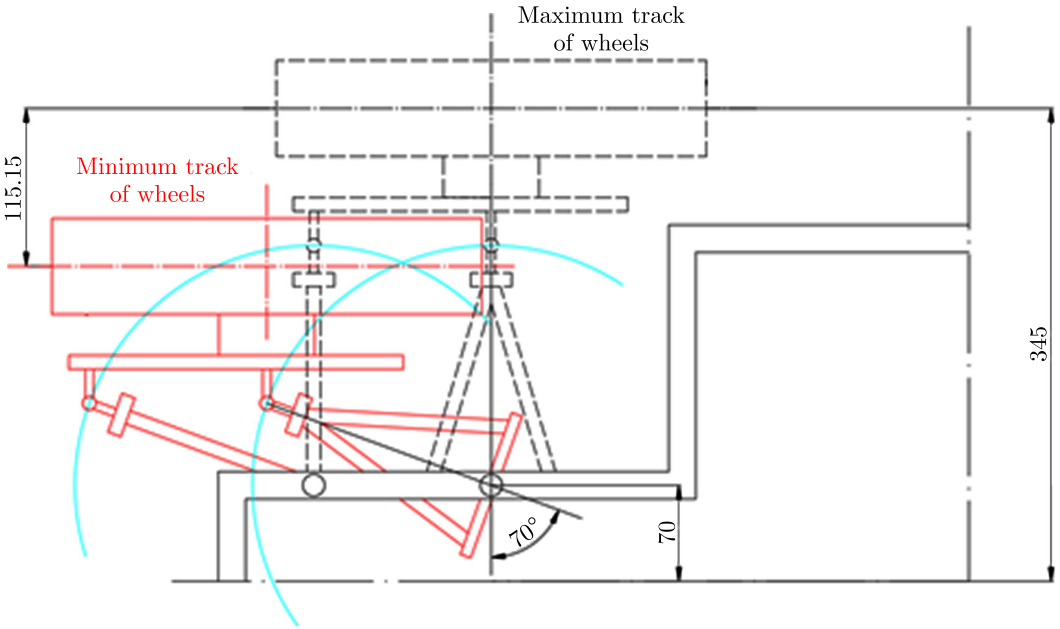


Fig. 4. Scheme presenting the working principle of the mechanism changing the track of wheels

setting any dimension in the obtainable range. It is also crucial that this process does not change length of the shock absorber and therefore does not have a negative impact on its operation. In order to ensure this, the shock absorber is mounted in accordance with the wishbone. Therefore during the process, one end of the shock absorber (Fig. 2 – pair III_{010}) is exactly in the axis of rotation (Fig. 2 vertical axis of rotation for pairs A, D), while the other end (Fig. 2 – pair I_{95}) moves in a circle (along the right arch in Fig. 4).

3. Numerical studies

3.1. Built numerical model

In order to perform test trials of the designed mobile platform, a spatial multibody dynamic numerical model with a semi-active suspension and system for changing the track of wheels has been built in Adams simulation software (Fig. 5). The model is based on elements with simple shapes, but all crucial features are reproduced according to the developed device kinematics/dimensions, as well as to the approximated mass of elements and mass distribution (possibly close to reality). Basing on diagrams and sketches, an initial 3D model of the vehicle

was created, assuming the frame to be made of aluminum profiles and sheets. The mass of the preliminary 3D model was 28 kg. The weight of additional equipment in the form of drives, batteries, sensors, cameras, electronics, transported load and/or manipulator was assumed to be another 25 kg. The final mass of the modeled vehicle (64 kg) includes also a 1.2 safety factor. All the additional mass is placed in the centre of the vehicle (as yellow cylinder in Fig. 5).

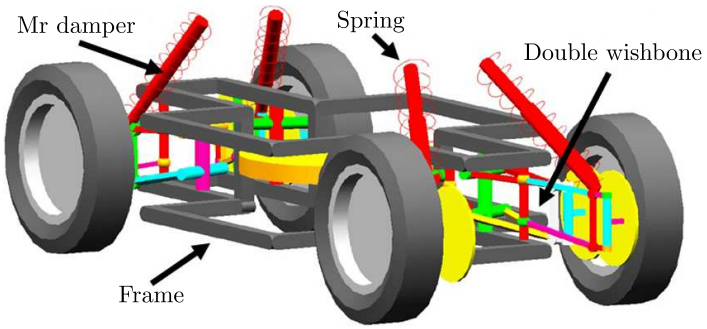


Fig. 5. Multibody dynamic numerical model of the platform built in Adams software

The model in Fig. 5 includes also visual presentation of the spring and damper, but the simulated complex force of shock absorber F_{SA}, consisting of spring F_S and MR damper F_D, is calculated in the model according to the below equation

$$F_{SA} = F_S + F_D = (F_P + k\Delta l_S) + bV_D \tag{3.1}$$

where: F_P – spring preload force, k – spring stiffness coefficient, Δl_S – spring length change, b – damper damping coefficient, V_D – velocity of damper length change.

Since MR dampers are applied, in some simulations the damping coefficient is changed and controlled. Therefore, to obtain an adaptable damping force, a control system minimizing the vertical velocity of chassis, is also implemented in the model. Specifically, for each damper, the vertical velocity of the frame corner is measured and minimised. In addition, b is limited to values between minimum and maximum (1.35 Ns/mm to 2.9 Ns/mm) achieved by the assumed MR damper. At this point of research, a proportional regulator (P regulator) is applied to calculate the damping coefficient b as

$$b = K_P(0 - V_y) \tag{3.2}$$

where: K_P – proportional gain of the regulator, V_y – vertical velocity of the chosen point (corner) of chassis.

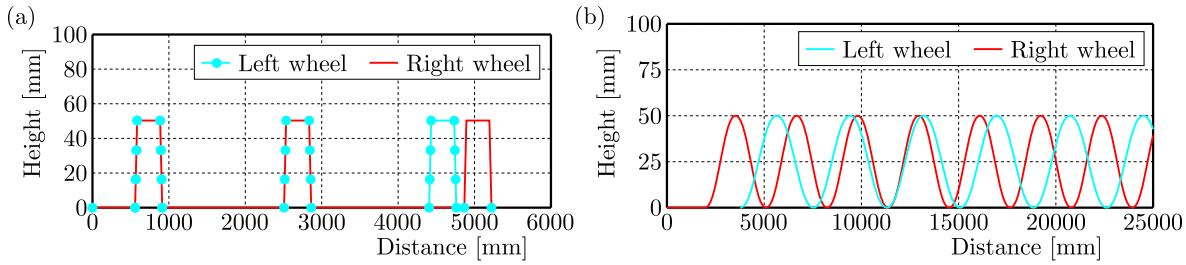


Fig. 6. Profile of road: (a) with step obstacles, (b) with sinusoidal obstacles

In order to perform the planned simulations, the numerical model is supplemented with definitions (in additional files for usage with Adams) of 2 road profiles: a) road with step obstacles (Fig. 6a) – 3 obstacles with height 50 mm, length 300 mm, third obstacle is shifted for right and

left wheel; b) road with sinusoidal obstacles (Fig. 6b) – height 50 mm, length about 1600 mm, in this case the obstacles for right and left wheels are shifted unevenly.

For all types of roads the tire features and tire/road contact conditions are assumed. Basing on literature (Blundell and Harty, 2004) a tire Fial model is applied, with parameters specified in another file for Adams – unloaded tire radius ($R_1 = 157.5$ mm), width ($d = 70$ mm), vertical stiffness ($k_z = 24$ N/mm), vertical damping ($\zeta = 0.6$ Ns/mm), rolling resistance ($C_r = 27$ mm), longitudinal force with respect to the longitudinal slip ratio ($C_S = 300$ N), lateral force connected with the slip angle ($C_a = 22$ N/rad), coefficients of friction at zero slip ($U_{min} = 0.65$) and when the tire is sliding ($U_{max} = 0.95$).

3.2. Performed simulations

For the built numerical model, with the defined road profiles and tire/road contact conditions, a couple of simulations was performed. The experimental modes with a particular set and measured parameters are specified in Table 1, where: k, b are the set values of spring and damper coefficients, K_P – proportional gain of the regulator, F_P – spring preload force, a_V – vehicle acceleration, V_{max} – maximum device velocity kept after 0.5 s of acceleration, d_b, V_b, a_b – measured body displacement, velocity and acceleration, F_j, F_D – measured forces in joints and dampers, T_M – torques of motors.

Table 1. Planned experiments with specified model/simulation conditions and collected variables

No.	Experimental mode	Set values of variables		Collected variables
1	Road with step obstacles Constant damping coefficient	$b = 2.1$ Ns/mm	¢$k = 1750$ N/mm $F_P = 156$ N Vehicle's $V_{max} = 1.5$ m/s Vehicle's $a_V = 3$ m/s^2	¢d_b, V_b, α_b F_j, F_D, T_M
2	Road with step obstacles Variable damping coefficient	¢b – controlled $K_P = 10^5$		
3	Road with sinusoidal obstacles Variable damping coefficient			

The simulations were conducted with a 100 Hz frequency, partially in order to obtain a realistic operation of MR dampers – time of damping coefficient change is therefore limited to 0.01 s. The spring stiffness ($k = 1750$ N/mm) and its preload force ($F_P = 156$ N) were selected on the basis of intermediate simulations. These values enabled the vehicle to maintain a stable vertical position of the chassis (without fluctuations) in the absence of external loads, and enabled the spring to deform freely when overcoming the obstacles. An average damping coefficient $b = 2.1$ Ns/mm was selected for the MR damper in cases when it was kept as constant.

Simulations concerning the changing of the track of wheels during ride and overcoming obstacles with various values of track of wheels were also performed in Adams. Details are not reported in this paper, but the obtained results (e.g. max. value of torque for the mechanism changing the track of wheels was 33 Nm) proved the feasibility of the system as it does not collide with the chassis and allows selecting linear motors shown in Section 5.

4. Results of simulations

According to the details presented in Table 1, the 3 planned simulations were carried out in Adams by using the built numerical model. The results include, among others, the acceleration of the body center of mass. The simulations allowed also for the assessment of the correct operation of the built P regulator, which controls the magnetorheological damper. In addition, taking as

criteria the minimization of body vibrations and the maximization of safety, understood as the vehicle ability to maintain its wheels in contact with terrain, the ground reaction forces and time of possible losses of contact between the wheels and ground were verified. A part of the obtained simulation results has been presented in Figs. 7-10.

For the first simulation of the vehicle movement on the road with step obstacles and with a constant damping coefficient, the platform centre of mass accelerations in 3 axes are measured and presented in Fig. 7. Overall, the mean value of vertical acceleration as a_{RMS} (root mean square) is equal to 1.81 m/s^2 and the maximum value does not exceed 9.8 m/s^2 (Fig. 7). As can be seen in the plot, the road profile (Fig. 5a) is mapped in the results of vertical acceleration. At the road beginning there is the first obstacle, symmetric for right and left wheels. The vehicle is overcoming it during accelerating to max. velocity $V_{max} = 1.5$ m/s. This constant speed is kept, and after 2.5 m there is another symmetric obstacle. At the end, there are 2 asymmetric obstacles.

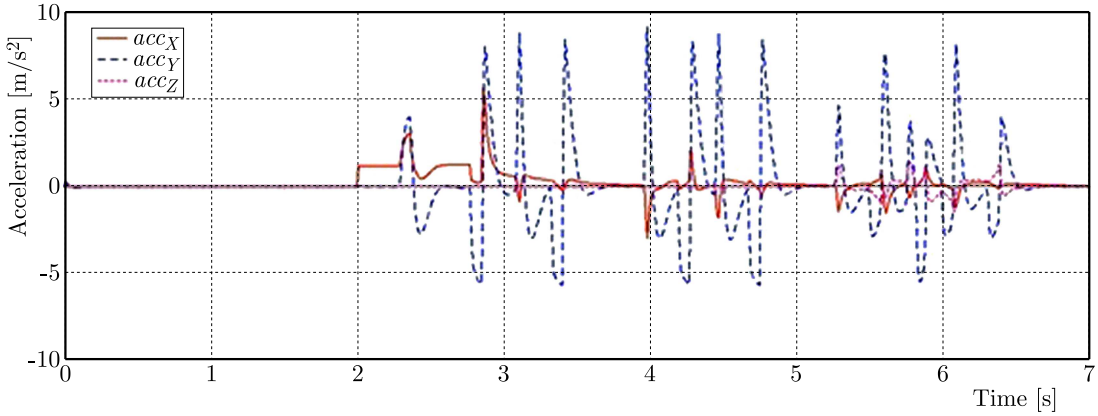


Fig. 7. Accelerations of the vehicle chassis centre of mass during the first simulation of riding on a road with step obstacles with a constant value of the damping coefficient

During the second simulation, the correct operation of the damper regulator was verified by measuring the damping coefficient value while overcoming the obstacles 2.2 s-6.7 s (Fig. 8). It can be noticed that the controller is working and the value of the damping coefficient is adjusted in real time to minimize the set error, i.e. to minimise the body vertical speed. However, the coefficient takes only the limit values in the assumed allowable range (1.35 Ns/mm to 2.9 Ns/mm), indicating the need for a more advanced control system.

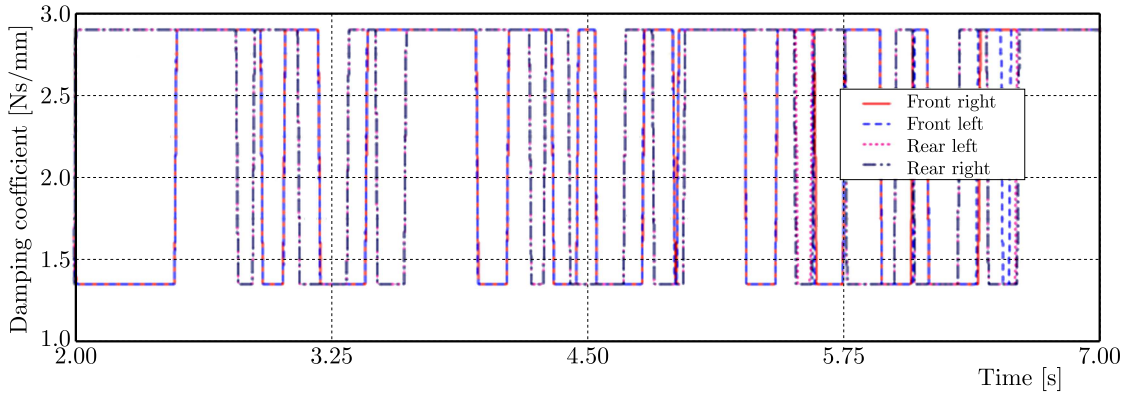


Fig. 8. Result of simulation 2 with values of controlled individually for each wheel damping coefficient while driving over the 3 step obstacles

Furthermore, in this case, the chassis accelerations were also checked and compared with the results of previous simulation. In addition, interesting values of accelerations obtained during the 3 experiments have been gathered and presented in Table 2, where a_{max} is the measured maximum value of vertical acceleration, a_{RMS} – calculated root mean square of the vertical acceleration, a_{mag} – magnitude value of accelerations calculated as a square root of the sum of squares of a_{RMS} for the acceleration in each axis.

As can be seen in Table 2, a small improvement in each of the acceleration parameters is achieved. For instance, the maximum values decreased from 9.8 m/s^2 to 9.7 m/s^2 for the case with the semi-active suspension and, similarly, the a_{RMS} dropped from 1.81 m/s^2 to 1.77 m/s^2. It can be concluded that the semi-active suspension with MR dampers proved to be useful, and even better results are expected when control of the damping coefficient is improved. However, it should be noted that these cases with very steep obstacles enabled above all evaluation of the operation of the semi-active suspension system and regulator in exceptionally unfavorable terrain conditions, which resulted in relatively high values of accelerations.

Table 2. Numerical results of chassis accelerations for each simulation

No.	Type of simulation	Acceleration of chassis in [m/s^2]		
		Vertical (in y axis)		Magnitude
		a_{max}	a_{RMS}	a_{mag}
1	Step obstacles – Constant damping coefficient	9.8	1.81	1.85
2	Step obstacles – Variable damping coefficient	9.7	1.77	1.81
3	Sinusoidal obstacles – Variable damping coefficient	0.39	0.06	0.08

Furthermore, the vehicle must be able to move in unstructured terrain, so it is important that it maintains contact with the ground even when overcoming rough obstacles. For this reason, the vertical ground reaction forces are verified for each wheel and presented in Fig. 9 (No. 2). These allowed one to check when the wheels are separated from the ground, since it corresponds to a zero force value and entails temporary inability of the wheel to generate the traction force. Duration of losses of contact is checked for each wheel: front left (0.18 s), front right (0.19 s), rear left (0.19 s), rear right (0.19 s). The sum of time when at least one wheel is not in contact with the road amounts to about 0.5s. This indicates decent vehicle behaviour during overcoming step obstacles, also in view of the fact that the design assumes a motor in each wheel. Therefore, even if one of the wheels loses traction, then the other three are still in contact with the ground (e.g. visible in Fig. 9 while overcoming the asymmetric obstacle at about 6 s).

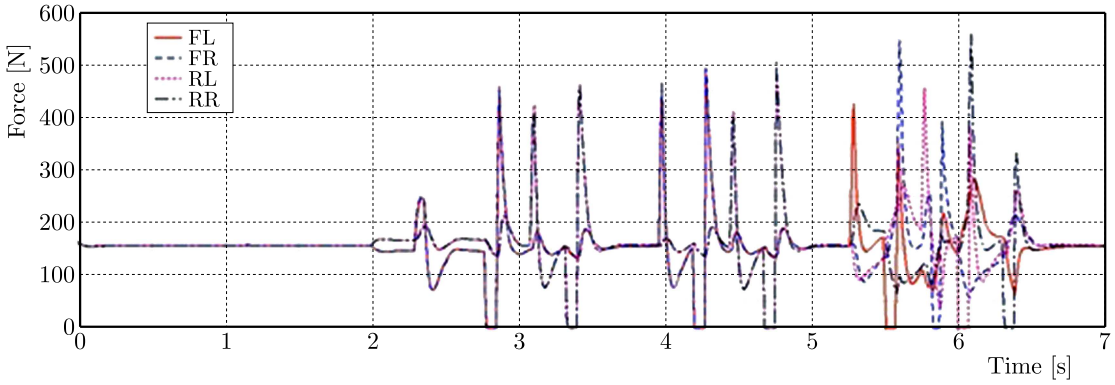


Fig. 9. Vertical ground reaction force on each wheel (FL, FR – front left and right, RL, RR – rear left and right) for simulation 2 while driving over the 3 step obstacles with the controlled damping coefficient

Additionally, thanks to the advantageous work of the suspension system and its components, for most of the time, the left and right wheels are equally loaded (Fig. 9), which enables achieving

favorable contact conditions for traction force generation. Moreover, when stationary, the vertical force is about 157 N (Fig. 9) on each wheel which sums to the effect of the whole vehicle mass (64 kg).

Simulated is also movement on the road with relatively high, but not steep sinusoidal obstacles (Table 1 – No. 3). This way, a realistic terrain and work conditions for the device have been reproduced. The simulations results are among others the motor torques for each wheel (Fig. 10) and forces in joints. For the case of constant vehicle velocity, the average wheel torque is about 3.75 Nm (Fig. 10) and the maximum values do not exceed 7 Nm (omitting the results in initial seconds probably distorted by excessive slipping). In the case of this less demanding road with sinusoidal obstacles, which do not appear suddenly, but their height increases gradually, the vertical accelerations are minor and much smaller than those in previous simulations (Table 2). The maximum value does not exceed 0.4 m/s^2 and the RMS is 0.06 m/s^2. What is more, in this case, it was possible to continuously keep all wheels in contact with the ground, which proves the feasibility of the designed semi-active suspension system.

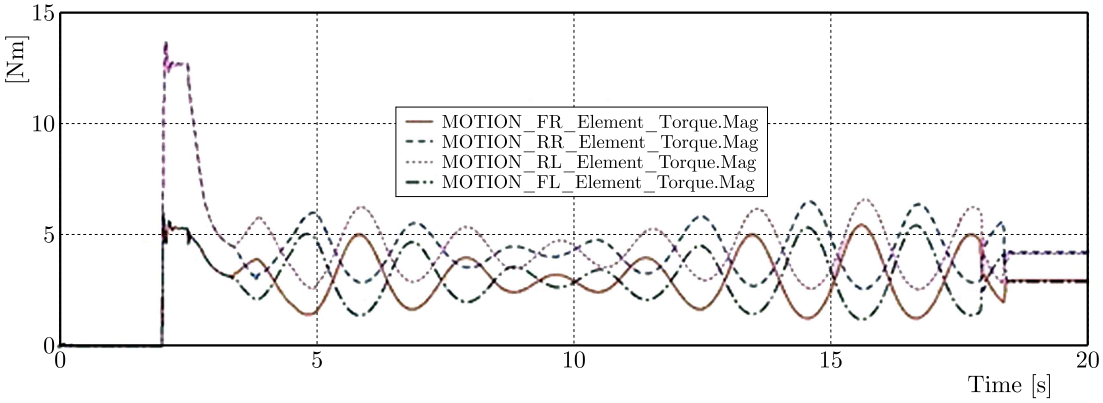


Fig. 10. Torques in wheels for simulation No. 3, moving on a road with sinusoidal obstacles while the damping coefficient is controlled and changed

All the results proved that the semi-active suspension system works properly and that the designed platform has the ability to move in uneven terrains, as well as is characterised by decent contact with the ground, overall good vertical stability, acceptable acceleration amplitude, frequency and vibration damping.

5. Built CAD model

A 3D CAD model of the designed mobile platform has been created in Autodesk Inventor software with particular emphasis put on presenting one wheel suspension (Fig. 11a) and the whole device with selected elements (Fig. 11b). It serves as an illustrative presentation of the vehicle concept and enables one to verify that no collisions occur between the main components.

The model was made taking into account the selected, on basis of simulation results, parts like: wheel drive as BDLC motor Dunkermotoren BG45X15SI with gearhead PLG40LB; clutch ROTEX KTR 98ShA,14 and electric linear cylinder FESTO EPCS BS-32-75-3P for changing the track of wheels, etc. Individual elements were also checked for their strength/stress features to select proper standardized parts that can be used in the final design and become a basis for building a prototype.

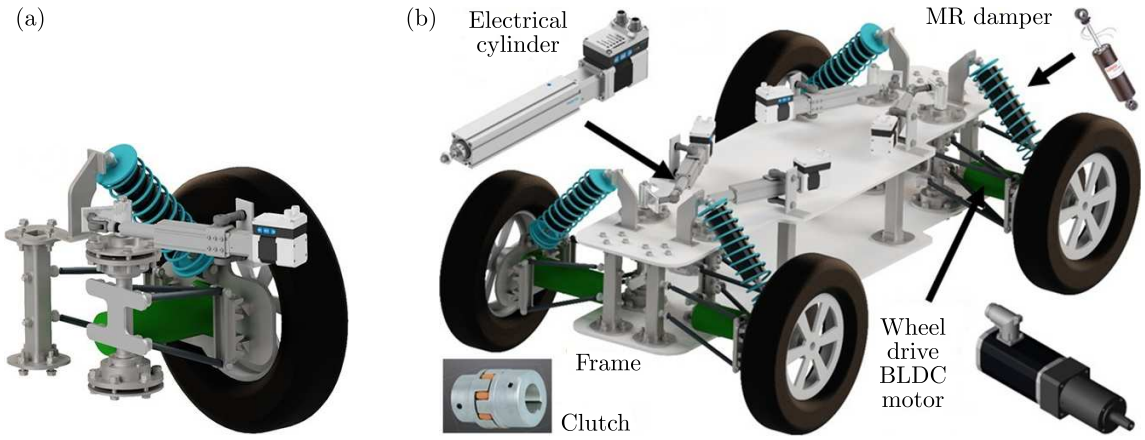


Fig. 11. A spatial 3D model built in CAD software (Inventor): (a) details of the suspension system for one wheel, (b) view of the whole vehicle with a part of selected elements

6. Conclusions

The paper is focused on the development and design of a mobile wheeled platform able to move in an uneven terrain and particularly intended for a supporting work in agriculture. The research is conducted partially, in view of the increasing demand for food production and the desire to achieve smart and precise farming with the usage of autonomous machines.

The exact planned application of the platform is combining it with a manipulator (like Kinova Gen3 Lite, or similar in workspace, payload and weight), as well as equipping it with a camera and various sensors, e.g. lidar, distance sensors, inertial IMU sensors. This way an autonomous multi-purpose mobile device could be obtained. An independent stabilization system for the camera or other elements would not be desirable, since it would limit the vehicle suitable cooperating equipment to special models. Therefore, the concept of minimizing the uncontrolled platform movements and vibrations is studied. Strictly agricultural applications could possibly focus only on providing the best ground contact conditions – the least loses of contact, omitting the overall vibrations problems, since no comfort for the operator would be necessary. This could be obtained by maximising spring stiffness and damping coefficients. However, the intended operation of the platform is wider, concerning various terrains and applications.

For these reasons, it was decided to design a 4-wheeled platform and emphasis was placed on the concept (application of MR dampers) and kinematic structure of the suspension system to provide flexibility for the vehicle mobility, as well as operation in various ground conditions and farmlands in agricultural applications. Constructional and functional assumptions were specified determining many features of the designed device. A semi-active independent double wishbone suspension mechanism with mobility of 2 was proposed, enabling not only the vertical movement of the wheel, but also a supplementary rotation of the suspension system around the vertical axis, realizing the change of track of wheels. These will allow the vehicle to avoid too large obstacles, increase its stability and elasticity of application.

Then, the vehicles dimensions were specified and in the next step, a preliminary design of the device was elaborated as a 3D dynamic numerical model of the platform (64 kg) with a 4-wheel suspension system formulated in a multibody simulation program (Adams). In order to carry out the experiments on the semi-active suspension with an adaptable damping force, a control system has also been implemented in the model. Semi-active suspensions and MR dampers have been modelled and simulated in other research like (Klockiewicz and Ślaski, 2023), where friction, hysteresis and actuation delay have been taken into consideration. However, this paper is focused on the design of the wheeled platform for uneven terrain in which the semi-active

suspension system is a part of the developed idea. Therefore, at this stage of research, the MR dampers are modelled as simple elements and the control strategy for changing, in real time, the damping coefficient is based on a P controller, minimizing the body vertical velocity.

While simulating the driving on uneven terrain, the obtained vehicle parameters were verified and compared. The occurrences and durations of possible losses of contact between the individual wheels and the ground surface were also checked. The three presented experimental modes included different settings of road with obstacles, as well as cases with constant damping characteristics and with active damping regulation. Particularly, in order to evaluate the suspension system capability, plots and numerical results were derived from simulations. The vehicle kinematics and dynamics were examined, and the results were used to perform strength calculations of elements, as well as to select standardized parts and motors.

The first simulation evaluated the vehicle behaviour without the semi-active suspension in exceptionally unfavourable terrain conditions, with step obstacles. A satisfactory level of uncontrolled vertical movements was obtained. Furthermore, for comparison, the same terrain conditions were simulated with application of the controlled damping coefficient. Due to straightforward control (P regulator) the damping coefficient assumed only max/min values. Despite this fact, a slight but significant improvement in the uncontrolled vertical movements level was observed (maximum acceleration dropped from $9.8\,\mathrm{m/s^2}$ to $9.7\,\mathrm{m/s^2}$ and its RMS from $1.81\,\mathrm{m/s^2}$ to $1.77\,\mathrm{m/s^2}$). More realistic terrain with sinusoidal obstacles was also simulated. The vertical acceleration was lower than $0.4\,\mathrm{m/s^2}$ and the average wheels torque was about 3.75 Nm. Vertical ground reaction forces were also checked and, consequently, the time when the wheels were separated from the ground. For the vehicle, favourable contact conditions for traction force generation in unstructured terrain were obtained. A decent level was observed when overcoming rough obstacles and continuous contact with the ground was kept for all wheels for sinusoidal obstacles.

To conclude, the results proved feasibility of the device and that the designed semi-active suspension enabled to decrease the main body undesired vertical movement and acceleration, as well as to possibly adjust the vehicle to various and individual cases of applications in uneven/unstructured terrains and in agriculture. The features enabling this, are the proposed two main advantages: the possibility of changing the track of wheels (functionality distinguishing the vehicle from many currently available), as well as application of MR dampers to obtain a semi-active suspension.

As the final result, a conceptual design of the vehicle has been proposed by selecting motors, constructional standardized parts, as well as some elements of electric and electronic systems. The device 3D CAD model has also been built, providing a good basis for constructing a research prototype in the future. Completing it would require suitable sensors to be selected, which combined with appropriate control would allow the device to operate autonomously. Focus should be placed on application of technologies and elements such as lidar to create maps of the environment and to detect objects, cameras for image recognition and analysis, but also ultrasonic and inertial IMU sensors, which would allow for proper control and minimizing body movements when overcoming unevenness. Moreover, the CAD model could still be used to improve the numerical model and perform simulations checking various tire/terrain contact conditions, as well as a more advanced MR damper model and control strategies (e.g. PID controller).

References

1. Ackerman E., 2015, *Bosch's Giant Robot Can Punch Weeds to Death, IEEE Spectrum, 0-1*, https://spectrum.ieee.org/automaton/robotics/industrial-robots/bosch-deepfield-robotics-weed-control

2. Blundell M., Harty D., 2004, Active systems [In:] *The Multibody Systems Approach to Vehicle Dynamics*, Elsevier, New York, 441-451
3. Bruzzone L., Quaglia G., 2012, Review article: locomotion systems for ground mobile robots in unstructured environments, *Mechanical Sciences*, **3**, 49-62
4. Ceccarelli M., Cafolla D., Carbone G., Russo M., Ferrante F., *et al.*, 2017, Heritage bot service robot assisting in cultural heritage, *First IEEE International Conference on Robotic Computing*, 440-445
5. Cholewa K., 2023, Design of a mobile platform for agricultural purposes (in Polish), Master Thesis, Faculty of Mechanical Engineering, Wroclaw University of Science and Technology, Wroclaw, Poland
6. Chołodowski J., 2023, Modelling and experimental identification of spring-damping properties of the off-road vehicle rubber tracks, rubber belts, and rubber-bushed tracks subjected to flexural vibrations, *Journal of Terramechanics*, **110**, 101-122
7. Ecorobotix, 2019, Technology for environment: an innovative, autonomous and economical machine, https://www.ecorobotix.com/wpcontent/uploads/2017/02/ECOX_FlyerPres18-EN-1_RVB-1.pdf
8. Fischer D., Isermann R., 2003, Mechatronics semi-active and active vehicle suspensions, *Control Engineering Practice*, **12**, 11, 1353-1367
9. Garimella S.S, Revzen S., 2021, *Dandelion-Picking Legged Robot*, arXiv:2112.05383
10. Harrington B.D., Voorhees C., 2004, The challenges of designing the rocker-bogie suspension for the Mars exploration rover, *Proceedings of the 37th Aerospace Mechanisms Symposium*, 185-195
11. Husti I., 2019, Possibilities of using robots in agriculture, *Hungarian Agricultural Engineering*, **35**, 59-67
12. Klockiewicz Z., Ślaski G., 2023, Comparison of vehicle suspension dynamic responses for simplified and advanced adjustable damper models with friction, hysteresis and actuation delay for different comfort-oriented control strategies, *Acta Mechanica et Automatica*, **17**, 1, Special Issue "Machine Modeling and Simulations 2022"
13. Mohd Yamin A., Ab Talib M.H., Mat Darus I.Z., Mohd Nor Nur S., 2022, Magneto-rheological (MR) damper – parametric modelling and experimental validation for LORD RD 8040-1, *Jurnal Teknologi*, **84**, 2, 27-34
14. Niu J., Wang H., Shi H., Pop N., Li D., Li S., Wu S., 2018, Study on structural modeling and kinematics analysis of a novel wheel-legged rescue robot, *International Journal of Advanced Robotic Systems*, **15**, 1
15. Olinski M., Ziemba J., 2014, Hybrid quadruped robot – mechanical design and gait modelling, [In:] *New Advances in Mechanisms, Transmissions and Applications, Mechanisms and Machine Science*, Petuya V., Pinto C., Lovasz E.C. (Eds.), **17**, 183-190
16. Raibert M.H., Blankespoor K., Nelson G.M., Playter R., 2008, BigDog, the rough-terrain quadruped robot, *IFAC Proceedings Volumes*, **41**, 2, 10822-10825
17. Raper R.L., 2004, *Agricultural Traffic Impacts on Soil*, USDA-ARS-National Soil Dynamics Laboratory
18. Roldán J.J., del Cerro J., Garzón-Ramos D., Garcia-Aunon P., Garzón M., de León J., Barrientos A., 2018, Robots in agriculture: state of art and practical experiences, [In] *Service Robots*, A.J.R. Neves (Edit.)
19. Sarami S., 2009, *Developement and Evaluation of a Semi-Active Suspension System for Full Suspension Tractors*, Berlin
20. Shah R., Ozcelik S., Challoo R., 2012, Design of a highly maneuverable mobile robot, *Procedia Computer Science*, **12**, 170-175

21. Shamshiri R.R., Weltzien C., Hameed I,A., Yule I.J., Grift T.E., *et al.*, 2018, Research and development in agricultural robotics: A perspective of digital farming, *International Journal of Agricultural and Biological Engineering*, **11**, 4, 1-14

22. Sperzyński P., Szrek J., Muraszkowski A., 2018, Simulation research of a mobile robot walking on stairs (in Polish), *Modelowanie Inżynierskie*, **67**

Manuscript received November 14, 2023; accepted for print December 4, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 293-306, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185614

INTERNAL HEAT SOURCES IN LARGE STRAIN THERMO-ELASTO-PLASTICITY – THEORY AND FINITE ELEMENT SIMULATIONS[1]

Balbina Wcisło
Cracow University of Technology, Chair of Computational Engineering, Cracow, Poland
e-mail: balbina.wcislo@pk.edu.pl

Marzena Mucha
Cracow University of Technology, Chair of Computational Engineering, Cracow, Poland, and
TU Dortmund University, Institute of Mechanics, Department of Mechanical Engineering, Dortmund, Germany

Jerzy Pamin
Cracow University of Technology, Chair of Computational Engineering, Cracow, Poland

The paper deals with theoretical description and numerical simulations of internal sources of heating/cooling in large strain thermo-elasticity and thermo-elasto-plasticity. The attention is paid to metallic materials which undergo cooling in the elastic range and heating during plastic yielding. Theoretical description can be derived from thermodynamic considerations based on the first and second laws of thermodynamics and assumed forms of the Helmholtz free energy. Numerical simulations within the Finite Element Method are performed for a uniaxial tension test and elongation of a dogbone-shape sample. For the latter specimen, a comparison with experimental results is performed, and good agreement is obtained.

Keywords: thermo-mechanics, thermo-elastic cooling, plastic dissipation, AceGen/FEM

1. Introduction

In non-isothermal conditions, the response of an elastic-plastic material is usually described by using two governing equations: the balance of linear momentum and the balance of energy. The equations can be coupled, the thermal field can influence the mechanical one, and conversely. In particular, a change of temperature causes thermal expansion of the material and influences material parameters, both the mechanical (e.g. Young's modulus or initial yield threshold) and thermal ones (e.g. heat conductivity or heat capacity coefficients). It should be noted that the decreasing value of the yield limit can lead to softening of the material and to strain localization, see e.g. Duszek and Perzyna (1991), which additionally complicates the constitutive description, since the response can then be incorrectly represented by a classical (local) constitutive theory.

On the other hand, large deformation of a sample can influence heat transfer in the material and can involve internal sources of heating or cooling. In particular, in the elastic regime, a thermo-elastic coupling, related to the Gough-Joule effect, can be observed. Very often, the phenomenon is related to so-called entropic materials like rubber or polymers, see Holzapfel (2000), which undergo heating during rapid stretching. However, it is observed in experiments that metals also show thermo-elastic coupling in the elastic regime but in a different way – their temperature decreases during elongation (Mucha *et al.*, 2023). A second example of the internal heat source in a material is plastic deformation which causes heat generation due to energy

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

dissipation. This matter is widely analysed in literature in the context of experiments, theory and simulations, for example (Taylor and Quinney, 1934; Ristinmaa *et al.*, 2007; Oliferuk *et al.*, 2013; Rose and Menzel, 2021; Musiał *et al.*, 2022).

The subject of this paper is the analysis of fully thermo-mechanically coupled elasto--plasticity, with special attention paid to internal heat sources observed in elastic and plastic regimes for materials undergoing large strains. The formulation is derived in a thermodynamically consistent manner on the basis of the Helmholtz free energy which consists of the following parts: reversible (including elasticity and thermal expansion), purely thermal and plastic. The internal heat sources related to the thermo-elastic coupling and to energy dissipation during yielding of the material can be derived from the first and second laws of thermodynamics according to, for example, Ristinmaa *et al.* (2007), Oppermann *et al.* (2022). A simplified specification of the plastic dissipation heat source, i.e. the application of the Taylor-Quinney factor, is also used for comparison.

The attention in the paper is focused on metals/alloys, and isotropy of the material is assumed. The thermo-mechanical coupling involves thermal expansion, internal heat sources in elasticity and plasticity, Fourier's law in the deformed configuration and thermal softening in plasticity. It does not include the dependence of material parameters on temperature. The description of plasticity in this work is based on Huber-Mises-Hencky theory with associative flow rule without viscous terms.

The presented model is implemented within the Finite Element Method (FEM) (Zienkiewicz *et al.*, 2005), and two specimens are tested: a cube and a dogbone-shape sample, both in tension. Simulations for the latter specimen are compared with the results of laboratory experiments presented in Mucha *et al.* (2023). Although in the experiments by (Mucha *et al.*, 2023) propagative instabilities are observed (i.e. the Lueders bands and Portevin-Le Chatelier effect), which are not reproduced by the presented model, the comparison of reactions and temperature is performed in a general way, and a good agreement is obtained.

The thermo-elastic and thermo-elasto-plastic models are described using different stress and strain measures. This approach is applied here intentionally: thermo-elasticity can easily be described using quantities related to the reference configuration, and this allows for efficient implementation within FEM. In turn, thermo-elasto-plasticity at large strain is based on plasticity description developed in (Simo and Hughes, 1998) and (Simo and Miehe, 1992), which involves spatial quantities, and can be effectively applied in the chosen FEM software (Korelc and Wriggers, 2016).

The large strain elasto-plasticity theory, which takes into account the full thermo-mechanical coupling, leads to strongly non-linear problems. Numerical simulations using such a material model require an advanced solution approach. In this work, finite element procedures are implemented within symbolic-numerical package AceGen for Wolfram Mathematica (Korelc and Wriggers, 2016). The most significant feature of the package is automatic differentiation which allows for computation of the tangent operator in the Newton-Raphson procedure as a derivative of the residual vector with respect to the vector of unknowns.

The paper is laid out as follows. In Section 2, basic quantities and thermodynamic laws are presented, which are further used in the specific models. In Section 3, the thermo-elastic model is presented with special attention paid to thermo-elastic coupling and its numerical simulation. The thermo-elasto-plastic model is presented, in turn, in Section 4. Two variants of the model are investigated using FEM, the first in which plastic dissipation is calculated directly from thermodynamic considerations and the second one related to the simplified approach with the Taylor-Quinney coefficient. The paper is closed with Section 5 including final remarks.

2. Fundamentals

The description of large strain kinematics applied in the presented model can be found e.g. in (Bonet and Wood, 2008; Haupt, 2002; Wriggers, 2008). Let us consider a deformable continuous body whose particles in the reference configuration occupy material points denoted with the vector $\mathbf{X}$. At a time t, the placement of the particle $\mathbf{X}$ in the current configuration is described with the vector $\mathbf{x}(\mathbf{X},t)$. The displacement vector is defined as $\mathbf{u}(\mathbf{X},t)=\mathbf{x}(\mathbf{X},t)-\mathbf{X}$, whereas the deformation gradient and its determinant are

$$\mathbf{F}=\frac{\partial\mathbf{x}}{\partial\mathbf{X}}=\mathbf{I}+\frac{\partial\mathbf{u}}{\partial\mathbf{X}}\qquad J=\det(\mathbf{F}) \tag{2.1}$$

Symbol $\mathbf{I}$ in the above equation denotes the second-order identity tensor. The left Cauchy-Green deformation tensor, its determinant and the right Cauchy-Green deformation tensor are defined as follows

$$\mathbf{b}=\mathbf{F}\mathbf{F}^{\mathrm{T}}\qquad J^{b}=\det(\mathbf{b})\qquad \mathbf{C}=\mathbf{F}^{\mathrm{T}}\mathbf{F} \tag{2.2}$$

The velocity of a particle is defined as the time derivative of $\mathbf{x}$ with respect to time $\mathbf{v}=\partial\mathbf{x}/\partial t=\partial\mathbf{u}/\partial t$.

The first law of thermodynamics in the referential setting has the form, see e.g. (Holzapfel, 2000; Simo, 1998)

$$\dot{e}=\mathbf{P}:\dot{\mathbf{F}}+\mathcal{R}-\operatorname{Div}(\mathbf{Q}) \tag{2.3}$$

where e is the internal energy per unit initial volume, the dot over symbol denotes time derivative (i.e. rate), $\mathbf{P}$ denotes the first Piola-Kirchhoff stress tensor, $\mathcal{R}$ is an external source of heat per unit of the initial volume, and $\mathbf{Q}$ is the Piola-Kirchhoff heat flux density vector (for detailed discussion on heat flux density measures, see Wcisło *et al.* (2023)). Symbol $\operatorname{Div}(\cdot)$ denotes the divergence of vector $\cdot$ in the reference configuration, and the colon is used for the scalar product of two second-order tensors.

The second law of thermodynamics in the form of Clausius-Duhem inequality reads (Truesdell and Toupin, 1960)

$$\dot{s}-\frac{\mathcal{R}}{T}+\operatorname{Div}\Big(\frac{\mathbf{Q}}{T}\Big)\geqslant 0 \tag{2.4}$$

where the rate of entropy (per unit of initial volume) is denoted with $\dot{s}$, and the absolute temperature with T.

The constitutive equation for the heat flow applied in this work is spatial Fourier's law for Kirchhoff heat flux density vector $\widehat{\mathbf{q}}$ which can be equivalently expressed using the Piola Kirchhoff heat flux, see (Wcisło *et al.*, 2023)

$$\widehat{\mathbf{q}}=-k\operatorname{grad}(T)\qquad\Longleftrightarrow\qquad\mathbf{Q}=-k\mathbf{C}^{-1}\operatorname{Grad}(T) \tag{2.5}$$

In the above equation, the parameter k is the heat conductivity of the material, whereas symbols $\operatorname{grad}(\cdot)$ and $\operatorname{Grad}(\cdot)$ denote the spatial and referential gradient of quantity $\cdot$, respectively.

The balance of linear momentum for the static case with mass forces neglected has the following form (Simo, 1998)

$$\operatorname{Div}(\mathbf{P})=\mathbf{0} \tag{2.6}$$

3. Thermo-elasticity

3.1. Model of the thermo-elastic material

For the thermo-elastic material model, the free energy is assumed to be a function of the deformation gradient and temperature $\psi = \psi(\mathbf{F}, T)$. Using the Legendre transformation $\psi = e - Ts$, see e.g. (Marsden and Hughes, 1983), the following form of the second law of thermodynamics can be written

$$\Big[\mathbf{P} - \frac{\partial \psi}{\partial \mathbf{F}}\Big] : \dot{\mathbf{F}} - \Big[s + \frac{\partial \psi}{\partial T}\Big]\dot{T} - \frac{1}{T}\mathbf{Q} \cdot \mathrm{Grad}\,(T) \geqslant 0 \tag{3.1}$$

The state equations for the first Piola-Kirchhoff stress tensor and entropy are as follows

$$\mathbf{P} = \frac{\partial \psi}{\partial \mathbf{F}} \qquad s = -\frac{\partial \psi}{\partial T} \tag{3.2}$$

Then the reduced form of the second law of thermodynamics is obtained

$$\mathcal{D}_{therm} = -\frac{1}{T}\mathbf{Q} \cdot \mathrm{Grad}\,(T) \geqslant 0 \tag{3.3}$$

and it is fulfilled for the assumed Fourier law in Eq. (2.5).

Further derivations lead to temperature form of the energy balance equation

$$c\dot{T} + \mathrm{Div}\,(\mathbf{Q}) - \mathcal{R} - \mathcal{H} = 0 \qquad c = -T\frac{\partial^2 \psi}{\partial T^2} \qquad \mathcal{H} = T\frac{\partial^2 \psi}{\partial \mathbf{F}\partial T} : \dot{\mathbf{F}} \tag{3.4}$$

where c is the heat capacity per unit volume and $\mathcal{H}$ is a thermo-elastic source of heating/cooling.

In the following analysis, the specific form of the free energy function is applied, cf. (Simo and Miehe, 1992)

$$\psi(\mathbf{F}, T) = \widehat{\psi}(\mathbf{b}(\mathbf{F}), T) = \psi^{eT}(\mathbf{b}, T) + \psi^{T}(T) \tag{3.5}$$

The first component of the free energy ψ^{eT} is related to elastic deformation and thermal expansion, whereas the second component ψ^{eT} is the purely thermal part

$$\begin{aligned} \psi^{eT}(\mathbf{b}, T) &= \frac{1}{2}K\ln^2(\sqrt{J^b}) + \frac{1}{2}G\Big[\,\mathrm{tr}\,([J^b]^{-1/3}\mathbf{b}) - 3\Big] - 3K\alpha_T(T - T_0)\ln(\sqrt{J^b}) \\ \psi^{T}(T) &= c_0\Big[(T - T_0) - T\ln Bigl(\frac{T}{T_0}\Big)\Big] \end{aligned} \tag{3.6}$$

where K and G are the bulk and shear moduli, respectively, α_T is the coefficient of linear thermal expansion, c_0 is the initial specific heat capacity expressed per unit of material volume and T_0 is the initial (reference) temperature of the material. For the assumed form of the purely thermal part of the free energy function, the heat capacity in Eq. (3.4) equals $c = c_0$.

The thermo-elastic heating/cooling from Eq. (3.4) can now be derived as

$$\mathcal{H} = -3TK\alpha_T\mathbf{F}^{-1} : \dot{\mathbf{F}} \tag{3.7}$$

The value of $\mathcal{H}$ can be positive or negative: if $\mathbf{F}^{-1} : \dot{\mathbf{F}} < 0$ then the material undergoes heating, if $\mathbf{F}^{-1} : \dot{\mathbf{F}} > 0$ the material undergoes cooling. Note that the thermo-elastic heating/cooling is rate-dependent: the higher the rate of the deformation tensor, the greater effect of heating/cooling produced during deformation process. It is worth emphasizing that the thermo-elastic source of heating/cooling is dependent on the material parameters related to elasticity, i.e. the bulk modulus and thermal expansion, and there are no additional parameters which can control the heating/cooling source during elastic deformation.

3.2. Numerical simulations for thermo-elasticity

The thermo-elastic model presented in Section 3.1 is implemented within the FEM for a 3D space in *Wolfram Mathematica* packages AceGen/FEM (Korelc and Wriggers, 2016). The former package is a code generator capable of automatic differentiation, whereas the latter is a FEM engine.

The fundamental unknowns for the problem are displacement vector $\mathbf{u}$ and temperature T. For the two-field problem, hexahedral finite elements with linear interpolation of both unknown fields are applied. To avoid volumetric locking, the code includes the modification called *F-bar* (de Souza Neto *et al.*, 20O8). The implementation aspects of thermomechanical models are discussed in (Wcisło and Pamin, 2017).

3.2.1. Uniaxial tension test

The first computations are performed for a uniaxial tension test simulated with one cubic finite element with dimensions $L = W = H = 10\,\text{mm}$, see Fig. 1a. The mechanical boundary conditions are applied in such a way that the homogeneous stress state is preserved. The insulation on all sides of the cube is assumed. The enforced displacement has the maximum value ΔL=0.05 mm (the sample remains in the elastic regime) and is applied monotonically within 10 , 1 s or 0.1 s, thus there are three rates of elongation under consideration called further as slow, medium rate and fast processes. The following material data related to aluminium are used: $K = 57.133 \cdot 10^9\,\text{Pa}$, $G = 26.369 \cdot 10^9\,\text{Pa}$, $k = 121\,\text{W/(m·K)}$, $c_0 = 2.423 \cdot 10^6\,\text{J/(m}^3\text{·K)}$, $\alpha_T = 23.2 \cdot 10^{-4}\,\text{K}^{-1}$.

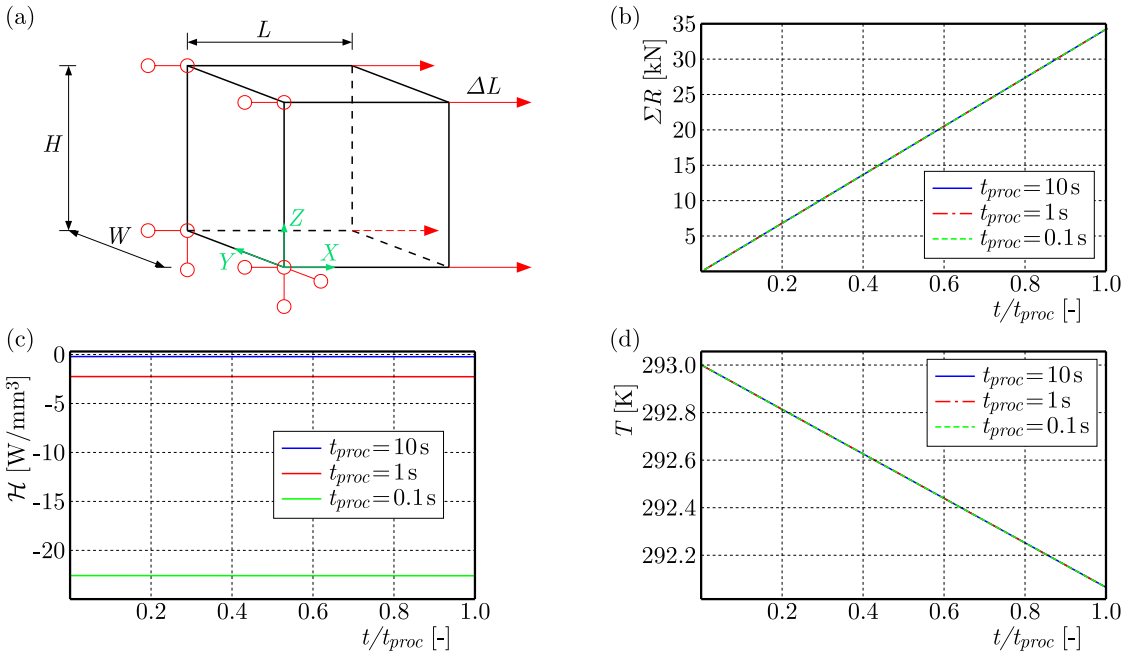


Fig. 1. Uniaxial tension test for 1 FE: (a) specimen and mechanical boundary conditions, (b) sum of reactions vs. time, (c) thermo-elastic source of heating/cooling vs. time, (d) temperature vs. time

The sum of reactions vs. (relative) process time is presented in Fig. 1b. The diagrams are almost linear, and for different elongation rates the curves coincide in spite of different values of thermo-elastic source of heating/cooling in Fig. 1c which depend on the process rate. It can be observed that the value of $\mathcal{H}$ is negative, thus indeed the tests reproduce the thermo-elastic cooling observed in metallic materials. The diagram presenting the evolution of temperature in the sample is shown in Fig. 1d. Although for different process rates, different intensity of cooling

is produced, and it does not influence the value of temperature in the sample. This behaviour can be explained by the fact that for the homogeneous temperature distribution in the sample, the energy balance in Eq. (3.4) reduces to equation $c\dot{T} = -3TK\alpha_T \mathbf{F}^{-1} : \dot{\mathbf{F}}$. The application of the backward Euler time integration $\dot{T} = (T - T_n)/\Delta t$ and $\dot{\mathbf{F}} = (\mathbf{F} - \mathbf{F}_n)/\Delta t$, where T and $\mathbf{F}$ are the values at the current time step, T_n and $\mathbf{F}_n$ are values from the previous time step, and Δt is the time increment, leads to the following closed-form formula for temperature at the current time step

$$T = T_n\Big[1 + \frac{3K\alpha_T \mathbf{F}^{-1} : (\mathbf{F} - \mathbf{F}_n)}{c}\Big]^{-1} \tag{3.8}$$

It can be noted that the current value of temperature does not depend directly on time, but only on the increment of deformation.

At this stage, it is worth mentioning that for the isotropic model, the thermo-elastic heating/cooling can also be alternatively calculated using formula (Simo and Miehe, 1992)

$$\mathcal{H} = T\frac{\partial^2 \psi}{\partial J \partial T} : \dot{J} = -3TK\alpha_T \frac{1}{J}\dot{J} \tag{3.9}$$

In this case, the rate of determinant of the deformation gradient $\dot{J}$ can be approximated straightforwardly using formula $\dot{J} = (J - J_n)/\Delta t$. However, it can be shown, see e.g. (Bonet and Wood, 2008; Wood, 2008), that $\dot{J} = J\,\mathrm{tr}\,(\mathbf{d})$, where $\mathbf{d}$ is the symmetric part of the velocity gradient. If the velocity is approximated as $\mathbf{v} = (\mathbf{u} - \mathbf{u}_n)/\Delta t$ then the obtained results can be different. It has been tested numerically that for aluminium and the elastic range, the choice of approximation of $\dot{J}$ does not influence the results significantly.

3.2.2. Dogbone sample – comparison with experiment in elastic range

This Subsection includes a comparison of experiments and simulations performed for a dogbone-shape sample in tension. The description of the laboratory experiments, which are used here, can be found in (Mucha *et al.*, 2023). The sample of thickness 2 mm presented in Fig. 2 is made of aluminium AW5083. The applied material parameters are the same as for the uniaxial tension test in the previous Subsection. In the numerical tests, the specimen is insulated on all sides. The experiments reported in (Mucha *et al.*, 2023) were performed for three displacement rates: slow $6 \cdot 10^{-5}$ m/s (experiments No. 1, 2 and 3), medium rate $6 \cdot 10^{-4}$ m/s (experiments No. 4, 5 and 6) and fast $6 \cdot 10^{-3}$ m/s (experiments No. 7, 8 and 9). The comparison of experimental and numerical results is presented in Fig. 3. For clarity, the experimental results obtained for the same process rates are marked with the same colour but different line style.

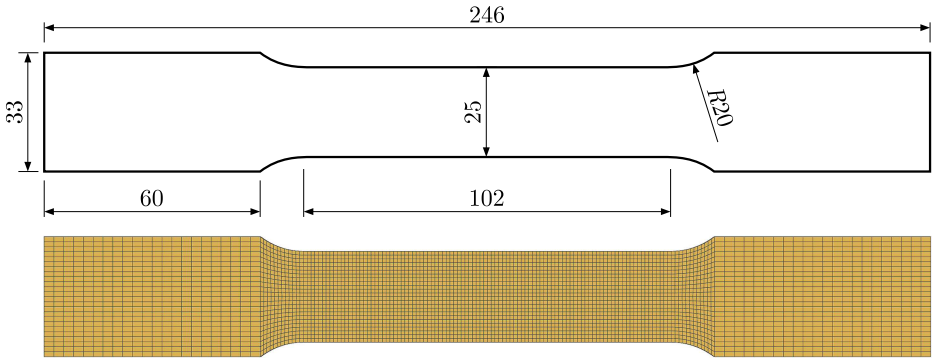


Fig. 2. Geometry, dimensions (in milimeters) and discretization of the dogbone sample

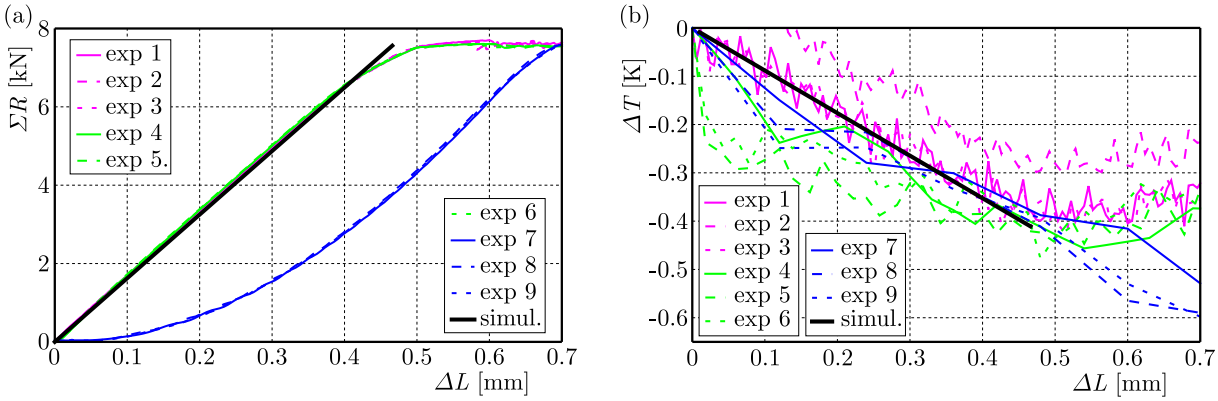


Fig. 3. Comparison of experiments and simulations for the dogbone sample in the elastic regime: (a) sum of reactions vs. enforced displacement, (b) temperature at the central material point of the sample vs. displacement

The constitutive model used for the simulation does not include viscous effects in the elastic regime, and only dependence of the material response on the process rate is related to thermo-elastic heating/cooling. However, in the analysed case, this factor has negligible influence on reactions, thus the black diagram in Fig. 3a is the result obtained for the three analysed rates. It can be observed that the reaction diagrams for simulations almost coincide with the experiments performed with slow and medium rates. In turn, the blue curves for the fast process are significantly different from the rest of results, which can be attributed to an unwanted loading machine effect.

The results of experiments presented in (Mucha *et al.*, 2023) include measurement of temperature on the surface at the central point of the sample. The comparison of experimental and simulation outcome is performed in Fig. 3b. Also in this case the results obtained from simulations performed for different displacement rates coincide (the black diagram). The reason is as follows: the simulated sample and boundary conditions are symmetric, thus the obtained temperature distribution is symmetric as well. As a result, at the central point of the sample the temperature gradient is zero, and taking the considerations from the previous Subsection into account, temperature is not dependent on the process rate. It can be observed in Fig. 3b that the numerical simulations reproduce temperature evolution at the center of the sample very well. The black line is in the middle of all experimental curves.

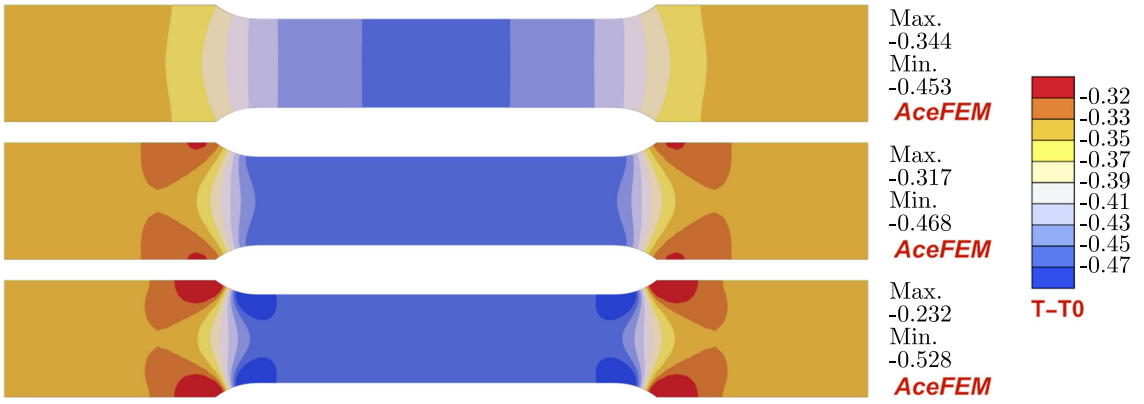


Fig. 4. Temperature distribution for slow, medium rate and fast processes (top, middle and bottom, respectively) at the end of simulation

Although at the central point of the sample temperature does not depend on the deformation rate, it is worthwhile to investigate its value obtained in simulations for the whole specimen.

Fig. 4 shows the temperature distribution at the end of simulation for the three analysed displacement rates. It can be observed that the whole sample undergoes cooling. In the central part of the specimen temperature is similar for each process rate, however significant differences are observed in the areas where the web widens. For the fast process, the difference between the minimum and maximum temperature is the highest.

4. Thermo-elasto-plasticity

4.1. Model of thermo-elasto-plastic material

The description of the thermo-elasto-plastic material is based on the assumption of multiplicative decomposition of the deformation gradient into a reversible part (related to elastic deformation and thermal expansion) and a plastic part $\mathbf{F} = \mathbf{F}^r\mathbf{F}^p$, see (Ristinmaa *et al.*, 2007), although an alternative formula $\mathbf{F} = \mathbf{F}^\theta\mathbf{F}^e\mathbf{F}^p$, where $\mathbf{F}^\theta$ is a thermal part and $\mathbf{F}^e$ an elastic one could also be applied, cf. (Wcisło and Pamin, 2017). Now, the left reversible Cauchy-Green deformation tensor and its determinant are defined as

$$\mathbf{b}^r = \mathbf{F}^r[\mathbf{F}^r]^{\mathrm{T}} \qquad J^{br} = \det(\mathbf{b}^r) \tag{4.1}$$

Further, the spatial velocity gradient and its decomposition are written, cf. (Ristinmaa *et al.*, 2007)

$$\mathbf{l} = \dot{\mathbf{F}}\mathbf{F}^{-1} \qquad \mathbf{l} = \mathbf{l}^r + \mathbf{l}^p \qquad \mathbf{l}^r = \dot{\mathbf{F}}^r[\mathbf{F}^r]^{-1} \qquad \mathbf{l}^p = \mathbf{F}^{\mathbf{r}}\dot{\mathbf{F}}^p[\mathbf{F}^p]^{-1}[\mathbf{F}^r]^{-1} \tag{4.2}$$

The symmetric part of the velocity gradient and its plastic part are as follows

$$\mathbf{d} = \mathrm{sym}\,(\mathbf{l}) \qquad \mathbf{d}^p = \mathrm{sym}\,(\mathbf{l}^p) \tag{4.3}$$

The Helmholtz free energy for the isotropic thermo-elasto-plastic material is assumed here as a function of the left reversible Cauchy-Green deformation tensor, internal variable associated with isotropic hardening α and temperature: $\psi = \psi(\mathbf{b}^r, \alpha, T)$, see (Ristinmaa *et al.*, 2007). The dissipation inequality for this case is

$$\mathcal{D} = \boldsymbol{\tau} : \mathbf{d} - s\dot{T} - \dot{\psi} - \frac{1}{T}\mathbf{Q} \cdot \mathrm{Grad}\,(T) \geqslant 0 \tag{4.4}$$

where $\boldsymbol{\tau}$ is the Kirchhoff stress tensor. After derivations, the above equation can be presented in the following form

$$\mathcal{D} = \left[\boldsymbol{\tau} - 2\frac{\partial\psi}{\partial\mathbf{b}^r}\mathbf{b}^r\right] : \mathbf{d} + 2\frac{\partial\psi}{\partial\mathbf{b}^r}\mathbf{b}^r : \mathbf{d}^p - \frac{\partial\psi}{\partial\alpha}\dot{\alpha} - \frac{1}{T}\mathbf{Q} \cdot \mathrm{Grad}\,(T) \geqslant 0 \tag{4.5}$$

The state equations for the Kirchhoff stress tensor and thermodynamic force conjugated to α are specified as

$$\boldsymbol{\tau} = 2\frac{\partial\psi}{\partial\mathbf{b}^r}\mathbf{b}^r \qquad h = \frac{\partial\psi}{\partial\alpha} \tag{4.6}$$

Now, the reduced form of dissipation inequality can be written

$$\mathcal{D} = \boldsymbol{\tau} : \mathbf{d}^p - h\dot{\alpha} - \frac{1}{T}\mathbf{Q} \cdot \mathrm{Grad}\,(T) \geqslant 0 \tag{4.7}$$

The total dissipation can be divided into mechanical and thermal contribution as follows

$$\mathcal{D}_{mech} = \boldsymbol{\tau} : \mathbf{d}^p - h\dot{\alpha} \geqslant 0 \qquad \mathcal{D}_{therm} = -\frac{1}{T}\mathbf{Q} \cdot \mathrm{Grad}\,(T) \geqslant 0 \tag{4.8}$$

Next, the energy balance equation in temperature form can be written for the thermo-elasto-plastic material

$$
\begin{aligned}
&c\dot{T} + \operatorname{Div}(\mathbf{Q}) = \underbrace{\mathcal{H} + \mathcal{D}_{mech} + \mathcal{A}}_{Q_{mech}} + \mathcal{R} \qquad c = -T\frac{\partial^2 \psi}{\partial T^2} \\
&\mathcal{H} = \Big[T\frac{\partial \boldsymbol{\tau}}{\partial T}\Big] : [\mathbf{d} - \mathbf{d}^p] \qquad \mathcal{A} = T\frac{\partial h}{\partial T}\dot{\alpha}
\end{aligned}
\tag{4.9}
$$

In the above equation, Q_{mech} represents the mechanical heat production rate which consists of thermo-elastic heating/cooling $\mathcal{H}$, mechanical dissipation related to the plastic process $\mathcal{D}_{mech}$ and contribution $\mathcal{A}$ which is related to the temperature dependence of the thermodynamic force conjugated to the hardening variable.

In the subsequent analysis, we assume free energy $\psi = \psi^r(\mathbf{b}^r, T) + \psi^p(\alpha) + \psi^T(T)$ which is additively decoupled into reversible, plastic and purely thermal parts as follows, see (Ristinmaa *et al.*, 2007)

$$
\begin{aligned}
\psi^r &= \frac{1}{2}K\ln^2(J) + \frac{1}{2}G\Big[\operatorname{tr}([J^{br}]^{-1/3}\mathbf{b}^r) - 3\Big] - 3K\alpha_T(T - T_0)\ln(J) \\
\psi^p &= \frac{1}{2}H\alpha^2 + [\sigma_{yf} - \sigma_{y0}]\Big[\alpha + \frac{1}{\delta}\exp(-\delta\alpha)\Big] \\
\psi^T(T) &= c_0\Big[(T - T_0) - T\ln\Big(\frac{T}{T_0}\Big)\Big]
\end{aligned}
\tag{4.10}
$$

The introduction of the above specific form of the free energy allows for derivation of the thermodynamic force conjugated to the hardening variable as

$$
h = H\alpha + [\sigma_{yf} - \sigma_{y0}][1 - \exp(-\delta\alpha)] \tag{4.11}
$$

where σ_{y0} is the initial yield threshold, σ_{yf} is the final yield threshold and δ is a saturation constant. Moreover, the heat capacity from Eq. (4.9) remains constant $c = c_0$.

To complete the description of plasticity, the yield function and the flow rule have to be specified. In this work, the volume preserving Huber-Mises-Hencky yield function is applied

$$
\begin{aligned}
&F_p(\boldsymbol{\tau}, \alpha, T) = f(\boldsymbol{\tau}) - \sqrt{\frac{2}{3}}\sigma_y(\alpha, T) \leqslant 0 \qquad f(\boldsymbol{\tau}) = \sqrt{\boldsymbol{\tau}_{dev} : \boldsymbol{\tau}_{dev}} \\
&\boldsymbol{\tau}_{dev} = \boldsymbol{\tau} - \frac{1}{3}\operatorname{tr}(\boldsymbol{\tau})\mathbf{I} \qquad \sigma_y(\alpha, T) = \Big[1 - H_T[T - T_0]\Big]\sigma_{y0} + h
\end{aligned}
\tag{4.12}
$$

where H_T is the thermal softening modulus and T_0 is the reference temperature.

The flow rule is defined through the Lie derivative of $\mathbf{b}^r$, cf. (Simo and Miehe, 1992)

$$
-\frac{1}{2}\mathcal{L}_v\mathbf{b}^r = \dot{\gamma}\mathbf{N}^p\mathbf{b}^r \qquad \mathbf{N}^p = \frac{\partial F_p}{\partial \boldsymbol{\tau}}
$$

where $\dot{\gamma}$ is a plastic multiplier related to the hardening variable by formula $\dot{\alpha} = \sqrt{2/3}\dot{\gamma}$.

Now, having the thermo-elasto-plastic model in hand, the structural sources of heating can be analysed in more detail. The thermo-elastic source of heat from Eq. (4.9) can be written as

$$
\mathcal{H} = -3K\alpha_T[\operatorname{tr}(\mathbf{d}) - \operatorname{tr}(\mathbf{d}^p)] = -3K\alpha_T\Big[\frac{1}{J}\dot{J} - \frac{1}{J^p}\dot{J}^p\Big] \qquad J^p = \det(\mathbf{F}^p) \tag{4.13}
$$

The yield function leads to $\dot{J}^p = 0$, so that the above equation can be rewritten as

$$
\mathcal{H} = -3TK\alpha_T\operatorname{tr}(\mathbf{d}) = -3TK\alpha_T\frac{1}{J}\dot{J} \tag{4.14}
$$

The second structural source of heat, i.e. plastic dissipation, can be derived as

$$\mathcal{D}_{mech} = \underbrace{\boldsymbol{\tau} : \mathbf{d}^p}_{\dot{w}^p} - h\dot{\alpha} = \dot{\gamma}\Big[f - \sqrt{\frac{2}{3}}h\Big] = \sqrt{\frac{2}{3}}\dot{\gamma}\Big[1 - H_T[T - T_0]\Big]\sigma_{y0} \tag{4.15}$$

The fraction of rate of plastic work $\dot{w}^p$ which is converted into heat can be calculated as, cf. (Ristinmaa *et al.*, 2007)

$$\eta = 1 - \frac{h\dot{\alpha}}{\boldsymbol{\tau} : \mathbf{d}^p} = \frac{[1 - H_T[T - T_0]]\sigma_{y0}}{\sigma_y} \tag{4.16}$$

As it was mentioned in the introduction, an alternative estimation for the plastic dissipation often used in literature is, see e.g. (Wriggers *et al.*, 1992)

$$\mathcal{D}_{mech} = \chi\dot{w}^p = \chi[\boldsymbol{\tau} : \mathbf{d}^p] = \sqrt{\frac{2}{3}}\chi\dot{\gamma}\sigma_y \tag{4.17}$$

where χ is the fraction of rate of plastic work converted into heat, often called the Taylor--Quinney coefficient, which is usually assumed to be a constant material parameter with value from interval 0.8-0.95.

4.2. Numerical simulations for thermo-elasto-plasticity

Numerical verification of the presented thermo-elasto-plastic model is performed similarly to the thermo-elasticity using two specimens: the cube presented in Fig. 1a and the dogbone sample, see Fig. 2. The main attention is now paid to the influence of transition from the elastic to plastic regime on the temperature change in the sample and the impact of the applied formulation of plastic dissipation, in particular plastic dissipation calculated:

a) straightforwardly from thermodynamic, i.e. using Eq. (4.15), called further *Model 1*
b) using Taylor-Quinney coefficient according to Eq. (4.17) called further *Model 2*.

Both samples are simulated using the same material properties. The elastic and thermal parameters related to aluminium are taken from Section 3.2.1, whereas the parameters describing the plastic behaviour are as follows: $\sigma_{y0} = 150 \cdot 10^6\,\mathrm{Pa}$, $\sigma_{yf} = 390 \cdot 10^6\,\mathrm{Pa}$, $H = 0$, $\delta = 12$, $H_T = 0.0016\,\mathrm{K}^{-1}$. For Model 2, the value of the Taylor-Quinney coefficient equals $\chi = 0.9$.

4.2.1. Uniaxial tension test

The cubic sample from Fig. 1a is now elongated by $\Delta L = 1\,\mathrm{mm}$ within 16.7 s which gives the displacement rate equal to $6 \cdot 10^{-5}$ m/s. The results obtained in simulations are presented in Fig. 5. The diagram presenting the sum of reactions (Fig. 5a) shows that the onset of plasticity takes place when $t/t_{proc} = 0.02$ and from this point the diagram related to plasticity with significant strain hardening is observed. The reactions are very close for the two applied models of plastic dissipation, however, differences are observed for the temperature diagram in Fig. 5b. In this case, Model 1 manifests higher temperature in the sample at the beginning of the plastic process and lower in the second part, and this is consistent with the amount of dissipated energy during the process, cf. Fig. 5c. As the elongation progresses, the difference in the plastic dissipation for the two models becomes greater. In the last diagram presented in Fig. 5, the fraction of the plastic work converted into heat is presented for the two analysed models. For Model 1, the fraction with coefficient η defined in Eq. (4.16) is shown. It can be observed that the constant fraction of the plastic work defined by coefficient χ used in Model 2 can be treated as an average of the fraction calculated with Model 1 for the initial part of plastic deformation (for the whole process the averaged value of η is lower than the applied value of χ).

It is worth noting that in the elastic regime the thermo-elastic cooling is properly reproduced as a decrease in temperature at the beginning of the process, see Fig. 5b. The onset of plasticity can also be recognized as the moment at which temperature in the sample starts to grow.

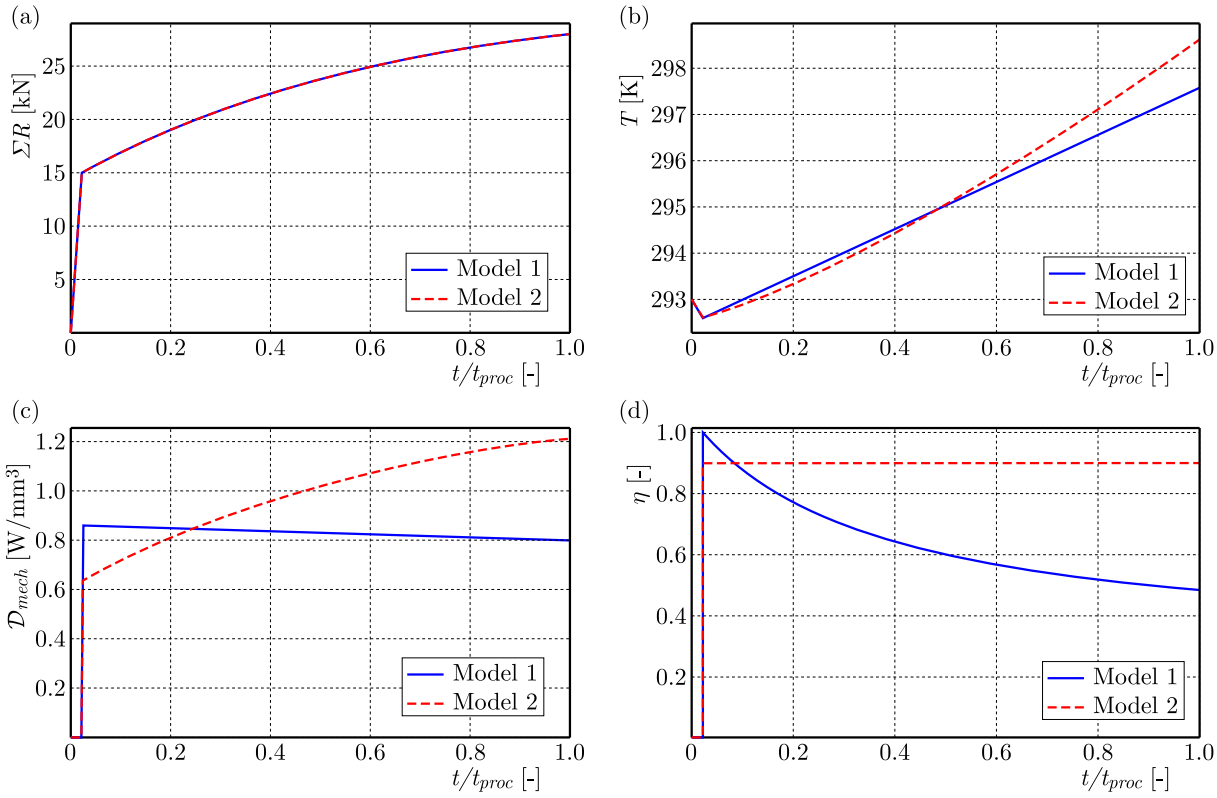


Fig. 5. Results for the uniaxial tension test for thermo-elasto-plasticity: (a) sum of reactions, (b) temperature, (c) plastic dissipation, (d) fraction of plastic work converted into heat source

4.2.2. Dogbone sample – comparison with experiment for elasto-plasticity

The dogbone sample is elongated now with the displacement rate $6{\cdot}10^{-5}$ m/s. The simulations are performed for the slow process due to the fact that the plasticity description does not include viscosity, which can have a significant impact on the results for faster deformation. The results of computational tests for the dogbone specimen are presented in Fig. 6. At the initial stage of the plastic process, see Fig. 6a, a slight softening and plateau in the diagram is visible. This behaviour is related to the formation and expansion of Lueders bands. The material model which is used in the simulation does not reproduce this phenomenon, however, it correctly it simulates the onset of plasticity and the overall plastic behaviour until the failure, see Fig. 6b. The oscillations visible in the experimental diagrams in Fig. 6b are a result of the PLC effect. The differences between Model 1 and Model 2 are negligible in the displacement-force diagram, however the choice of the model has a significant influence on temperature evolution, see Fig. 7. Model 1, which predicts plastic dissipation directly from thermodynamics, shows higher temperature at the beginning of the plastic process and lower in the following part of the process with respect to Model 2. The analysis of the diagrams in Fig. 7 clearly shows that Model 1 is closer to the experimental results.

Note that the temperature evolution in the elastic regime is very close to experimental measurements and, what is of great importance, the increase of temperature starts in simulations and experiments exactly at the same moment. Thus the analysis of temperature evolution (the end of elastic cooling) allows for detection of plasticity onset.

The deformed samples with temperature distribution in the middle of the elongation process (enforced displacement equals 15 mm) are depicted in Fig. 8. Both the plots of temperature, obtained with Model 1 and Model 2, are presented using the same scale, which allows for a

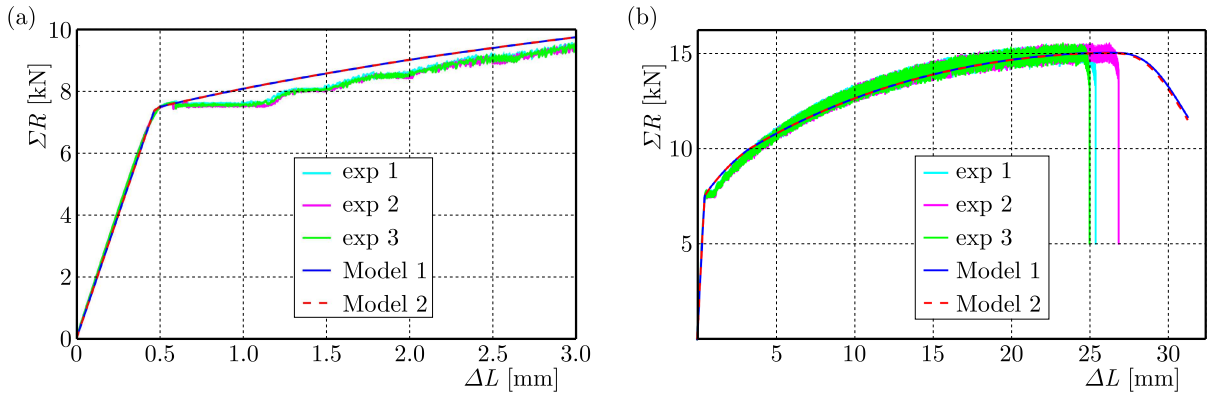


Fig. 6. Sum of reactions vs. enforced displacement: (a) initial stage, (b) whole process

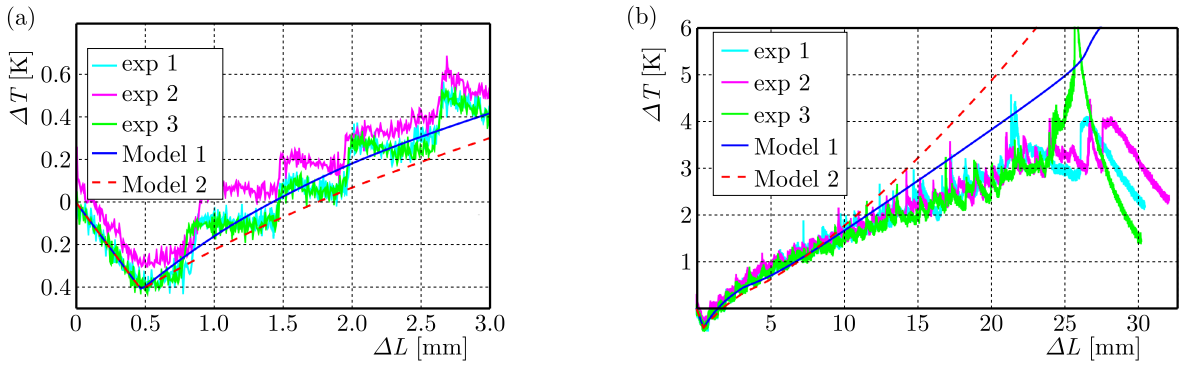


Fig. 7. Temperature at the central point of the sample vs. enforced displacement: (a) initial stage, (b) whole process

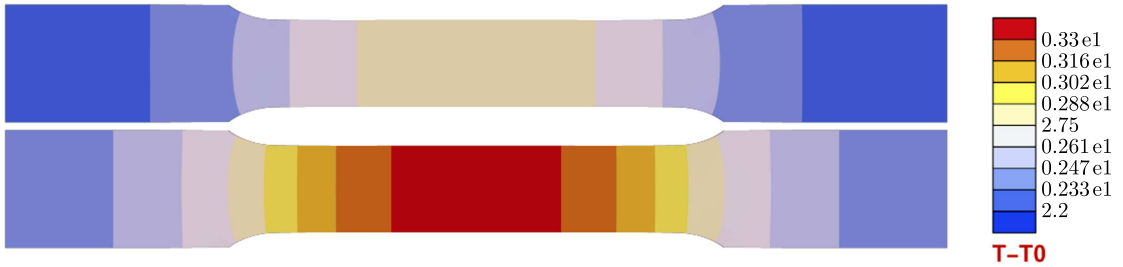


Fig. 8. Deformed mesh with temperature distribution in the middle of the deformation process for Model 1 (top) and Model 2 (bottom)

comparison. Model 2 manifests significantly higher temperatures in the sample than Model 1. For both models, the highest temperature is located in the central part of the sample and it is a precursor of necking.

5. Conclusion

The paper is focused on internal heat sources observed in large strain thermo-elastic and thermo-elasto-plastic materials. It includes description of material models and presentation of numerical simulations with remarks aimed at deeper understanding of the phenomena. Computations have been performed using AceGen/AceFEM packages in Wolfram Mathematica, starting from a one-element test. Then a dogbone-shape sample has been analysed, for which the results have been compared with experimental ones.

In the elastic regime, the presented and implemented model reproduces thermo-elastic cooling characteristics for metals, e.g. for aluminium, which is in general a rate-dependent phenomenon. However, it has been shown in the analysis of uniaxial tension tests that for samples with homogeneous temperature distribution the diagrams do not depend on the tension rate. Experimental and computational results for the dogbone sample show a good agreement for both the sum of reactions and temperature diagrams, even though there is no direct material parameter to control the cooling rate in the model.

For the plastic material, the attention has been focused on the heating source related to plastic dissipation. Two models have been used, Model 1 in which dissipation is calculated directly from thermodynamic derivation and Model 2 where estimation of plastic dissipation is obtained using the Taylor-Quinney coefficient. The force-displacement diagrams for the two thermo-elasto-plastic models overlap, however, the diagrams of temperature differ from the beginning of the plastic process, which is a result of different amount of dissipated energy in the two analysed models. Overall, good agreement between experimental and computational results has been obtained, and Model 1 gives results closer to experimental ones especially for the advanced stage of the process.

The incorporation of thermo-elastic coupling in the thermo-elasto-plastic model allows one to recognize the beginning of the plastic process at a material point, and the presented simulations properly reproduce the transition from elastic to plastic regime observed in the experimental sample.

Acknowledgement

The authors acknowledge valuable discussions on the research with Prof. Andreas Menzel (TU Dortmund University/Lund University) and Dr. Lars Rose (TU Dortmund University).

References

1. Bonet J., Wood R.D., 3008, *Nonlinear Continuum Mechanics for Finite Element Analysis*, 2nd ed., Cambridge University Press, Cambridge
2. de Souza Neto E.A., Peric D., Owen D.R.J., 2008, *Computational Methods for Plasticity. Theory and Applications*, John Wiley & Sons, Chichester, UK
3. Duszek M., Perzyna P., 1991, The localization of plastic deformation in thermoplastic solids, *International Journal of Solids and Structures*, **27**, 11, 1419-1443
4. Haupt P., 2002, *Continuum Mechanics and Theory of Materials*, Springer
5. Holzapfel G.A., 2000, *Nonlinear Solid Mechanics. A Continuum Approach for Engineering*, John Wiley & Sons, Chichester
6. Korelc J., Wriggers P., 2016, *Automation of Finite Element Methods*, Springer International Publishing Switzerland
7. Marsden J.E., Hughes T.J.R., 1983, *Mathematical Foundations of Elasticity*, Dover Publications, Inc., New York
8. Mucha M., Rose L., Wcisło B., Menzel A., Pamin J., 2023, Experiments and numerical simulations of Lueders bands and Portevin-Le Chatelier effect in aluminium alloy AW5083, *Archives of Mechanics*, **75**, 3, 301-336
9. Musiał S., Maj M., Urbański L., Nowak M., 2022, Field analysis of energy conversion during plastic deformation of 310S stainless steel, *International Journal of Solids and Structures*, **238**, 1, 111411
10. Oliferuk W., Maj M., Zembrzycki K., 2013, Determination of the energy storage rate distribution in the area of strain localization using infrared and visible imaging, *Experimental Mechanics*, **55**, 4, 753-760

11. Oppermann P., Denzer R., Menzel A., 2022, A thermo-viscoplasticity model for metals over wide temperature ranges – application to case hardening steel, *Computational Mechanics*, **69**, 541-563

12. Ristinmaa M., Wallin M., Ottosen N.S., 2007, Thermodynamic format and heat generation of isotropic hardening plasticity, *Acta Mechanica*, **194**, 103-121

13. Rose L., Menzel A., 2021, Identification of thermal material parameters for thermo-mechanically coupled material models, *Meccanica*, **56**, 393-416

14. Simo J.C., 1998, Numerical analysis and simulation of plasticity, [In:] P.G. Ciarlet and J.L. Lions, Edit., *Handbook of Numerical Analysis. Numerical Methods for Solids (Part 3)*, Vol, VI, 183-499, Elsevier Science, Boca Raton

15. Simo J.C., Hughes T.J.R., 1998, *Computational Inelasticity. Interdisciplinary Applied Mathematics*, Vol. 7, Springer-Verlag, New York

16. Simo J.C., Miehe C., 1992, Associative coupled thermoplasticity at finite strains: Formulation, numerical analysis and implementation, *Computer Methods in Applied Mechanics and Engineering*, **98**, 1, 41-104

17. Taylor G.I., Quinney H., 1934, The latent energy remaining in a metal after cold working, *Proceedings of the Royal Society of London. Series A*, **143**, 307-326

18. Truesdell C., Toupin R.A., 1960, The classical field theories, [In:] S. Flügge, Edit., *Encyclopedia of Physics. Vol. III Principles of Classical Mechanics and Field Theory*, 226-788, Springer-Verlag, Berlin Heidelberg

19. Wcisło B., Pamin J., 2017, Local and non-local thermomechanical modeling of elastic-plastic materials undergoing large strains, *International Journal for Numerical Methods in Engineering*, **109**, 1, 102-124

20. Wcisło B., Pamin J., Rose L., Menzel A., 2023, On spatial vs. referential isotropic Fourier's law in finite deformation thermomechanics, *Engineering Transactions*, **71**, 1, 111140

21. Wriggers P., 2008, *Nonlinear Finite Element Methods*, Springer-Verlag, Berlin Heidelberg

22. Wriggers P.A., Miehe C., Kleiber M., Simo J.C., 1992, On the coupled thermomechnical treatment of necking problems via finite element methods, *International Journal for Numerical Methods in Engineering*, **33**, 869-883

23. Zienkiewicz O.C., Taylor R.L., Zhu J.Z., 2005, *The Finite Element Method: Its Basis and Fundamentals*, 6th ed., Elsevier Butterworth-Heinemann

Manuscript received October 25, 2023; accepted for print January 22, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 307-319, Warsaw 2024
https://doi.org/10.15632/jtam-pl/184322

DYNAMIC RESPONSE OF A GUY LINE OF A GUYED TOWER TO STOCHASTIC WIND EXCITATION: 3D NON-LINEAR SMALL-SAG CABLE MODEL[1]

Hanna Weber, Anna Jabłonka
West Pomeranian University of Technology, Szczecin, Poland
e-mail: hanna.weber@zut.edu.pl; anna.jablonka@zut.edu.pl

Radosław Iwankiewicz
Calisia University, Kalisz, Poland
e-mail: r.iwankiewicz@uniwersytetkaliski.edu.pl

In the proposed approach, a 3D response of the guy line treated as a small-sag cable is considered. The strong dynamic wind action leads to the base motion excitation of the guy line. Longitudinal cable displacements are coupled with lateral ones. Hamilton's principle and Galerkin method are used to obtain the set of differential equations of motion. The cable excitation is assumed as a narrow-band stochastic process modelled as a response of an auxiliary linear filter to a Gaussian white noise process. The equivalent linearization technique is applied to obtain approximate analytical results verified against the numerical Monte Carlo simulation.

Keywords: equivalent linearization technique, non-linear system, small-sag cable, spatial response, stochastic dynamics

1. Introduction

The equations and numerical results included in this paper concern the problem presented at the 5th Polish Congress of Mechanics and 25th Conference on Computer Methods in Mechanics (PCM-CMM) which was held in September 2023. Nowadays, the use of cables in various civil engineering structures has become very popular. Classic examples include suspended or cable--stayed-bridges (Larsen and Larose, 2015). On the other hand, modern cable roof coverings are becoming more and more common (Xue *et al.*, 2022). What is worth mentioning, steel ropes are often used as flexible supports or system stabilizing elements, such as hangers (Zhu *et al.*, 2023) or guying elements in masts and towers (Shi and Salim, 2015). In each type of structures mentioned above, the function and behavior of the cable is different and requires a different approach at the design stage. Therefore, in the literature many articles dedicated to various methods of analyzing rope structures may be found, from analytical approach to complex finite element models (Ha *et al.*, 2018; Biliszczuk *et al.*, 2021).

Structural cables are flexible elements that can carry only tensile forces, however, depending on their function, types of support in the system and, above all, cross-sectional area can be considered as elements with some bending stiffness (Zhang *et al.*, 2021). It needs to be mentioned that the value of bending stiffness should be determined from experimental tests (Chen *et al.*, 2015). Due to their use in the structure, cables are exposed to external factors such as rain, snow, wind, and their slenderness makes them sensitive to various dynamic loads (Caracoglia and Zuo, 2009), which, due to randomness, should be considered using stochastic analysis (Georgakis and Taylor, 2005; Li and Chen, 2009).

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

Analytical or finite element models give good results in static analysis. However, dynamics of these systems create many problems. Results from nonlinear models often differ from those obtained by experimental measurements, while complex finite element models are characterized by a large amount of time that needs to be spent on modeling and conducting the analysis. Therefore, there is a constant search for methods and tools enabling quick dynamic analysis of cables, taking into account random loads, which would support the design process.

In the presented approach, a simplified model of a single guy line in guyed towers and its 3D response to the base-motion excitation modelled as a response of an auxiliary linear filter to Gaussian white noise excitation is considered. In the guyed lines with significant length that are exposed to external factors like wind and temperature changes, most of the time some sag can be observed, even if the value of the pre-tension force is large. Therefore, the nonlinear model based on a small-sag cable is developed where longitudinal vibrations of the guyed line are coupled with transverse ones that are considered in two different directions: in and out of the cable plane. Next, the equivalent linearization technique (Socha, 2007; Roberts and Spanos, 1990) is used to solve the set of nonlinear differential equations of motion and obtain variances and cross-covariances of particular random state variables. The received results are compared with those obtained by the Monte Carlo simulation (Proppe *et al.*, 2003).

2. Nonlinear equation of motion – 3D response of a small-sag cable

In the presented approach, the initial tension in the guy line denoted as H is assumed to be very high in comparison to the effect of own weight of the rope (gravity forces), therefore the line is regarded as a small-sag cable. It is the case when the ratio of the sag to the initial length of the rope is equal or less than 1:8 (Irvine, 1981). Moreover, the 3D response of the cable is considered, where $u(x,t)$ are longitudinal displacements of the guy line, while $z(x)$ and $w(x,t)$ are the initial shape of the cable in its plane in the direction perpendicular to the guy line and the transverse displacements of the cable resulting from deformation, respectively (see Fig. 1a). The displacement out of the cable plane is denoted as $v(x,t)$ (Fig. 1b). The axial stiffness of the cable and its total length are denoted as EA and L, respectively, while mass per unit length of the rope is denoted as μ.

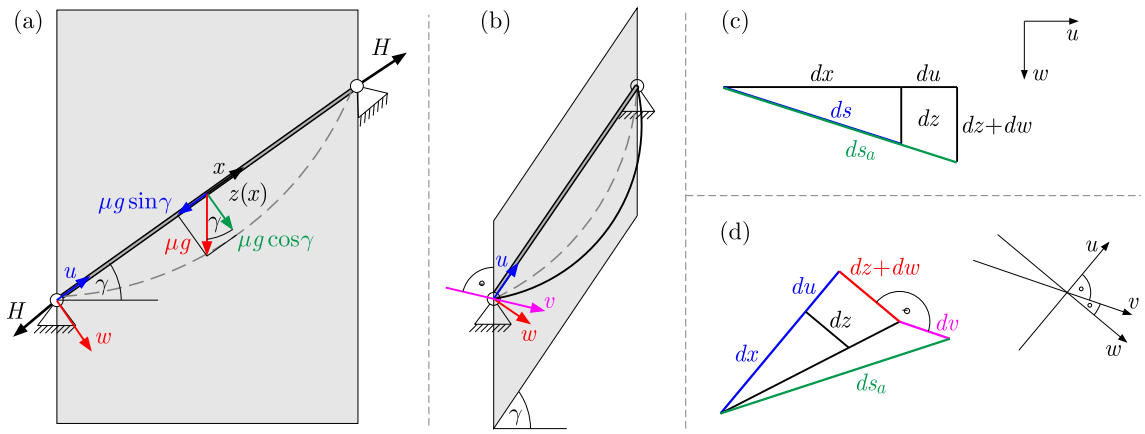


Fig. 1. Small-sag cable model of the guy line model under gravity forces: (a) planar view, (b) 3D view. Differential element of the small-sag cable: (c) planar view, (d) 3D view

If g is the gravity acceleration, the second derivative of $z(x)$ with respect to x is defined as

$$\frac{d^2z}{dx^2} = -\frac{\mu g}{H}\cos\gamma \tag{2.1}$$

Since the small-sag cable is considered, the initial shape of the guy line can be treated as a parabola. Based on that we can assume at for the support points $z(x=0)=0$ and at the mid span where the maximum lateral displacement can be observed $dz/dx(x=L/2)=0$. The integration process with using these two conditions results in the following equation

$$z = \frac{\mu g}{2H} x(L-x)\cos\gamma \tag{2.2}$$

If V and T denote the potential and kinetic energies, Hamilton's principle is given by

$$\int_0^t \delta(V-T)\,dt = 0 \tag{2.3}$$

For a 3D response of a cable with small sag, the kinetic energy can by expressed as

$$T = \frac{\mu}{2}\int_0^L \left(\left(\frac{\partial u}{\partial t}\right)^2 + \left(\frac{\partial w}{\partial t}\right)^2 + \left(\frac{\partial v}{\partial t}\right)^2\right) ds \tag{2.4}$$

Using the initial shape of the cable, according to Fig. 1c, leads to

$$ds = \sqrt{dx^2+dz^2} = \sqrt{dx^2\left(1+\left(\frac{dz}{dx}\right)^2\right)} = \sqrt{1+(z')^2}\,dx \tag{2.5}$$

The variation of the kinetic energy is then given by

$$\delta T = \frac{1}{2}\mu\int_0^L \delta\Big(2\frac{\partial u}{\partial t}\delta\dot{u} + 2\frac{\partial w}{\partial t}\delta\dot{w} + 2\frac{\partial v}{\partial t}\delta\dot{v}\Big)\sqrt{1+(z')^2}\,dx \tag{2.6}$$

Taking into account that the variation of the derivative equals the derivative of the variation, and assuming the vanishing of the variations δu and δw because of the fixed states at the initial time 0 and at the final time t, one obtains

$$\int_0^t \delta T = -\mu\int_0^t\int_0^L \Big(\frac{\partial^2 u}{\partial t^2}\delta u + \frac{\partial^2 w}{\partial t^2}\delta w + \frac{\partial^2 v}{\partial t^2}\delta v\Big)\sqrt{1+(z')^2}\,dx\,dt \tag{2.7}$$

If $V^{(g)}$ denotes the gravitational potential energy and $N(x)$ is the cable initial static tension, the elastic potential energy of the system is given by

$$V = \underbrace{\int_0^L N(x)\varepsilon(u',w',v')\,ds}_{V^{(1)}} + \underbrace{\frac{EA}{2}\int_0^L \varepsilon^2(u',w',v')\,ds}_{V^{(2)}} + V^{(g)} \tag{2.8}$$

where ε is the normal strain defined as $\varepsilon = (ds_a - ds)/ds$, (see Fig. 1d). After neglecting the term $(\partial z/\partial x)^2$ due to its insignificant value compared to the others, the term ds_a is obtained in the following form

$$\begin{aligned} ds_a &= \sqrt{(dx+du)^2+(dz+dw)^2+dv^2} \\ &= dx\sqrt{1+2\frac{\partial u}{\partial x}+\left(\frac{\partial u}{\partial x}\right)^2+2\frac{\partial z}{\partial x}\frac{\partial w}{\partial x}+\left(\frac{\partial w}{\partial x}\right)^2+\left(\frac{\partial v}{\partial x}\right)^2} \end{aligned} \tag{2.9}$$

If the small-sag cable is considered, the simplification $1/\sqrt{1+(\partial z/\partial x)^2} \approx 1$ can be assumed. Using Eq. (2.4) and the Taylor series expansion results in the equation for normal strain in the presented form

$$\varepsilon \cong \frac{\partial u}{\partial x} + \frac{1}{2}\Big(\frac{\partial u}{\partial x}\Big)^2 + \frac{\partial z}{\partial x}\frac{\partial w}{\partial x} + \frac{1}{2}\Big(\frac{\partial w}{\partial x}\Big)^2 + \frac{1}{2}\Big(\frac{\partial v}{\partial x}\Big)^2 = u' + \frac{1}{2}(u')^2 + z'w' + \frac{1}{2}(w')^2 + \frac{1}{2}(v')^2 \tag{2.10}$$

Assuming that $N(x)/\sqrt{1+(z')^2} = H(x)$, the variation of the first term of potential energy is obtained

$$\begin{aligned}\delta V^{(1)} &= \delta \int_0^L N(x)\varepsilon(u', w', v)\, ds = \int_0^L N(x)\frac{\delta u' + u'\delta u' + z'\delta w' + w'\delta w' + v'\delta v'}{1+(z')^2}\sqrt{1+(z')^2}\, dx \\ &= \int_0^L H(x)(\delta u' + u'\delta u' + z'\delta w' + w'\delta w' + v'\delta v')\, dx\end{aligned} \tag{2.11}$$

Taking into account that $\delta u' = (\delta u)' = \partial\delta u/\partial x$, $\delta w' = (\delta w)' = \partial\delta w/\partial x$ and $\delta v' = (\delta v)' = \partial\delta v/\partial x$, the terms of Eq. (2.11) that depend on $u(x,t)$, $w(x,t)$ and $v(x,t)$, respectively, are defined by

$$\begin{aligned}&\int_0^L H(x)(\delta u' + u'\delta u')\, dx = H(x)(\delta u)\Big|_0^L - \int_0^L \frac{\partial H}{\partial x}\, dx\delta u + Hu'\delta u\Big|_0^L - \int_0^L \frac{\partial}{\partial x}(Hu')\, dx\delta u \\ &\int_0^L H(x)(z'\delta w' + w'\delta w')\, dx = H(x)z'(\delta w)\Big|_0^L - \int_0^L \frac{\partial}{\partial x}(H(x)z')\, dx\delta w + H(x)w'\delta w\Big|_0^L \\ &\quad - \int_0^L \frac{\partial}{\partial x}(H(x)w')\, dx\delta w \\ &\int_0^L H(x)(v'\delta v')\, dx = H(x)v'\delta v\Big|_0^L - \int_0^L \frac{\partial}{\partial x}(H(x)v')\, dx\delta v\end{aligned} \tag{2.12}$$

The gravitational potential energy is given by the following equation

$$V^{(g)} = -\int_0^L \mu g w\, ds = -\int_0^L \mu g w\sqrt{1+(z')^2}\, dx \tag{2.13}$$

while its variation is obtained as

$$\delta V^{(g)} = -\int_0^L \mu g\sqrt{1+(z')^2}\, dx\, \delta w \tag{2.14}$$

The below self-satisfied equation of equilibrium is subtracted from the final form of the equation of motion

$$-\frac{\partial}{\partial x}(H(x)z') - \mu g\sqrt{1+(z')^2} = 0 \tag{2.15}$$

Variation of the second term of potential energy is given by

$$\begin{aligned}\delta V^{(2)} &= \frac{EA}{2}\delta\int_0^L \varepsilon^2(u', w', v')\, ds = EA\int_0^L \varepsilon(u', w', v')\delta\varepsilon(u', w', v')\, ds \\ &\cong EA\int_0^L \Big(u' + \frac{1}{2}(u')^2 + z'w' + \frac{1}{2}(w')^2 + \frac{1}{2}(v')^2\Big)(\delta u' + u'\delta u' + z'\delta w' + w'\delta w' + v'\delta v')\, dx\end{aligned} \tag{2.16}$$

Using the rule that the variation of the derivative is equal the derivative of the variation, particular terms of Eq. (2.16) that depend on $u(x,t)$, $w(x,t)$ and $v(x,t)$, respectively, are defined as

$$\begin{aligned}
&EA\int_0^L \Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(1+u')\delta u'\,dx \\
&\quad = EA\frac{\partial}{\partial x}\Big[\Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(1+u')\Big] \\
&EA\int_0^L \Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(z'+w')\delta w'\,dx \\
&\quad = EA\frac{\partial}{\partial x}\Big[\Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(z'+w')\Big] \\
&EA\int_0^L \Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(v')\delta v'\,dx \\
&\quad = EA\frac{\partial}{\partial x}\Big[\Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(v')\Big]
\end{aligned} \tag{2.17}$$

If Hamilton's principle is used (Eq. (2.3)) for Eqs. (2.12) and Eqs. (2.17), the following set of equations is obtained

$$\begin{aligned}
&-\frac{\partial}{\partial x}(H(x)u') - EA\frac{\partial}{\partial x}\Big[\Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(1+u')\Big] + \mu\frac{\partial^2 u}{\partial t^2} = 0 \\
&-\frac{\partial}{\partial x}(H(x)w') - EA\frac{\partial}{\partial x}\Big[\Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(z'+w')\Big] + \mu\frac{\partial^2 w}{\partial t^2} = 0 \\
&-\frac{\partial}{\partial x}(H(x)v') - EA\frac{\partial}{\partial x}\Big[\Big(u' + \frac{(u')^2}{2} + z'w' + \frac{(w')^2}{2} + \frac{(v')^2}{2}\Big)(v')\Big] + \mu\frac{\partial^2 v}{\partial t^2} = 0
\end{aligned} \tag{2.18}$$

3. Base motion excitation – dynamics of a guy line

The displacement $U(t)$ of a tower at the point of attachment of the guy line is treated as a base motion excitation for guy line vibration (see Fig. 2). For the case that the horizontal displacement of the guyed tower is out the cable plane, the components $U_u(t) = U(t)\cos\gamma\cos\eta$ and $U_w(t) = U(t)\sin\gamma\cos\eta$, are

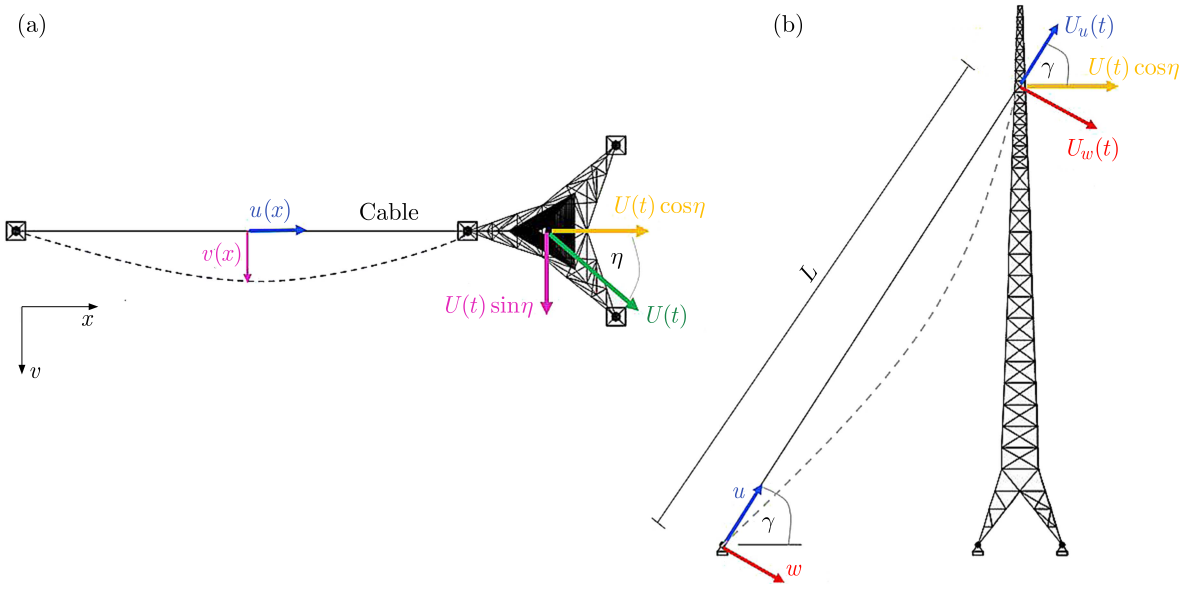


Fig. 2. Guy line base motion: (a) top view, (b) in cable plane view

excitations for motion in the longitudinal and transverse direction in the cable plane, respectively, while $U_v(t) = U(t)\sin\eta$ is the base motion excitation in the out of plane direction, where γ and η are the slope of the cable and the angle in the horizontal plane between the direction of displacements $U(t)$ and cable plane, respectively. If $\overline{u}(x,t)$, $\overline{w}(x,t)$ and $\overline{v}(x,t)$ denote absolute motions expressed in terms of the relative motions $u(x,t)$, $w(x,t)$ and $v(x,t)$, which are related with elastic deformations and base motion due to $U(t)$, they can be given by

$$\begin{aligned}
&\overline{u}(x,t) = \frac{x}{L}U(t)\cos\gamma\cos\eta + u(x,t) \qquad\qquad \overline{w}(x,t) = \frac{x}{L}U(t)\sin\gamma\cos\eta + w(x,t)\\
&\overline{v}(x,t) = \frac{x}{L}U(t)\sin\eta + v(x,t)
\end{aligned} \tag{3.1}$$

Equation (2.18) expressed by absolute motions is obtained as

$$\begin{aligned}
&\int_0^L \Big\{ -\frac{\partial H(x)}{\partial x}\frac{\partial\overline{u}}{\partial x} - H(x)\frac{\partial^2\overline{u}}{\partial x^2} - EA\frac{\partial}{\partial x}\Big[\frac{\partial\overline{u}}{\partial x} + \frac{3}{2}\Big(\frac{\partial\overline{u}}{\partial x}\Big)^2 + \frac{\partial z}{\partial x}\frac{\partial\overline{w}}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{w}}{\partial x}\Big)^2\\
&\quad + \frac{1}{2}\Big(\frac{\partial\overline{v}}{\partial x}\Big)^2 + \frac{1}{2}\Big(\frac{\partial\overline{u}}{\partial x}\Big)^3 + \frac{\partial\overline{u}}{\partial x}\frac{\partial z}{\partial x}\frac{\partial\overline{w}}{\partial x} + \frac{1}{2}\frac{\partial\overline{u}}{\partial x}\Big(\frac{\partial\overline{w}}{\partial x}\Big)^2 + \frac{1}{2}\frac{\partial\overline{u}}{\partial x}\Big(\frac{\partial\overline{v}}{\partial x}\Big)^2\Big] + \mu\frac{\partial^2\overline{u}}{\partial t^2}\Big\}\delta\overline{u}\,dx\\
&+ \int_0^L \Big\{ -\frac{\partial H(x)}{\partial x}\frac{\partial\overline{w}}{\partial x} - H(x)\frac{\partial^2\overline{w}}{\partial x^2} - EA\frac{\partial}{\partial x}\Big[\frac{\partial\overline{u}}{\partial x}\frac{\partial z}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{u}}{\partial x}\Big)^2\frac{\partial z}{\partial x} + \Big(\frac{\partial z}{\partial x}\Big)^2\frac{\partial\overline{w}}{\partial x}\\
&\quad + \frac{3}{2}\frac{\partial z}{\partial x}\Big(\frac{\partial\overline{w}}{\partial x}\Big)^2 + \frac{1}{2}\Big(\frac{\partial\overline{v}}{\partial x}\Big)^2\frac{\partial z}{\partial x} + \frac{\partial\overline{u}}{\partial x}\frac{\partial\overline{w}}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{u}}{\partial x}\Big)^2\frac{\partial\overline{w}}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{w}}{\partial x}\Big)^3\\
&\quad + \frac{1}{2}\Big(\frac{\partial\overline{v}}{\partial x}\Big)^2\frac{\partial\overline{w}}{\partial x}\Big] + \mu\frac{\partial^2\overline{w}}{\partial t^2}\Big\}\delta\overline{w}\,dx + \int_0^L \Big\{ -\frac{\partial H(x)}{\partial x}\frac{\partial\overline{v}}{\partial x} - H(x)\frac{\partial^2\overline{v}}{\partial x^2}\\
&- EA\frac{\partial}{\partial x}\Big[\frac{\partial\overline{u}}{\partial x}\frac{\partial\overline{v}}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{u}}{\partial x}\Big)^2\frac{\partial\overline{v}}{\partial x} + \frac{\partial z}{\partial x}\frac{\partial\overline{w}}{\partial x}\frac{\partial\overline{v}}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{w}}{\partial x}\Big)^2\frac{\partial\overline{v}}{\partial x} + \frac{1}{2}\Big(\frac{\partial\overline{v}}{\partial x}\Big)^3\Big] + \mu\frac{\partial^2\overline{v}}{\partial t^2}\Big\}\delta\overline{v}\,dx = 0
\end{aligned} \tag{3.2}$$

Using the relationships $\delta\overline{u}(x,t) = \delta u(x,t)$, $\delta\overline{w}(x,t) = \delta w(x,t)$ and $\delta\overline{v}(x,t) = \delta v(x,t)$, the time derivatives are expressed by

$$\begin{aligned}
&\frac{\partial^2\overline{u}}{\partial t^2} = \frac{x}{L}\ddot{U}(t)\cos\gamma\cos\eta + \frac{\partial^2 u}{\partial t^2} \qquad\qquad \frac{\partial^2\overline{w}}{\partial t^2} = \frac{x}{L}\ddot{U}(t)\sin\gamma\cos\eta + \frac{\partial^2 w}{\partial t^2}\\
&\frac{\partial^2\overline{v}}{\partial t^2} = \frac{x}{L}\ddot{U}(t)\sin\eta + \frac{\partial^2 v}{\partial t^2}
\end{aligned} \tag{3.3}$$

When the derivatives with respect to x are considered, the base motion terms vanish. Using Galerkin's method and single-mode approximation, the particular displacements are defined as

$$u(x,t) = p(t)\sin\frac{\pi x}{L} \qquad w(x,t) = q(t)\sin\frac{\pi x}{L} \qquad v(x,t) = r(t)\sin\frac{\pi x}{L} \tag{3.4}$$

and, consequently, their variations are given by

$$\delta u(x,t) = \delta p(t)\sin\frac{\pi x}{L} \qquad \delta w(x,t) = \delta q(t)\sin\frac{\pi x}{L} \qquad \delta v(x,t) = \delta r(t)\sin\frac{\pi x}{L} \tag{3.5}$$

In the considered small-sag cable model, the initial tension is much more significant in comparison to the dead load of the line, therefore $H = \text{const}$ can be assumed. After including damping forces depending on relative velocities $-c_u\partial u/\partial t$, $-c_w\partial w/\partial t$ and $-c_v\partial v/\partial t$ together with Eqs. (3.3)-(3.5) in Eq. (3.2), and after integration, the following set of nonlinear equations is obtained

$$\begin{aligned}
&\ddot{p}(t) + a_1 p(t) + a_2 p^3(t) - 2a_3 p(t)q(t) + a_2 p(t)q^2(t) + a_2 p(t)r^2(t) + \frac{c_u}{\mu}\dot{p}(t) + Ha_1 p(t) = -\beta_u\ddot{U}(t)\\
&\ddot{q}(t) - a_3 p^2(t) - a_4 q(t) - 3a_3 q^2(t) - a_3 r^2(t) + a_2 p^2(t)q(t) + a_2 q^3(t) + a_2 r^2(t)q(t)\\
&\quad + \frac{c_w}{\mu}\dot{q}(t) + Ha_1 q(t) = -\beta_w\ddot{U}(t)\\
&\ddot{r}(t) + a_2 p^2(t)r(t) - 2a_3 q(t)r(t) + a_2 q^2(t)r(t) + a_2 r^3(t) + \frac{c_v}{\mu}\dot{r}(t) + Ha_1 r(t) = -\beta_v\ddot{U}(t)
\end{aligned} \tag{3.6}$$

where the particular constant terms are denoted as

$$a_1 = \frac{\pi^2}{\mu L^2} \qquad a_2 = EA\frac{3\pi^4}{8\mu L^4} \qquad a_3 = -EA\frac{14g\pi}{9HL^2}\cos\gamma$$
$$a_4 = -EA\Big(\frac{\mu g}{H}\cos\gamma\Big)^2\Big(\frac{6+\pi^2}{12\mu}\Big) \qquad \beta_u = \frac{2}{\pi}\cos\gamma\cos\eta$$
$$\beta_w = \frac{2}{\pi}\sin\gamma\cos\eta \qquad \beta_v = \frac{2}{\pi}\sin\eta$$

4. Stochastic governing equations

The structure displacement $U(t)$ is assumed to be dominated by the fundamental mode shape of the tower with the corresponding natural frequency Ω_o. Since the stochastic wind excitation in the form of a strong wind gust can be treated as a stationary wide band process, the process $U(t)$ is considered as a narrow-band one, with the central frequency Ω_o. In the presented approach, it is assumed as the Gaussian white noise passed through the first-order linear filter, giving the process $X(t)$, which is subsequently passed through the second-order linear filter. Therefore, the process $U(t)$ is governed by the stochastic equations defined as

$$\ddot{U} + 2\zeta_f\Omega_o\dot{U} + \Omega_o^2 U = X(t) \qquad \dot{X} + \alpha X = \alpha\sqrt{2\pi S_o}\xi(t) \tag{4.1}$$

where ζ_f is damping of the linear filter, $\xi(t)$ denotes a Gaussian white noise while S_0 is its spectral density. It should be noted that the process $U(t)$, as the displacement response, must be twice differentiable. That condition will be fulfilled if

$$\int_{-\infty}^{\infty}\omega^4 S_{UU}(\omega)\,d\omega < \infty \qquad \text{where} \qquad S_{UU}(\omega) = \frac{S_0\alpha^2}{(\omega^2+\alpha^2)[(\Omega_0^2-\omega^2)^2+(2\zeta_f\Omega_0\omega)^2]} \tag{4.2}$$

$S_{UU}(\omega)$ is the spectral density of the process $U(t)$ while its steady-state variance σ_U^2 is given by the following expression

$$\sigma_U^2 = \frac{\alpha\pi S_0(2\zeta_f\Omega_o+\alpha)}{2\zeta_f\Omega_o^3(2\alpha\zeta_f\Omega_o+\alpha^2+\Omega_o^2)} \qquad \text{with} \qquad \alpha = \Omega_o\left(-\zeta_f + \sqrt{\zeta_f^2 + \frac{\zeta_f\Omega_o^3A_0^2}{\pi S_0 - \zeta_f\Omega_o^3A_0^2}}\right) \tag{4.3}$$

The expression for α is obtained from the condition of the mean-square equivalence of the horizontal displacement response $U(t)$ to the harmonic process with the amplitude A_0, frequency Ω_0 and variance $\sigma_U^2 = A_0^2/2$. Using Eqs. (3.6) together with Eqs. (4.1) leads to the set of differential equations of motion

$$\begin{aligned}
\ddot{p}(t) &= -a_1(EA+H)p(t) - a_2p^3(t) + 2a_3p(t)q(t) - a_2p(t)q^2(t) - a_2p(t)r^2(t) - \frac{c_u}{\mu}\dot{p}(t) - \beta_u\ddot{U}(t)\\
\ddot{q}(t) &= a_3p^2(t) + a_4q(t) + 3a_3q^2(t) + a_3r^2(t) - a_2p^2(t)q(t) - a_2q^3(t) - a_2r^2(t)q(t)\\
&\quad - \frac{c_w}{\mu}\dot{q}(t) - a_1Hq(t) - \beta_w\ddot{U}(t)\\
\ddot{r}(t) &= -a_2p^2(t)r(t) + 2a_3q(t)r(t) - a_2q^2(t)r(t) - a_2r^3(t) - \frac{c_v}{\mu}\dot{r}(t) - a_1Hr(t) - \beta_v\ddot{U}(t)\\
\ddot{U}(t) &= X(t) - 2\zeta_f\Omega_o\dot{U}(t) - \Omega_o^2U(t) \qquad \dot{X} = -\alpha X + \alpha\sqrt{2\pi S_o}\xi(t)
\end{aligned} \tag{4.4}$$

The stochastic equations of motion in state space form are

$$\dot{\mathbf{Y}}(t) = \mathbf{c}(\mathbf{Y}(t))dt + \boldsymbol{\sigma}dW(t) \tag{4.5}$$

where $W(t)$ denotes the standard Wiener process, $\mathbf{c}(\mathbf{Y}(t))$ is the drift vector and $\boldsymbol{\sigma}$ means the diffusion vector. If the state vector is assumed as $\mathbf{Y}(t) = [p(t), \dot{p}(t), q(t), \dot{q}(t), r(t), \dot{r}(t), U(t), \dot{U}(t), X(t)]^{\mathrm{T}}$, the particular elements of the drift vector are obtained as

$$
\begin{aligned}
&c_1(\mathbf{Y}(t)) = \dot{p}(t)\\
&c_2(\mathbf{Y}(t)) = -a_1(EA+H)p(t) - a_2p^3(t) + 2a_3p(t)q(t) - a_2p(t)q^2(t) - a_2p(t)r^2(t) - \frac{c_u}{\mu}\dot{p}(t)\\
&\quad + \beta_u(\Omega_o^2U(t) + 2\zeta_f\Omega_o\dot{U}(t) - X(t))\\
&c_3(\mathbf{Y}(t)) = \dot{q}(t)\\
&c_4(\mathbf{Y}(t)) = a_3p^2(t) + a_4q(t) + 3a_3q^2(t) + a_3r^2(t) - a_2p^2(t)q(t) - a_2q^3(t) - a_2r^2(t)q(t)\\
&\quad - \frac{c_w}{\mu}\dot{q}(t) - a_1Hq(t) + \beta_w(\Omega_o^2U(t) + 2\zeta_f\Omega_o\dot{U}(t) - X(t))\\
&c_5(\mathbf{Y}(t)) = \dot{r}(t)\\
&c_6(\mathbf{Y}(t)) = -a_2p^2(t)r(t) + 2a_3q(t)r(t) - a_2q^2(t)r(t) - a_2r^3(t) - \frac{c_v}{\mu}\dot{r}(t) - a_1Hr(t)\\
&\quad + \beta_v(\Omega_o^2U(t) + 2\zeta_f\Omega_o\dot{U}(t) - X(t))\\
&c_7(\mathbf{Y}(t)) = \dot{U}(t)\\
&c_8(\mathbf{Y}(t)) = -\Omega_o^2U(t) - 2\zeta_f\Omega_o\dot{U}(t) + X(t)\\
&c_9(\mathbf{Y}(t)) = -\alpha X(t)
\end{aligned}
\tag{4.6}
$$

and the diffusion vector is defined as

$$
\boldsymbol{\sigma} = [0,0,0,0,0,0,0,0,\alpha\sqrt{2\pi S_o}]^{\mathrm{T}} \tag{4.7}
$$

5. Equivalent (statistical) linearization approach

The augmented state vector transformation to the centralized state vector is required to convert the original nonlinear set of differential equations into the linear one by using the equivalent linearization technique (ELT). The centralized state vector is defined as

$$
\mathbf{Y}^0(t) = [Y_1^0, Y_2^0, Y_3^0, Y_4^0, Y_5^0, Y_6^0, Y_7^0, Y_8^0, Y_9^0]^{\mathrm{T}} \tag{5.1}
$$

where its particular elements are given by $Y_i^0(t) = Y_i(t) - \mathrm{E}[Y_i(t)]$. The stochastic equation expressed in terms of the centralized state vector $\mathbf{Y}^0(t)$ and centralized drift vector $\mathbf{c}^0(\mathbf{Y}^0(t),t)$ is defined as

$$
d\mathbf{Y}^0(t) = \mathbf{c}^0(\mathbf{Y}^0(t),t)dt + \boldsymbol{\sigma}(t)dW(t) \tag{5.2}
$$

with

$$
\mathbf{c}^0(\mathbf{Y}^0(t),t) = \mathbf{c}(\mathbf{Y}^0(t),t) - \mathrm{E}[\mathbf{c}(\mathbf{Y}^0(t),t)]
$$

The idea of the equivalent linearization technique is the replacement of the original non-linear equation given by Eq. (4.5) with a linear one defined as

$$
d\mathbf{Y}^0(t) = \mathbf{B}\mathbf{Y}^0(t)dt + \boldsymbol{\sigma}dW(t) \tag{5.3}
$$

where the centralized drift terms are expressed by the linear function of the state variables $\mathbf{Y}^0(t)$ and equivalent coefficients $\mathbf{B}$. Using the condition of minimization of mean-square errors between the original model and the linear one, the equivalent coefficients can be determined from the following expression

$$
B_{im}\kappa_{mj} = \mathrm{E}[Y_j^0c_i^0(\mathbf{Y}^0)] \tag{5.4}
$$

where κ_{mj} denotes the covariance function of the state variables m and j. The centralized state variables $\mathbf{Y}^0$ are jointly Gaussian distributed, therefore in further consideration, the relationship given by Atalik Utku (1976) is used

$$
\mathrm{E}[\mathbf{X}f(\mathbf{X})] = \mathrm{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]\mathrm{E}[\nabla f(\mathbf{X})] \tag{5.5}
$$

where $\mathbf{X}$ is the zero-mean Gaussian random vector, $f(\mathbf{X})$ denotes a non-linear function and ∇ is given by the following expression $\nabla = [\partial/\partial X_1, \partial/\partial X_2, \ldots, \partial/\partial X_n]^{\mathrm{T}}$. If Eq. (5.5) is used in transposed form of Eq. (5.4), the following expression is obtained

$$
\boldsymbol{\kappa}(t)\mathbf{B}^{\mathrm{T}} = \boldsymbol{\kappa}(t)\mathrm{E}[\nabla\mathbf{c}^{0\mathrm{T}}(\mathbf{Y}^0(t))] \qquad \text{with} \qquad \mathbf{B}^{\mathrm{T}} = \mathrm{E}[\nabla\mathbf{c}^{0\mathrm{T}}(\mathbf{Y}^0(t))] \tag{5.6}
$$

The result of applying Eq. (5.6) to the elements of the centralized drift vector is the matrix $\mathbf{B}$ defined as

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_1 & -\frac{c_u}{\mu} & b_2 & 0 & b_3 & 0 & \beta_u\Omega_o^2 & 2\beta_u\zeta_f\Omega_o & -\beta_u \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ b_2 & 0 & b_4 & -\frac{c_w}{\mu} & b_5 & 0 & \beta_w\Omega_o^2 & 2\beta_w\zeta_f\Omega_o & -\beta_w \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ b_3 & 0 & b_5 & 0 & b_6 & -\frac{c_v}{\mu} & \beta_v\Omega_o^2 & 2\beta_v\zeta_f\Omega_o & -\beta_v \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\Omega_o^2 & -2\zeta_f\Omega_o & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\alpha \end{bmatrix}$$

where

$$\begin{aligned}
b_1 &= -a_1(EA+H) - a_2\Big(3\mathrm{E}[(Y_1^0)^2] + 3(\mathrm{E}[p(t)])^2\Big) + \mathrm{E}[(Y_3^0)^2] + (\mathrm{E}[q(t)])^2 + \mathrm{E}[(Y_5^0)^2] \\
&\quad + (\mathrm{E}[r(t)])^2) + 2a_3\mathrm{E}[q(t)] \\
b_2 &= -2a_2(\mathrm{E}[Y_1^0Y_3^0] + \mathrm{E}[q(t)]\mathrm{E}[p(t)]) + 2a_3\mathrm{E}[p(t)] \\
b_3 &= -2a_2(\mathrm{E}[Y_1^0Y_5^0] + \mathrm{E}[r(t)]\mathrm{E}[p(t)]) \\
b_4 &= -a_2\Big(\mathrm{E}[(Y_1^0)^2] + (\mathrm{E}[p(t)])^2 + 3\mathrm{E}[(Y_3^0)^2] + 3(\mathrm{E}[q(t)])^2 + \mathrm{E}[(Y_5^0)^2] + (\mathrm{E}[r(t)])^2\Big) \\
&\quad + 6a_3\mathrm{E}[q(t)] + a_4 - a_1H \\
b_5 &= -2a_2\Big(\mathrm{E}[Y_5^0Y_3^0] + \mathrm{E}[r(t)]\mathrm{E}[q(t)]\Big) + 2a_3\mathrm{E}[r(t)] \\
b_6 &= -a_1H - a_2\Big(3\mathrm{E}[(Y_5^0)^2] + 3(\mathrm{E}[r(t)])^2 + \mathrm{E}[(Y_1^0)^2] + (\mathrm{E}[p(t)])^2 + \mathrm{E}[(Y_3^0)^2] + (\mathrm{E}[q(t)])^2\Big) \\
&\quad + 2a_3\mathrm{E}[q(t)]
\end{aligned}$$

To obtain variances and covariances of particular random state variables, the following set of differential equations for the covariance matrix $\boldsymbol{\kappa}_{\mathbf{Y}^0\mathbf{Y}^0} = \mathrm{E}[\mathbf{Y}^0\mathbf{Y}^{0\mathrm{T}}]$ should be solved

$$\frac{d}{dt}\boldsymbol{\kappa}_{\mathbf{Y}^0\mathbf{Y}^0} = \mathbf{B}\boldsymbol{\kappa}_{\mathbf{Y}^0\mathbf{Y}^0} + \boldsymbol{\kappa}_{\mathbf{Y}^0\mathbf{Y}^0}\mathbf{B}^{\mathrm{T}} + \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}} \tag{5.7}$$

together with the differential equations for mean values defined by

$$\frac{d}{dt}\mathrm{E}[\mathbf{Y}(t)] = \mathrm{E}[\mathbf{c}(\mathbf{Y}^0(t), t)] \tag{5.8}$$

As a result, a set of 54 differential equations is obtained that can be solved numerically.

6. Numerical examples – results and discussion

In the considered problem, the simplified model of a steel guyed tower with a single guy line is examined. The tower with triangular cross-section supported on three pin supports that is presented in Fig. 2 was firstly considered by the finite element method (FEM). The total height of the structure is assumed as 300 m while the point at which the guy line is attached to the tower is located on the level 252 m. The slope of the cable is assumed as $\gamma = 57°$, that gives the total length of the guy line equal to $L = 300$ m. The mass-per unit length of the steel rope and its longitudinal stiffness are assumed as $\mu = 7.47$ kg/m and $EA = 195$ MN, respectively. In the FEM analysis, the particular bars are modeled as 3D beam elements while the guy line as an elastic cable with a given pre-tension. It turns out that the presence of the guy line in the model does not affect the result of the whole system fundamental frequency, that equals $\Omega_0 = 1.82$ rad/s, which corresponds to the assumption that the cable has no significant influence on the fundamental frequency of the tower. However the static analysis of the guyed tower under the dead load of structural elements and static wind load gives a conclusion that the value of the pre-tension force has influence on the maximum horizontal displacement of the guy line attachment point. It turns out that the higher the value of the pretension force the larger the horizontal displacement in the structure. It is

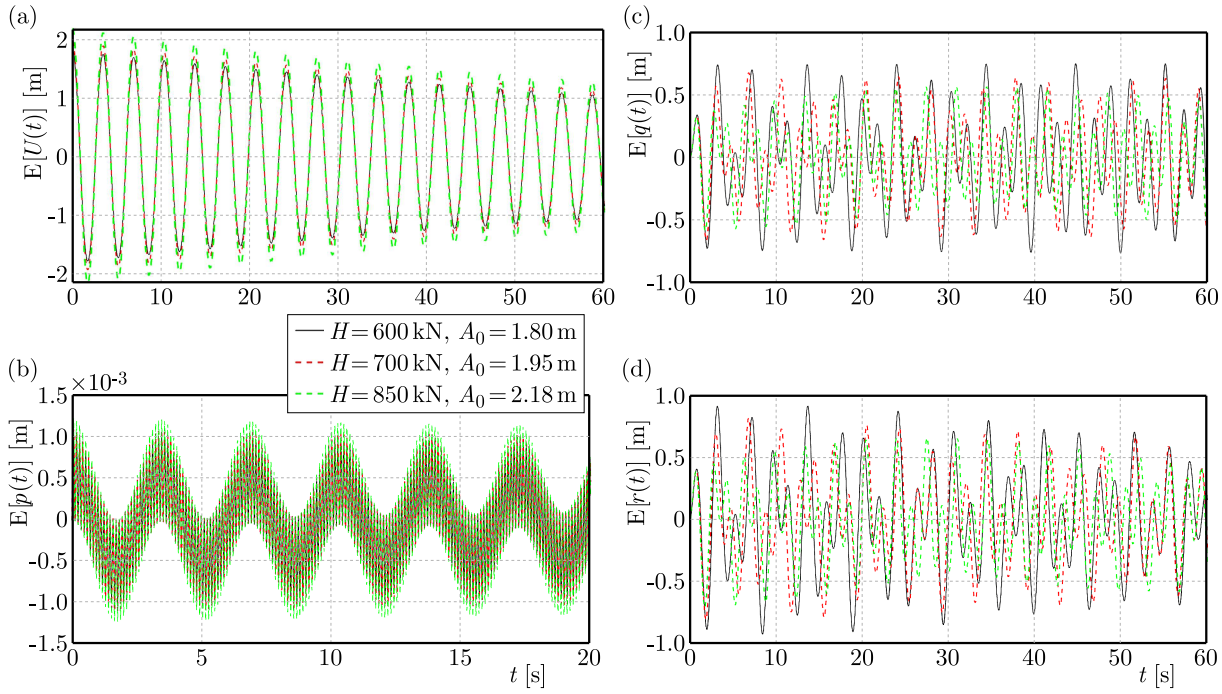


Fig. 3. Expected values of particular random state variables for various initial tension obtained by ELT

observed when the wind acts on the tower in the opposite direction to the action of the cable, it makes the rope compressed, so it is inactive, but its initial tension leads to an increase in displacement.

The maximum value of the tower horizontal displacement under the static load obtained by the FEM method is taken as the amplitude A_0 of the process $U(t)$ together with the corresponding pre-tension force H in the presented nonlinear model, Eqs. (5.7) and (5.8). The results of expected values obtained by the equivalent linearization technique (ELT) for selected cases are presented in Fig. 3. In every case, the Gaussian white noise process spectral density, damping of the linear filter and damping coefficients are assumed as $S_0 = 1$, $\zeta_f = 0.005$ and $c_u = c_w = c_v = 0.03\,\mathrm{Ns/m^2}$, respectively. It is considered that the wind is acting parallel to the cable plane, therefore $\eta = 0^\circ$ is assumed. The whole motion is examined during 60 s, however for clarity of presentation, some results with significant vibration frequency are presented in a shorter time interval. As it can be seen, the bigger the amplitude A_0, the larger the expected values of the tower horizontal displacement $\mathrm{E}[U(t)]$ (Fig. 3a) and the expected generalized coordinate of cable longitudinal vibrations $\mathrm{E}[p(t)]$ (Fig. 3b), which seems natural. It is worth noticing that even if the wind is acting parallel to the cable plane, the results of expected values of generalized coordinates of the cable lateral vibrations in and out of the plane, i.e. $\mathrm{E}[q(t)]$ (Fig. 3c) and $\mathrm{E}[r(t)]$ (Fig. 3d), respectively, are comparable. However, the behaviour of these random variables is opposite to the longitudinal displacement. Increasing the pre-tension force, which leads to increasing its stiffness due to the greater axial force, results in decreasing the expected values of generalized coordinate in the cable lateral vibration.

The same regularity can be observed in diagrams of the variances of particular random state variables. In the case of $\mathrm{Var}\,[X(t)]$ (Fig. 4a), $\mathrm{Var}\,[U(t)]$ (Fig 4b) and $\mathrm{Var}\,[p(t)]$ (Fig. 4c) increasing the amplitude of the tower horizontal displacement leads to increasing the value of the variance. On the other hand, the lower pre-tension force leads to decreasing the stiffness of the guy line and, consequently, the variances of generalized coordinates of cable lateral vibration in and out of the guy line plane increase (compare Fig. 4d,e), but they are also comparable.

However, for the lowest values of the initial tension some wrong negative results of the variance of velocity of the longitudinal cable vibrations $\mathrm{Var}\,[\dot{p}(t)]$ are obtained (not reported here in the figure). All results of expected values and variances of particular random state variables are obtained directly from numerical solution of the differential set of equations described by Eqs. (5.7) and (5.8). As is well known, no variance can be negative. Such behaviour may be caused by numerical errors that arise in the solution of the set of differential equations because of very small final results. On the other hand, it should be also admitted that the ELT method has some limitations, namely when non-linearity of the considered

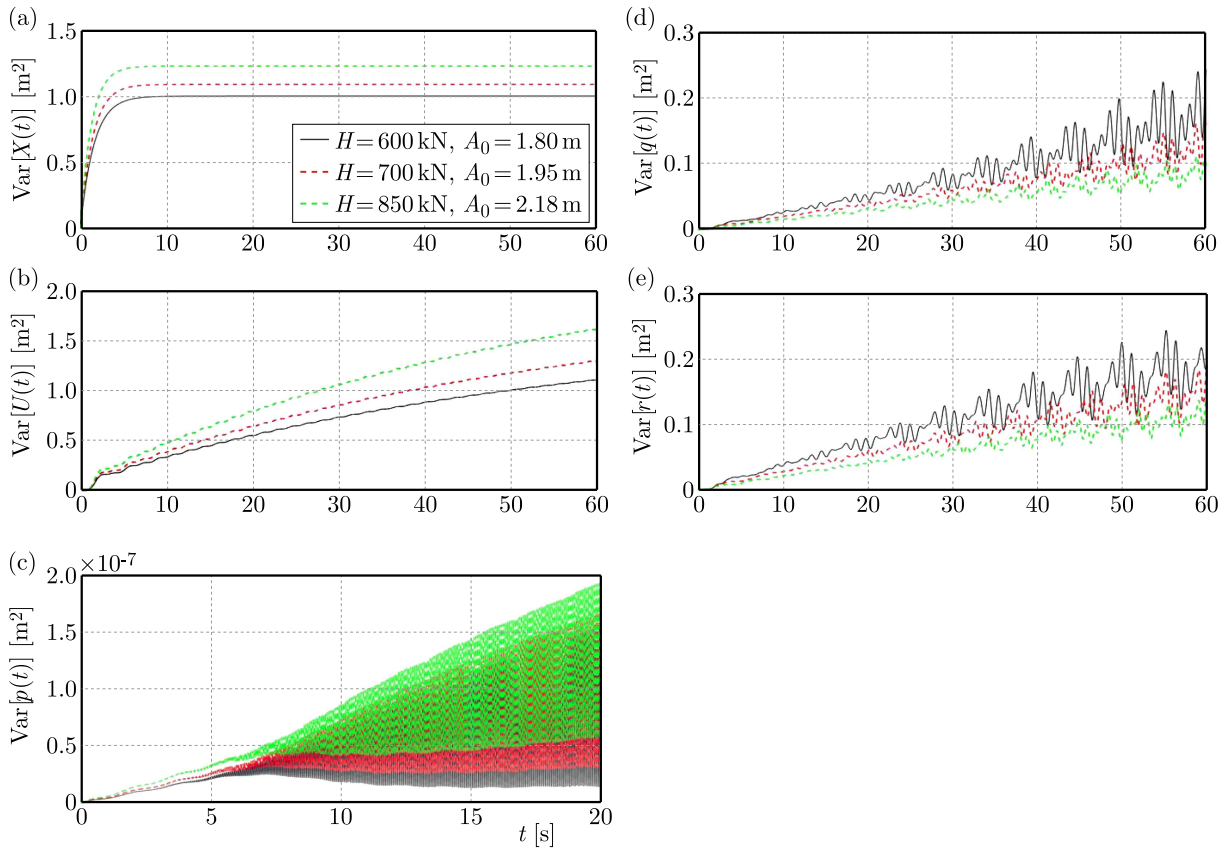


Fig. 4. Variances of particular random state variables for various initial tension

problem is strong, it can result in the incorrect results. Therefore, some additional numerical tests were conducted and the comparison of the course of variances $\mathrm{Var}\,[\dot{p}(t)]$ obtained for different directions of wind action, $H = 850\,\mathrm{kN}$ and $A_0 = 2.18\,\mathrm{m}$ were made. During the wind action in the cable plane ($\eta = 0°$), the results of variance are positive, but when the action of the wind is changed to $\eta = 45°$ and the other parameters of motion remain unchanged, the non-linearities arise and some results become negative. This confirmed the previous assumption.

The obtained results were verified by the Monte Carlo Simulation (MCS) conducted for the set of equations (4.4) with using 4000 sample functions and time step of computations $\Delta t = 0.005\,\mathrm{s}$. The value of standard deviation of the Gaussian white noise process simulated in the numerical computations is adopted as $1/\sqrt{\Delta t}$ (Weber *et al.*, 2021), to obtain results independent of the time step. Due to the very long time needed to conduct the simulation, only first 20 s of motion were examined. The comparison of diagrams obtained by both methods for $H = 850\,\mathrm{kN}$, $A_0 = 2.18\,\mathrm{m}$, $S_0 = 1$ and $\zeta_f = 0.005$ are presented in Figs. 5-6. As it can be seen, the expected values and variances of particular random state variables obtained from ELT and MCS are in good agreement. Only in the case of variance of the longitudinal cable vibration $\mathrm{Var}\,[p(t)]$ the results from ELT show a bigger amplitude of vibration in comparison to the results from the MCS. However, the MCS diagram course is exactly in the middle of that obtained by the ELT and additionally the meaningful values are very small, so the difference may be caused by numerical errors. The main advantage of the ELT method is easy application in numerical calculation and a very short time needed to obtain the result in comparison to the MCS, which takes many hours due to the large number of sample functions required to get smooth diagrams.

7. Concluding remarks

The presented approach shows that the nonlinear 3D response of the cable in the considered model of the guyed tower under stochastic excitation can be successfully solved by using an equivalent linearization technique. As it is shown, the obtained results are comparable with those obtained by the Monte Carlo

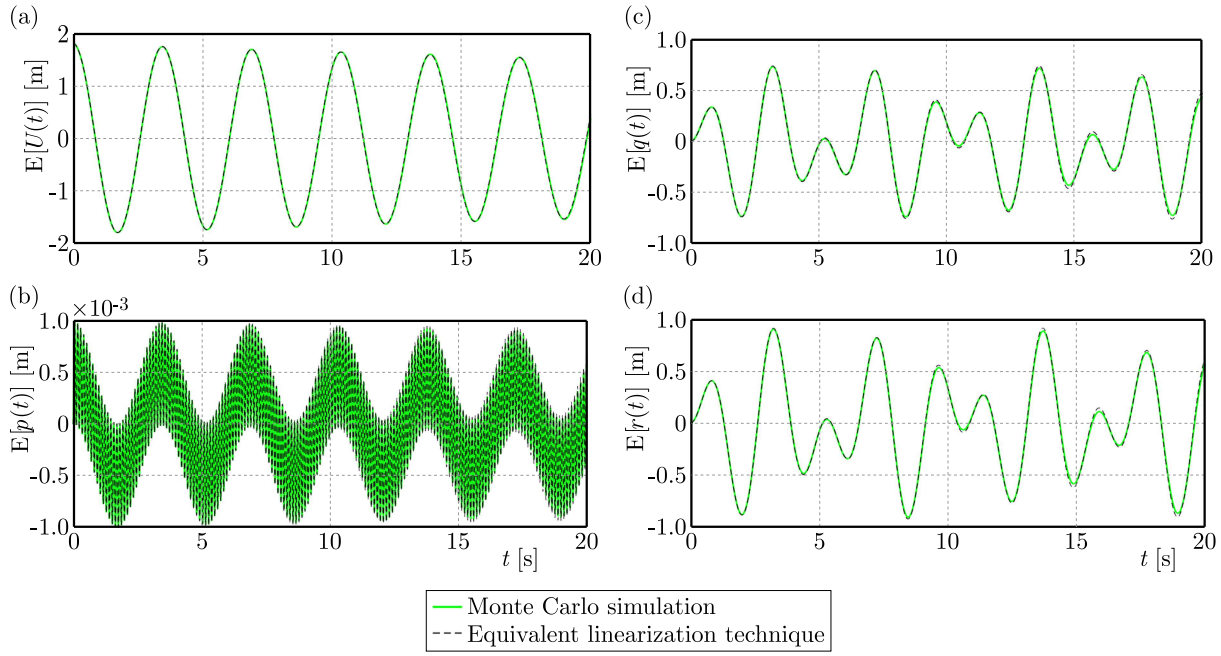


Fig. 5. Comparison of expected values of particular random state variables

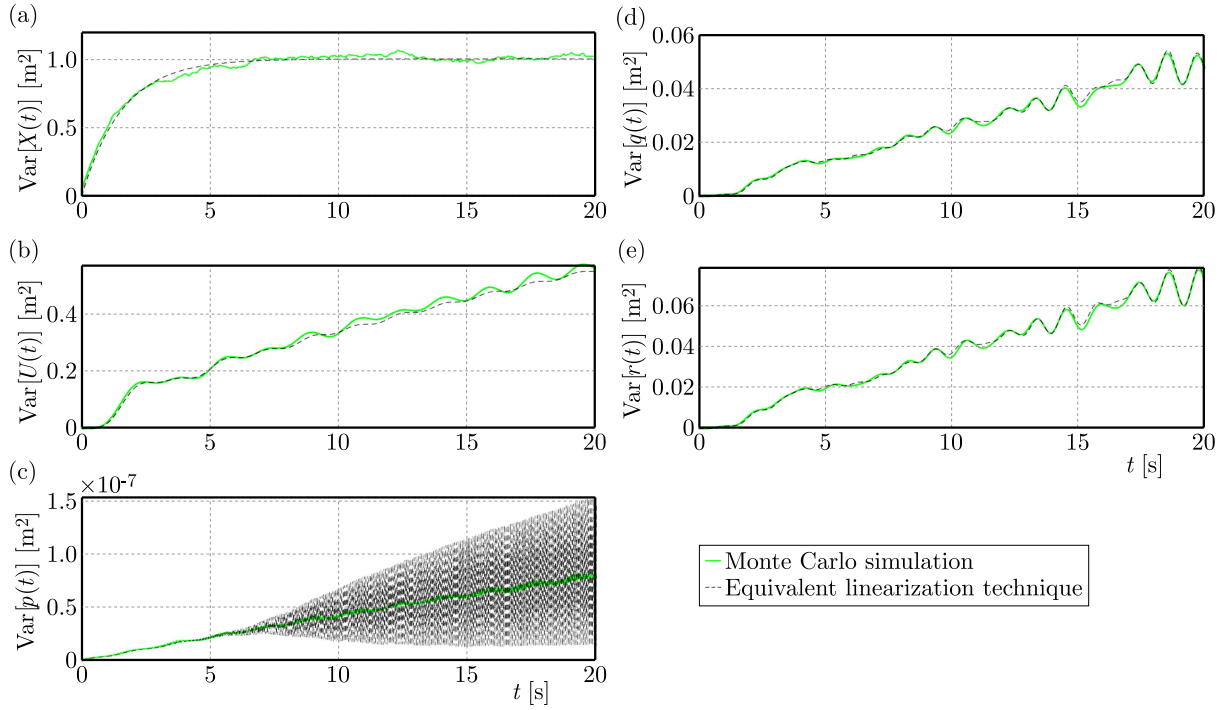


Fig. 6. Comparison of variances of particular random state variables

simulation and, furthermore, the time needed to conduct nonlinear analysis is significantly shorter. This fact together with easy application of this approach in numerical computations presents the opportunity to create a tool that can be very useful at the stage of designing structures with guy lines. A small-sag cable model more closely corresponds to the actual behaviour of the guy line in comparison to an elastic string. Additionally its 3D response under the wind excitation in various direction gives a possibility for deeper examination of the problem of random vibrations in cable systems.

References

1. Atalik T.S., Utku S., 1976, Stochastic linearization of multi-degree-of-freedom nonlinear systems, *Earthquake Enginering Structures Dynamics*, **4**, 411-420
2. Biliszczuk J., Hawryszków P., Teichgraeber M., 2021, SHM system and a FEM model-based force analysis assessment in stay cables, *Sensors*, **21**, 6, 1927
3. Caracoglia L., Zuo D., 2009, Effectiveness of cable networks of various configurations in suppressing stay-cable vibration, *Engineering Structures*, **31**, 12, 2851-2864
4. Chen Z., Yu Y., Wang X., Wu X., Liu H., 2015, Experimental research on bending performance of structural cable, *Construction and Building Materials*, **96**, 279-288
5. Georgakis C.T., Taylor C.A., 2005, Nonlinear dynamics of cable stays. Part 2: Stochastic cable support excitation, *Journal of Sound and Vibration*, **281**, 3-5, 565-591
6. Ha M.-H., Vu Q.-A., Truong V.-H., 2018, Optimum design of stay cables of steel cable-stayed bridges using nonlinear inelastic analysis and genetic algorithm, *Structures*, **16**, 288-302
7. Irvine H.M., 1981, *Cable Structures*, The MIT Press, Cambridge, Massachusetts and London
8. Larsen A., Larose G.L., 2015, Dynamic wind effects on suspension and cable-stayed bridges, *Journal of Sound and Vibration*, **334**, 2-28
9. Li J., Chen J., 2009, *Stochastic Dynamics of Structures*, John Wiley & Sons
10. Proppe C., Pradlwarter H.J., Schuëller G.I., 2003, Equivalent linearization and Monte Carlo simulation in stochastic dynamics, *Probabilistic Engineering Mechanics*, **18**, 1, 1-15
11. Roberts J.B., Spanos P.D., 1990, *Random Vibration and Statistical Linearization*, John Wiley and Sons, New York
12. Socha L., 2007, *Linearization Methods for Stochastic Dynamic Systems*, Springer, Berlin Heidelberg
13. Shi H., Salim H., 2015, Geometric nonlinear static and dynamic analysis of guyed towers using fully nonlinear element formulations, *Engineering Structures*, **99**, 492-501
14. Weber H., Kaczmarczyk R., Iwankiewicz R., 2121, Non-linear response of cable-mass-spring system in high-rise buildings under stochastic seismic excitation, *Materials*, **14**, 22, 6858
15. Xue S., Li X., Liu Y., 2022, Advanced form finding of cable roof structures integral with supporting frames: Numerical methods and case studies, *Journal of Building Engineering*, **60**, 105204
16. Zhang W., Lu X., Wang Z., Liu Z., 2021, Effect of the main cable bending stiffness on flexural and torsional vibrations of suspension bridges: Analytical approach, *Engineering Structures*, **240**, 112393
17. Zhu L., Chen T., Chen L., Lu Z., Hu X., Huang X., 2023, Experimental testing and residual performance evaluation of existing hangers with steel pipe protection taken from an in-service tied-arch bridge, *Applied Sciences*, **13**, 19, 11070

Manuscript received November 30, 2023; accepted for print January 8, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 321-335, Warsaw 2024
https://doi.org/10.15632/jtam-pl/186041

STUDY OF A HORIZONTAL SEAT SUSPENSION WITH A MODEL OF THE SEATED HUMAN BODY AND ENERGY RECOVERY BRAKING SUBSYSTEM[1]

Igor Maciejewski, Sebastian Pecolt, Andrzej Blazejewski, Bartosz Jereczek, Tomasz Krzyzynski

Koszalin University of Technology, Faculty of Mechanical Engineering, Koszalin, Poland

e-mail: igor.maciejewski@tu.koszalin.pl; sebastian.pecolt@tu.koszalin.pl; andrzej.blazejewski@tu.koszalin.pl; bartosz.jereczek@tu.koszalin.pl; tomasz.krzyzynski@tu.koszalin.pl

This article explains the mechanics, control strategies and main applications of a new concept which achieves a balance between energy saving and driver comfort. A physical and mathematical model of a suspension system with energy recovery is presented. It shows practical implementation of the BLDC braking system with energy recovery in a horizontal seat suspension and a detailed simulation analysis of their features and performance. The research involves a specific solution, with a specific BLDC motor, and experimental tests on a laboratory stand. The results of the simulation study using a simplified biomechanical model and experimental studies with human participation are presented.

Keywords: biomechanical model, seat suspension, energy recuperation

1. Introduction

One of the major challenges facing the world today is the energy crisis, which affects various sectors and regions. To address this issue, researchers are exploring new ways of storing and converting electricity which are more efficient and eco-friendlier. Some of the current research topics include developing better batteries (Zhang *et al.*, 2022), reducing power consumption (Farghali *et al.*, 2023), and implementing energy recovery systems (ERS) (Cipoletta *et al.*, 2021; Alhajri *et al.*, 2021). The ERS is mainly designed for the automotive industry (Gabriel-Buenaventura and Azzopardi, 2015; Bravo *et al.*, 2018; Salman *et al.*, 2018) and aims to convert some of kinetic energy into electrical energy. This allows one to power small devices such as sensors and microcontrollers, support the main power source, or store the recovered energy, which helps to lower operational costs. Such a process is called the energy harvesting.

The energy harvesters are devices that can capture and convert different forms of energy into electricity. They are often associated with renewable energy sources, such as solar, wind, thermal and geothermal energy. These sources are widely used in outdoor environments, but they depend on the availability and intensity of natural phenomena (Muscat *et al.*, 2022; Mescia *et al.*, 2014; Sudevalayam and Kulkarni, 2011). Another type of energy harvesters is based on mechanical vibrations, which are ubiquitous in indoor environments, where many machines and devices operate. These harvesters use transducers that transform kinetic energy of vibrations into electrical energy. Such transducers can be classified into three main categories: piezoelectric, electrostatic and electromagnetic, depending on the physical principle of their conversion process.

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

Piezoelectric vibrational energy harvesters have attracted a lot of interest in the recent years, and they have been applied in various fields, such as automotive, biomechanics, human body motion, architecture and construction engineering. These harvesters exploit the piezoelectric effect which is the generation of an electric potential by a material when it is subjected to a mechanical strain. The piezoelectric material deforms when it is exposed to vibrations, and this creates a charge imbalance that produces voltage (Hu *et al.*, 2013). However, piezoelectric harvesters have a limitation in their frequency range, as they work best at frequencies above 1 kHz, while most environmental vibrations are in the range of 1 Hz-100 Hz (Halm *et al.*, 2016). Therefore, piezoelectric harvesters need to be tuned to match the frequency of the vibration source (Priyn *et al.*, 2017). However, mechanical vibrations generated by technical devices, such as engines, can become a source of energy for these transducers.

Piezoelectric harvesters are devices that convert mechanical vibrations into electrical energy. They have benefits of being self-powered, producing relatively high voltage, having small size, and having a high efficiency of energy conversion. However, piezoelectric materials have some drawbacks, such as variable power output over time, which affects their performance (Fastier-Wooller *et al.*, 2022), and possible damage due to brittleness of the material (Matak, 2013; Beeby *et al.*, 2006). Electrostatic transducers are devices that generate voltage by changing their internal capacitance under an applied force (Zhu *et al.*, 2010b). Electrostatic technology includes electret-based vibration energy harvesting using MEMS, which are micro-scale mechanical systems, and triboelectric energy harvesting (Toshiyyoshi *et al.*, 2019). Electrostatic transducers usually need a high-voltage power source or an electret, which is a material that has a permanent electric charge or dipole polarization, to create strong electric fields that drive the electric current, which makes these systems more complex (Zhu *et al.*, 2010a). Moreover, since the gap between capacitor plates or surfaces is typically in a millimetre range, they are not suitable for higher amplitude vibrations without an additional system that adapts the input vibration motion to the appropriate amount of amplitude (Beeby *et al.*, 2006).

Electromagnetic vibrational energy harvesters (EVEH) are devices that convert low-frequency vibrations into electrical power using the principle of electromagnetic induction. They have a simple structure and can operate in various environments, which makes them attractive for many applications (Araujo and Nicoletti, 2015). For example, EVEH can be used to power wireless sensors, wearable devices, biomedical implants, or environmental monitoring systems. The basic mechanism of EVEH is that a magnet moves relative to a coil and induces an electric current according to Faraday's law (Bouendeu *et al.*, 2011). Another way to achieve electromagnetic energy conversion is by using inverse magnetostrictive materials which change their magnetization state when subjected to mechanical stress. By applying a bias magnetic field with permanent magnets, the strain-induced magnetic flux variation can be captured by a coil and converted into electricity (Akinaga, 2020; Ueno, 2019). Moreover, some vibration energy harvesters combine piezoelectric and electromagnetic effects to enhance their performance. Depending on their configuration, they can be classified into mono-stable, bi-stable, multi-stable, magnetic-plucking (contactless), or hybrid piezoelectric-electromagnetic energy harvesters (Jiang *et al.*, 2021).

Vibrations produced by devices in operation can be a source of energy replenishment. Researchers employ additional components to capture this energy, such as the piezoelectric vibrator proposed in work (Wang, 2020) as an energy converter from track vibrations caused by vehicle movement. The author developed a dynamic model of a vehicle coupled vertically to estimate displacements that affect the piezoelectric element in charge of energy recovery. The researcher then performed simulation tests and concluded that larger displacement amplitudes increase the amount of energy recovered. The vehicle speed and position of the piezoelectric elements also influence the maximisation of the energy output.

Another way to enhance the energy efficiency of automobiles was proposed by Hassan Fathabadi in his article (Fathabadi, 2019). He suggested two modifications: embedding electric coils in shock absorbers to capture and convert vibration energy into a steady DC voltage and adding a wind turbine to the vehicle condenser. The author showed that those two modifications could increase the energy production of electric vehicles and extend their travel range.

Alternatively, energy can be harvested by using the Energy Restore Braking System (ERBS) (Li *et al.*, 2021; Liu and Zhang, 2021). This concept involves converting kinetic energy of motors into electrical energy during deceleration (Taut *et al.*, 2013). This technique is commonly applied in electric vehicles. However, it often needs additional components in the system such as DC-DC converters, which are devices that convert the direct current (DC) from one voltage level to another (Kim, 2011; Onar and Khaligh, 2012), super-capacitors, which are high-capacity capacitors that can store and release large amounts of energy quickly and release it when connected to a chosen circuit (Naseri *et al.*, 2017; Song *et al.*, 2014), or gear shifts, which are mechanisms that change the speed ratio between the motor and wheels (Yang *et al.*, 2007). These components add to the weight and complexity of the system.

Another way to make energy recovery systems more simple is to use a single-stage converter that controls the BLDC motor, as suggested by Godfrey and Sankaranarayanan (2008). This method can switch to the regenerative braking mode by sending switching pulses in a specific sequence. This method does not need extra power converters, which is a benefit compared to other solutions. The authors of (Godfrey and Sankaranarayanan, 2008) present different switch topologies, such as H-bridge, half-bridge and full-bridge, and plugging combined to create a new braking strategy. The switch topologies determine how the current flows through motor windings and how the back electromotive force is generated. The simulation and experimental tests were done to show the effectiveness of the proposed solution.

On the contrary, this article explores how to combine the driver seat suspension system with a special BLDC motor braking system that can turn horizontal vibrations of the driver seat suspension into electric power. The stored power can be used for different vehicle subsystems as needed. This solution also helps one to optimize the active seat suspension systems that use electric motors to create vibration damping force. The idea is to use the motor as a generator to produce this force. The proposed suspension system can be tested using human biomechanical models (Maciejewski *et al.*, 2023). The model gives quick results on the power and vibration levels that affect the driver body parts, which allows for evaluating and comparing different suspension systems.

The first part of the article explains the mechanics, control strategies and main applications of this new concept, which achieves a balance between energy saving and passenger comfort. A physical model and a mathematical model of the suspension system with energy recovery are presented. This approach is based on multiple objectives aiming to accomplish. One of them is to enhance comprehension of complex interactions involved in the seat suspension that can reduce horizontal vibrations affecting comfort and health of work machine operators. Another one is to explore the efficiency and feasibility of incorporating energy recovery mechanisms into these systems, considering both the environmental demands of energy efficiency and the psychophysical health of drivers, operators and passengers. The next subsections of the article present practical implementation of the BLDC braking system with energy recovery in horizontal seat suspension and a detailed simulation analysis of their features and performance. Further research involves simulation analysis of a specific solution, with a specific engine (BLDC motor), and experimental tests on a laboratory stand. The data gathered during experiments can be used to broaden the knowledge about this type of systems and their practical applications. It will suggest the direction of further research, including the use of suitable biomechanical models.

2. Physical and mathematical model

Figure 1 shows the physical model of an active seat suspension and simple biomechanical human body model. The upper part of the human body is modelled as a lumped three mass system (roughly corresponds to the pelvis, torso and head) (Maciejewski *et al.*, 2022b). That inertial system responds to the external excitation x_s. In detail, the first mass m_1 represents the seat frame with pelvis, the cushions and motor inertia, the second mass m_2 represents the body part on the seat back rest, i.e the torso and the third mass m_3 represents the body part that moves freely, i.e. the head. Its mass m_3 has no contact with the backrest. The stiffnesses c_{12}, c_{23}, c_2

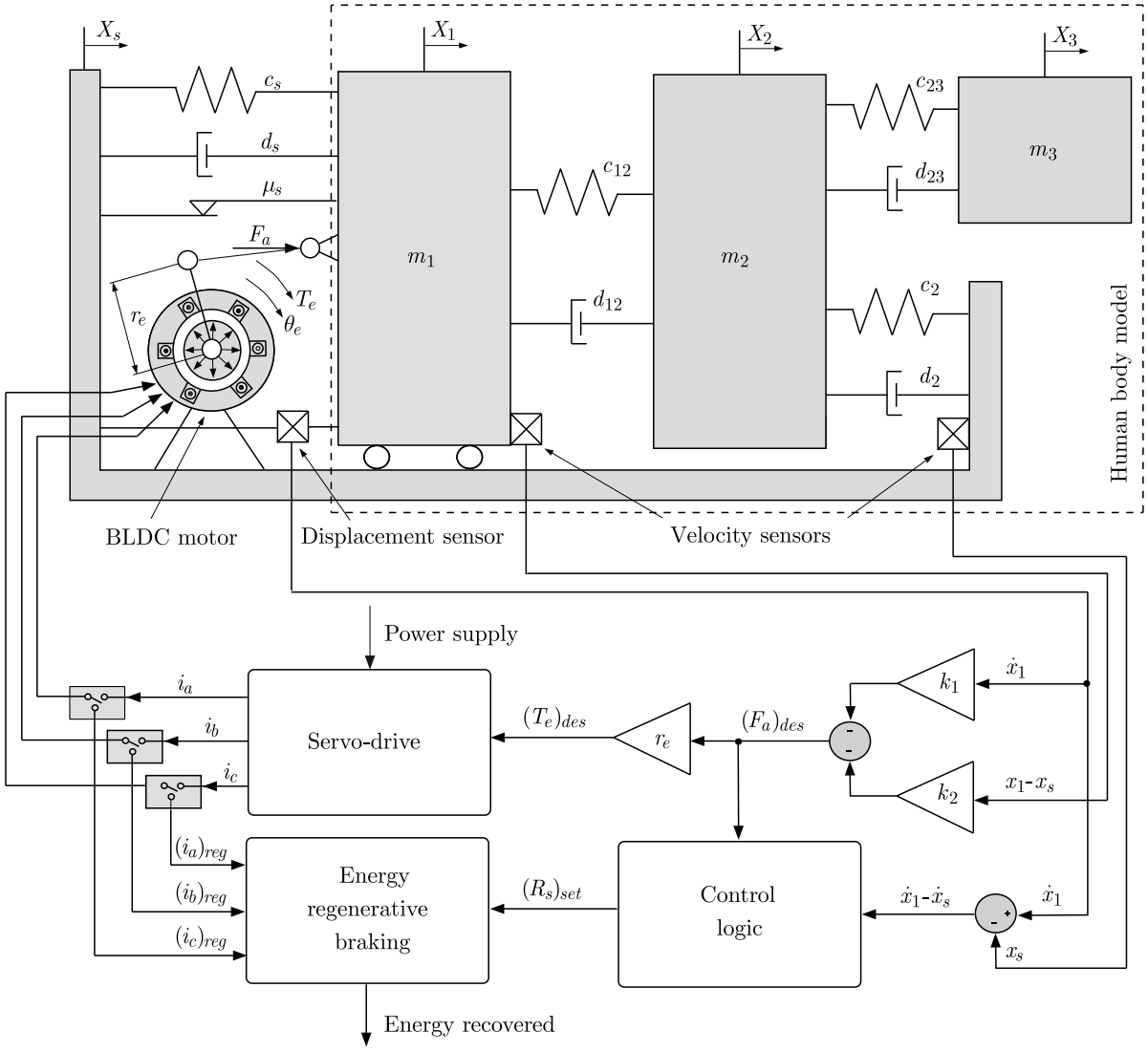


Fig. 1. Physical model of an active seat suspension with the BLDC motor

and damping coefficients d_{12}, d_{23}, d_2 capture visco-elastic properties of the human body tissues. Due to topology and simplicity of the model to identify the above-mentioned parameters, it was necessary to conduct experimental research with human participation and then the carry out the optimisation process. During the experiment, the tested individual was subjected to action of test vibration in the sitting position on the cushioned seat, occupied in the way providing the backrest contact. The identification process was conducted minimising by the error between numerically simulated and experimentally measured transmissibility functions over the frequency range of

1 Hz-12 Hz and introducing the Root Mean Square Error method. A detailed description of the model and its parameter sensitivity analysis is given in the work (Maciejewski *et al.*, 2022b).

This model structure and its proper parameters can be used to predict the biodynamical response of the seated human body in the frequency range of 1 Hz to 12 Hz. The model of the suspension system consists of a human body mass that is attached to two tension springs with stiffness c_s and a damper with damping ratio d_s. The friction in the suspension mechanism is modelled by a coefficient μ_s, which is determined experimentally for a specific seat type and described in the work (Jereczek *et al.*, 2022). Modern seat suspension systems are active systems. The aim of the active subsystem is to minimise harmful vibrations of the body mass. The active element here is a brushless DC motor (BLDC). Such an electric actuator generates an active force F_a which, in turn, is proportional to the electromagnetic torque T_e of this motor. The torque T_e is controlled by a servo-drive mode. The equations of motion for the mechanical structure are given by

$$\begin{aligned} m_1\ddot{x}_1 &= -d_s(\dot{x}_1-\dot{x}_s)-c_s(x_1-x_s)-\mu_s(m_1+m_2+m_3)g\,\mathrm{sgn}\,(\dot{x}_1-\dot{x}_s)+F_a \\ &\quad + d_s(\dot{x}_1-\dot{x}_s)+c_s(x_1-x_s) \\ m_2\ddot{x}_2 &= -d_{12}(\dot{x}_2-\dot{x}_1)-c_{12}(x_2-x_1)+d_{23}(\dot{x}_3-\dot{x}_2)+c_{23}(x_3-x_2) \\ &\quad + d_2(\dot{x}_s-\dot{x}_2)+c_2(x_s-x_3) \\ m_3\ddot{x}_3 &= -d_{23}(\dot{x}_3-\dot{x}_2)-c_{23}(x_3-x_2) \end{aligned} \tag{2.1}$$

Active force F_a in the suspension system is provided by three-phase currents (i_a, i_b and i_c) flowing through stator windings in the presence of a magnetic field from permanent magnets. The resulting force coming from an electric motor is therefore defined as the following function

$$F_a = \frac{p\lambda}{r_e}(\Phi_a i_a + \Phi_b i_b + \Phi_c i_c) \tag{2.2}$$

where: p is the number of pole pairs, λ is the amplitude of flux induced by permanent magnets, r_e is the lever arm of the motor, Φ_a, Φ_b and Φ_c are three-phase electromotive forces. The phase electromotive forces are assumed to be trapezoidal for most of the BLDC motors (Krause *et al.*, 2002).

The trapezoidal model is based on the assumption that the winding distribution and the magnetic flux created by permanent magnets generate three trapezoidal back electromotive forces. The back electromotive forces are the voltages induced in the stator windings by the changing magnetic field. The set of equations that describe the actual motor currents can be written in the phase reference frame (*abc* frame) as follows (Krause *et al.*, 2002)

$$\begin{aligned} \dot{i}_a &= \frac{1}{3L_s}[2v_{ab}+v_{bc}-3R_s i_a+\lambda p\dot{\theta}_r(-2\Phi_a+\Phi_b+\Phi_c)] \\ \dot{i}_b &= \frac{1}{3L_s}[-v_{ab}+v_{bc}-3R_s i_b+\lambda p\dot{\theta}_r(\Phi_a-2\Phi_b+\Phi_c)] \\ \dot{i}_c &= -(\dot{i}_a+\dot{i}_b) \end{aligned} \tag{2.3}$$

where: L_s is the inductance of stator windings, v_{ab} and v_{bc} are the phase to phase supply voltages, R_s is the resistance of stator windings. The currents in Eq. (2.3) need to get values that allow reducing the unwanted seat movements. The double-feedback loop system (Maciejewski *et al.*, 2020) is used to calculate the proper, desired active force $(F_a)_{des}$, which is needed to balance the kinematic excitation x_s.

The motor windings currents should produce the desired active force given by the following formula

$$(F_a)_{des} = \begin{cases} -k_1\dot{x}_1-k_2(x_1-x_s) & \text{for} \quad (F_a)_{des}(\dot{x}_1-\dot{x}_s)\geqslant 0 \quad \leftarrow \text{motoring} \\ 0 & \text{for} \quad (F_a)_{des}(\dot{x}_1-\dot{x}_s)<0 \quad \leftarrow \text{braking} \end{cases} \tag{2.4}$$

where $\dot{x}_1$ is the absolute velocity of the suspended body given by the mass acceleration sensor (after integration process) (Fig. 1), $x_1 - x_s$ is the relative displacement of the seat suspension measured by the displacement sensor (Fig. 1) k_1 and k_2 are the coefficients feedback that affect significance of reducing velocity of seat vibration (influence of vibration velocity criterion) and significance in limiting the suspension travel (influence of seat displacement criterion), respectively. Formula (2.4) also shows the conditions for which the damping force is generated in the active suspension cycle ("motoring"). and the force is not generated in the energy recovery cycle ("braking").

3. Energy recovery capabilities during braking mode of an induction motor

The force/velocity possibilities of the suspension system are represented as a velocity versus active force graph containing of four quadrants (Fig. 2). This figure illustrates four-quadrant operation of the horizontal seat suspension driven by an induction motor. In the chosen horizontal

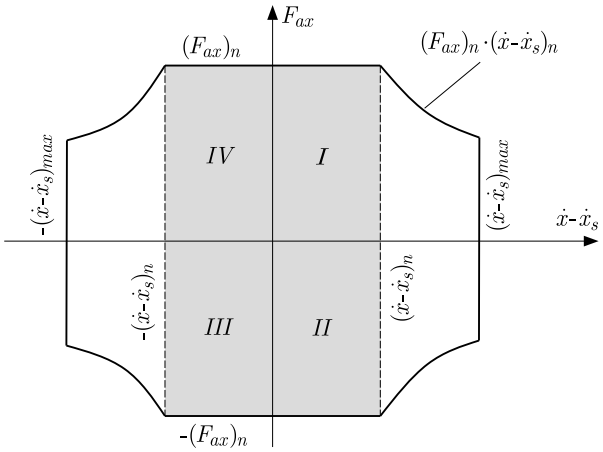


Fig. 2. Four-quadrant operation of the horizontal seat suspension driven by an induction motor

x direction, a constant force $(F_{ax})_n$ is given in grey colour region from 0 to nominal velocity $\pm(\dot{x}_1 - \dot{x}_s)_n$. The rest region (white colour) indicates a significant decrease of the force $(F_{ax})_n$ while an increase of velocity from the nominal $(\dot{x}_1 - \dot{x}_s)_n$ to the maximum value $(\dot{x}_1 - \dot{x}_s)_{max}$. Such a region is limited by a constant power level $(F_{ax})_n(\dot{x}_1 - \dot{x}_s)_n$ due to lowering of the motor magnetic flux. In the first (I) and third quadrant (III), the active force F_{ax} and the velocity $\dot{x}_1 - \dot{x}_s$ have the same signs, indicating the driving mode since the electric force is in the direction of motion. In the second (II) and fourth (IV) quadrant, the active force F_{ax} is opposite to the velocity $\dot{x}_1 - \dot{x}_s$, therefore the braking mode of an induction motor is applied. Operation in these quadrants means that the kinetic energy of the induction motor coupled to a mechanical load can be transformed into the electric energy. The control algorithm that corresponds to possible energy transfer between mechanical and electrical subsystems is defined as follows

$$(R_s)_{set} = \begin{cases} 0 & \text{for} \quad (F_{ax})_{des}(\dot{x} - \dot{x}_s) \geqslant 0 \quad \leftarrow \text{motoring} \\ \max\Big[0, \min\Big(\Big|\dfrac{(F_{ax})_{des}}{\dot{x} - \dot{x}_s}\Big|, k_{max}\Big)\Big]|\dot{x} - \dot{x}_s| & \text{for} \quad (F_{ax})_{des}(\dot{x} - \dot{x}_s) < 0 \quad \leftarrow \text{braking} \end{cases} \tag{3.1}$$

where: $(R_s)_{set}$ is the controllable external resistance of the induction motor, k_{max} is the constant coefficient representing the maximum braking force of the induction motor, the same BLDC motor working as a generator. Each constant model parameter of the seated human body together

with the horizontal seat suspension and energy recovery braking subsystem is specified in Appendix A.

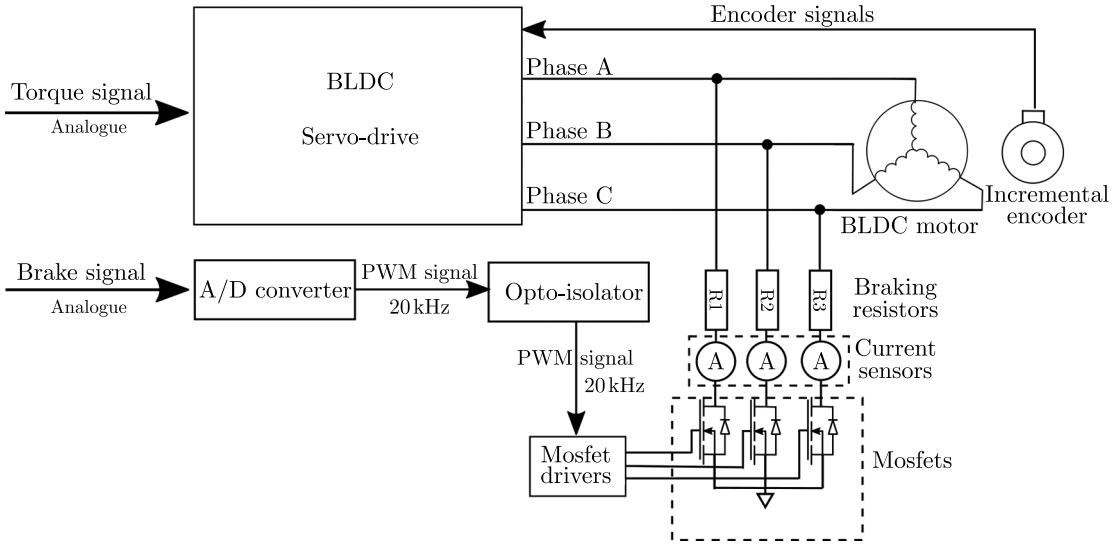


Fig. 3. Regenerative braking system for motor control

Physically, the braking system of the BLDC motor, realising energy transfer, works using resistors and transistors shown in Fig. 3. The system controller sends an analogue "braking" signal to initiate the braking process. This is happening when the controller gets a signal from seat velocity sensor and a torque signal, and calculates according to Eq. (3.1), which means second (II) and fourth (IV) quadrant (Fig. 2). The "braking" signal is converted into a digital PWM form with a frequency of 20 kHz. The PWM signal is then isolated from the transistors using an opto-isolator circuit. The transistors are connected to three braking resistors R_1, R_2 and R_3 which are also connected to the three motor phases A, B and C. When the transistors are turned on by the PWM signal, the motor phases are shorted through the resistors, which creates a braking torque on the BLDC motor. The currents flowing through the motor windings are measured by current sensors based on the Hall effect, which allows us to calculate the power dissipated by the resistors. In the same way, the available energy is determined.

4. Experimental versus simulation results

In the last phase of the research, experimental verification of the correctness of the model, simulation results and the effectiveness of the proposed energy recovery concept in such a system was carried out. The effectiveness and efficiency of the energy harvesting in the horizontal suspension systems was evaluated using the experimental set-up shown in Figs. 4a and 4b. As the physical model assumed, the suspension system (seat suspension system – Fig. 4a) was attached to a base platform (vibrating platform – Fig. 4a), driven by a PMSM motor (vibrations source motor PMSM – Fig. 4b) that produced a random vibration signal with a programmable electro-hydraulic shaker. Finally, the oscillatory seat motion transited to the mass (mass load – Fig. 4a, respectively mass m_1 – Fig. 1) was induced by an active force generated by the PMSM motor through a two-link mechanism. This mechanism transformed the rotational motion of the motor into the translational motion of the suspension system in the longitudinal direction. The accelerometer (platform acceleration sensor – Fig. 4a) was utilised to measure the input signal that was random vibration having a frequency between 0.5 Hz and 12.5 Hz. The data obtained from this accelerometer was recorded by a PC-based data acquisition system with the sampling

time of 1 ms. The output signal was measured by an accelerometer fixed to the mass (mass acceleration sensor – Fig. 4a). The recorded acceleration signals from both accelerometers were simultaneously digitally integrated to obtain velocities of the platform and mass. The rigid mass (mass load – Fig. 4a) was placed on the upper part of the suspension mechanism to load the system to simulate the driver (operator) presence during investigation. The element of the active seat suspension was the brushless motor (BLDC motor – Fig. 4b). Its task was, on the one hand, to generate an active vibration damping force in accordance with the selected algorithm in the control strategy (one of the control algorithms developed by the authors was presented in the work (Maciejewski *et al.*, 2022a), and, on the other hand, to obtain seat vibration energy and convert it into electrical energy, in accordance with the energy acquisition strategy. On the test stand, the energy of the seat movement (longitudinal vibrations) was transferred to the motor through a system of rigid rods and an eccentric system (centrifugal transfer of motion from the seat to the BLDC motor – Fig. 4b). To measure the amplitude of longitudinal seat vibrations directly, a dedicated displacement sensor was used (seat displacement sensor – Fig. 4a). Mounting this sensor enabled measuring the relative displacement of the platform and the seat.

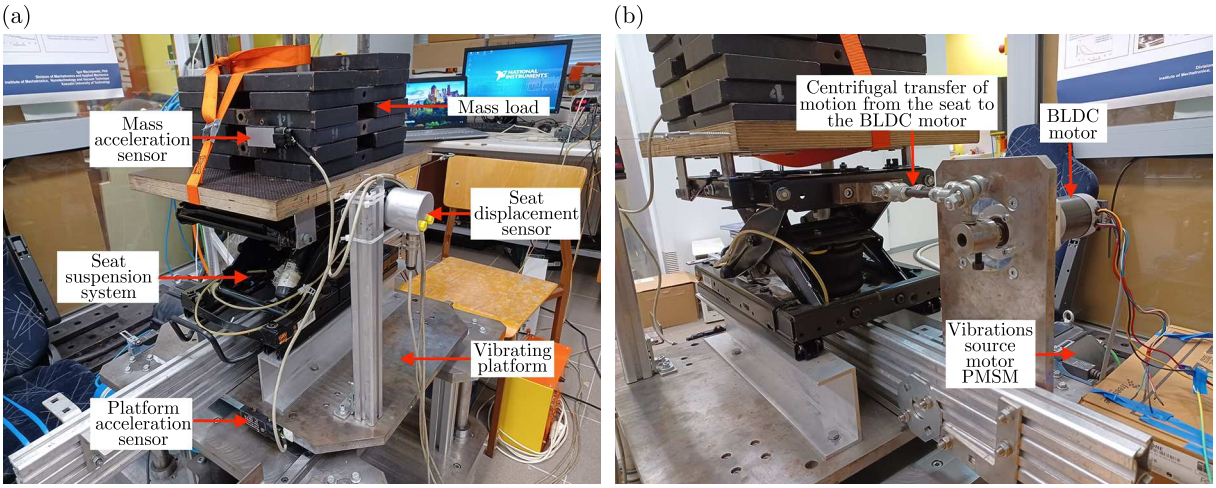


Fig. 4. Actual, overall view of: (a) the test rig with basic sensors, (b) the test stand with a BLDC motor (energy recovery) and a seat vibration motor

The aim of the experimental study was to examine the effect of the tested regenerative system on the suspension performance. Two common factors were selected for the analysis. The first one was the SEAT factor, which measured the seat isolation efficiency. The second one was the suspension travel (relative displacement), which evaluated the seat dynamic response (Maciejewski *et al.*, 2014). Generally, when the SEAT factor was higher than 1, the vibrations were amplified and transmitted from the road to the driver or operator. When the SEAT factor was lower than 1, the vibration isolation improved as the SEAT decreased. For the second factor, a smaller value was preferable. It ensured that the seat did not reach its movement limits when driving on rough roads. The test was carried for a chosen mass value (Mass load 80 kg). Figure 5 shows the tested person of weight 80 kg on the test rig. The same person was tested on passive, active and regenerative seat suspension configuration.

The comparative results of the passive, active and energy regenerative suspension are presented in Table 1. The data in this table show that the use of an energy recovery system generally reduces the vibration-insulating properties of the suspension in caparison to the active one. However, the reduction is approximately 18.1% for this particularly load mass, calculated on the basis of the SEAT factor. However, the dynamic response, calculated on the basis of the suspension travel value, is reduced by approximately 23.6% for the weight of 80 kg. At the

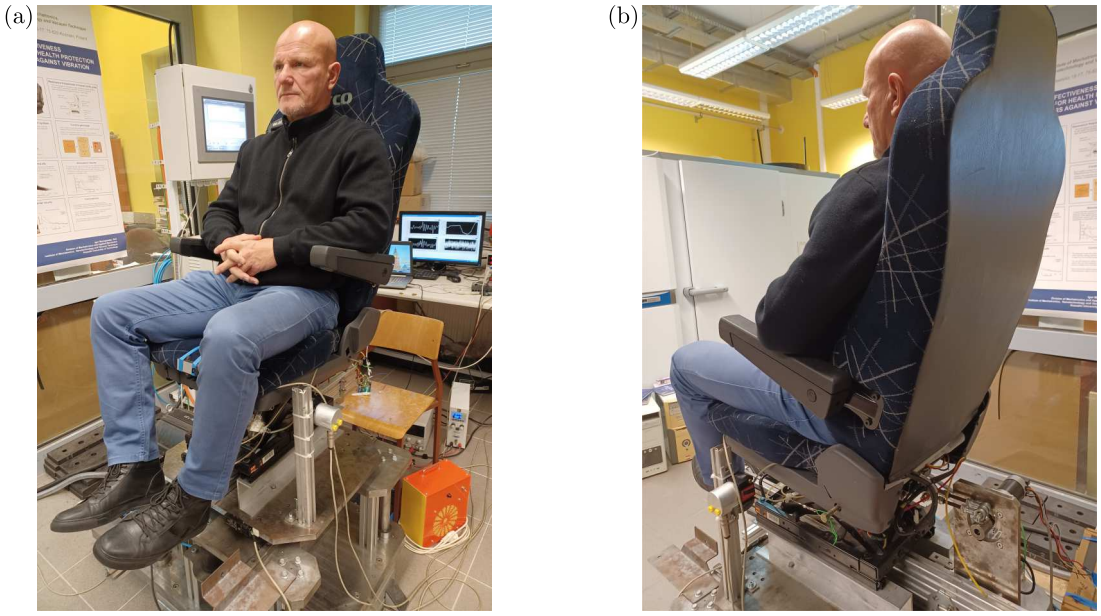


Fig. 5. The view of: (a) the front test rig with the tested person, (b) the rear part with the regenerative system

Table 1. Measured SEAT factors and suspension travels of the passive, active and regenerative suspension system for the mass load of 80 kg

Human weight	Horizontal seat suspension					
	Passive		Active		Regenerative	
	SEAT factor	Suspension travel	SEAT factor	Suspension travel	SEAT factor	Suspension travel
80 kg	0.877	29.3 mm	0.629	33.5 mm	0.743	25.6 mm

same time, as the values in the first two columns of Table 1 show, such a system still has better vibration isolation properties and dynamic response than the passive one.

At the same time, simulation tests were carried out using a biomechanical human model and a suspension system model. The results of the simulation using the biomechanical model are presented in Table 2. As can be seen from the values presented there, the model gives the same assessment of the vibro-isolation with the SEAT coefficient and the suspension travel value. All values obtained from the model are larger than those from the experiment. The greatest compliance is found in the case of the active suspension system. The largest relative error of the model in the case of SEAT was max 10%. However, in the case of suspension travel, the error reached 15%.

Table 2. Simulated SEAT factors and suspension travels of the passive, active and regenerative suspension system for a mass load of 80 kg

Human weight	Horizontal seat suspension					
	Passive		Active		Regenerative	
	SEAT factor	Suspension travel	SEAT factor	Suspension travel	SEAT factor	Suspension travel
80 kg	0.923	32.7 mm	0.630	35.5 mm	0.816	29.3 mm

In detail, the range in which an energy recovery suspension system works better than a passive system can be indicated by analysing power spectral densities (Fig. 6a) and transmissibility

functions graphs (Fig. 6b). The curves of both types of functions, for this case of mass load, indicate the frequency limit of 3 Hz. Below this frequency, the active and regenerative system represents similar vibro-isolation properties. The graph shows slightly better properties of the active system. In that frequency limit, the passive system demonstrates clearly inferior properties. In the case of frequencies above 3 Hz, the two systems, passive and regenerative, show similar vibration isolation properties and much worse than for the active one. However, as the transmissibility function curve for the regenerative system shows (Fig. 6b), it is below the value of 1 throughout the frequency range. Therefore, it significantly reduces vibrations coming from the ground or road.

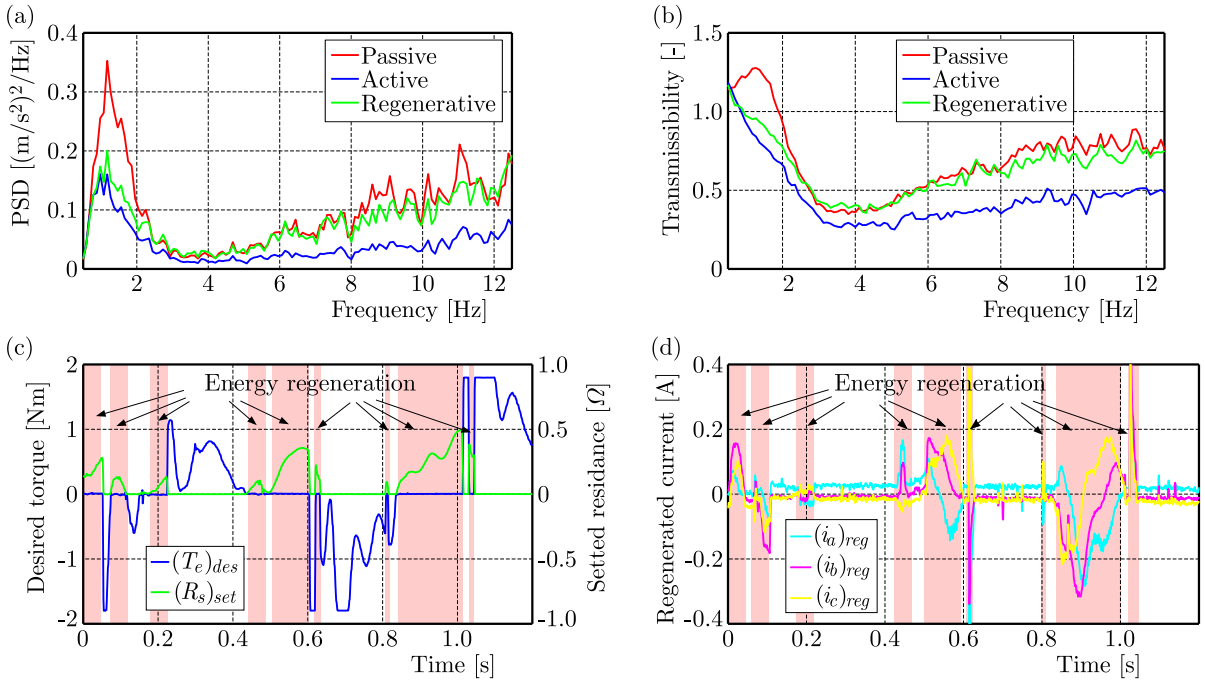


Fig. 6. Measured power spectral densities (a) and transmissibility functions (b) of the passive, active and regenerative suspension system, time histories of the desired motor torque versus the set resistance (c) and the corresponding regenerated phase currents (d)

At the same time on the test-rig, other essential parameters related to the regenerative seat suspension were measured (Figs. 6c and 6d). Figure 6c shows the desired torque generating during the damping period defined by the "motoring" condition in Eq. (2.4). In this figure, the torque values appear by multiplication by r_e the desired force (Fig. 1). The damping periods are marked with white vertical stripes along the selected 1.2 s time sample. On other the hand, pink stripes in Fig. 6c indicate the regenerative period. Particularly, it presents the resistance settled when the motor phases are shortened through the breaking resistors (Fig. 3). The current sensors shown in Fig. 3 measured the current values presented in Fig. 6d in pink strips time periods.

Figures 7a-f show the results of simulation of the passive, active and regenerative suspension system. In Fig. 7a, the power spectral densities of the three systems are compared, showing that the active system has the lowest vibration level (close to experimental results). In the case of regenerative and passive ones, the model returned higher values. The frequency limit 3 Hz is visible like in the case of the experiment. Above this limit, all simulated systems gave very similar results, differently than in the experiment. In Fig. 7b, the transmissibility functions of the three systems are plotted, indicating a similar to power density functions isolation performance of all systems. In this case, however, it should be noted that the regenerative system, in the range up to 3 Hz, reaches values greater than 1, contrary to what the experiment showed. In Fig. 7c, the

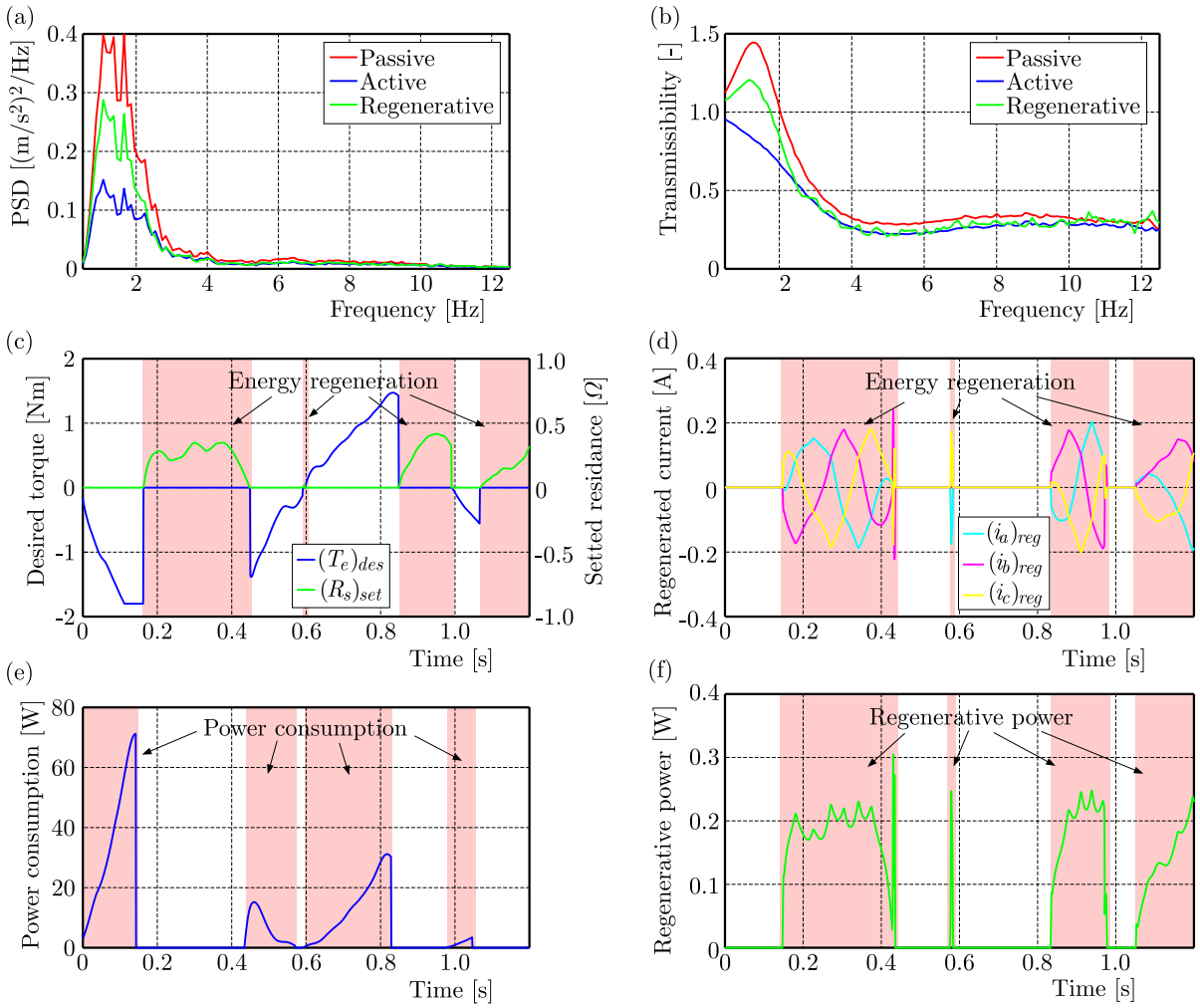


Fig. 7. Simulated power spectral densities (a) and transmissibility functions (b) of the passive, active and regenerative suspension system, time histories of the desired motor torque versus the set resistance (c) and the corresponding regenerated phase currents (d), power consumption (e) and the obtained regenerative power of the electric motor (f)

desired motor torque and the set resistance are shown, demonstrating that the motor torque is zero in the energy regeneration periods contrary to the resistance (pink strips). The opposite is true in the ranges marked by white strips. This confirms the validity of the regenerative model. In Fig. 7d, the phase currents of the electric motor are displayed, revealing that they are in phase with the motor torque. In Fig. 7e, the power consumption of the electric motor is calculated, proving that it is lower than the active system. In Fig. 7f, the regenerative power of the electric motor is estimated, showing that it can recover some energy from the suspension vibration.

5. Conclusions

The experimental study demonstrated that the energy regenerative suspension system can improve the suspension travel and reduce the seat movement limits compared to the passive system, while sacrificing some of the vibration isolation efficiency compared to the active system. The power spectral density and transmissibility function graphs showed the frequency ranges where the regenerative system performed better or similar to the passive system. The test-rig experiment demonstrated the feasibility of the proposed regenerative seat suspension system. The

torque and current measurements confirmed the energy harvesting potential during regenerative periods, as well as the damping performance during motoring periods. The results also validated the mathematical model and the control strategy of the system. In the case of biomechanical modelling, the proposed model is a simplified representation of the human body response to seat vibration, based on a three mass system. The model parameters were obtained by experimental tests and optimisation methods, using the Root Mean Square Error to minimise the discrepancy between simulation and measurement. The model can be used to evaluate the vibration comfort and performance of different seat suspension designs and control strategies. In this work, the operation of the seat suspension together with a human being presence was simulated. The simulation results of the passive, active and regenerative suspension systems were consistent with the experimental ones in terms of vibration reduction and energy recovery. The active system had the best performance in terms of vibration isolation, while the regenerative system had the advantage of lower power consumption and partial energy regeneration. The regenerative model was validated by comparison of the motor torque, resistance, phase currents and power of the electric motor. The simulation results also showed the influence of the frequency limit on the transmissibility and power spectral density functions of the three systems. The results suggest that the regenerative system can be a viable alternative to the active system, especially for applications where energy saving is important and vibration isolation is not critical.

Appendix A. Constant model parameters of a seated human body with horizontal seat suspension and energy recovery braking subsystem

Parameter	Value	Unit
Human body model		
Mass of pelvis m_1	25.18	kg
Mass of torso m_2	46.77	kg
Mass of head m_3	8.04	kg
Damping between pelvis and torso d_{12}	582	Ns/m
Stiffness between pelvis and torso c_{12}	15815	N/m
Damping between torso and head d_{23}	59	Ns/m
Stiffness between torso and head c_{23}	5809	N/m
Damping between torso and backrest d_2	5	Ns/m
Stiffness between torso and backrest c_2	50	N/m
Model of horizontal seat suspension		
Damping of suspension mechanism d_s	1000	Ns/m
Stiffness of suspension mechanism c_s	10000	N/m
Friction coefficient of suspension mechanism μ_s	0.05	–
Lever arm of motor r_e	0.045	m
Model of electrical subsystem		
Number of pole pairs p	3	–
Amplitude of flux induced by permanent magnets λ	0.00733	Vs
Inductance of stator windings L_e	0.0001	H
Resistance of stator windings R_e	0.0675	Ω
Coefficient of maximum braking force k_{max}	1	–

References

1. Akinaga H., 2020, Recent advances and future prospects in energy harvesting technologies, *Japanese Journal of Applied Physics*, **59**, 110201
2. Alhajri I.H., Gadalla M.A., Elazab H.A., 2021, A conceptual efficient design of energy recovery systems using a new energy-area key parameter, *Energy Reports*, **7**, 1079-1090
3. Araujo M., Nicoletti R., 2015, Electromagnetic harvester for lateral vibration in rotating machines, *Mechanical Systems and Signal Processing*, **52**, 685-699
4. Beeby S., Tudor M.J., White N., 2006, Energy harvesting vibration sources for microsystems applications, *Measurement Science and Technology*, **17**, R175"R195
5. Bouendeu E., Greiner A., Smith, P., Korvink J., 2011, Design synthesis of electromagnetic vibration-driven energy generators using a variational formulation, *Journal of Microelectromechanical Systems*, **20**, 466-475
6. Bravo R.R.S., De Negri V.J., Oliveira A.A.M., 2018, Design and analysis of a parallel hydraulic – pneumatic regenerative braking system for heavy-duty hybrid vehicles, *Applied Energy*, **225**, 60-77
7. Cipolletta G., Delle Femine A., Gallo D., Luiso M., Landi C., 2021, Design of a stationary energy recovery system in rail transport, *Energies*, **14**, 9, 2560
8. Farghali M., Osman A.I., Mohamed I.M.A., Chen Z., Chen L., *et al.*, Strategies to save energy in the context of the energy crisis: a review, *Environmental Chemistry Letters*, **21**, 2003-2039
9. Fastier-Wooller J.W., Vu T.-H., Nguyen H., Nguyen H.-Q., Rybachuk M., *et al.*, 2022, Multimodal fibrous static and dynamic tactile sensor, *ACS Applied Materials and Interfaces*, **14**, 27317-27327
10. Fathabadi H., 2019, Recovering waste vibration energy of an automobile using shock absorbers included magnet moving-coil mechanism and adding to overall efficiency using wind turbine, *Energy*, **189**, 116274
11. Gabriel-Buenaventura A., Azzopardi B., 2015, Energy recovery systems for retrofitting in internal combustion engine vehicles: A review of techniques, *Renewable and Sustainable Energy Reviews*, **41**, 955-964
12. Godfrey J.A., Sankaranarayanan V., 2018, A new electric braking system with energy regeneration for a BLDC motor driven electric vehicle, *Engineering Science and Technology, an International Journal*, **21**, 4, 704-713
13. Halim M.A., Cho H., Salauddin M., Park J.Y., 2016, A miniaturized electromagnetic vibration energy harvester using flux-guided magnet stacks for human-body-induced motion, *Sensors and Actuators A: Physical*, **249**, 23-29
14. Hu Y.T., Xue H., Hu H.P., 2013, Piezoelectric power/energy harvesters, [In:] *Analysis of Piezoelectric Structures and Devices*, Chapter 3, D. Fang, J. Wang, W. Chen (Edit.), De Gruyter: Berlin, Germany, 72-75
15. Jereczek B., Maciejewski I., Krzyżyński T., Królikowski T., 2022, Modeling and simulation of the horizontal seat suspension system under random vibration, *Procedia Computer Science*, **207**, 858-866
16. Jiang J., Liu S., Feng L., Zhao D., 2021, A review of piezoelectric vibration energy harvesting with magnetic coupling based on different structural characteristics, *Micromachines*, **12**, 4, 436
17. Krause P.C., Wasynczuk O., Sudhoff S.D., 2002, *Analysis of Electric Machinery and Drive Systems*, Wiley-IEEE Press
18. Kim T., 2011, Regenerative braking control of a light fuel cell hybrid electric vehicle, *Electric Power Components and Systems*, **39**, 5, 446-460

19. Li L., Ping X., Shi J., Wang X., Wu X., 2021, Energy recovery strategy for regenerative braking system of intelligent four-wheel independent drive electric vehicles, *IET Intelligent Transport Systems*, **15**, 1, 119-131

20. Liu C., Zhang K., 2021, Research on regenerative braking energy recovery strategy of electric vehicle, *Journal of Physics: Conference Series*, **2030**, 1, 012003

21. Maciejewski I., Blazejewski A., Pecolt S., Krzyzynski T., 2022a, A sliding mode control strategy for active horizontal seat suspension under realistic input vibration, *Journal of Vibration and Control*, **29**, 11-12, 25392551

22. Maciejewski I., Błażejewski A., Pecolt S., Królikowski T., 2022b, Multi-body model simulating biodynamic response of the seated human under whole-body vibration, *Procedia Computer Science*, **207**, 227-234

23. Maciejewski I., Błażejewski A., Pecolt S., Krzyżyński T., Zaporski P., 2023, Three--dimensional modelling and parameter identification of the seated human body exposed to random vibration, *Journal of Theoretical and Applied Mechanics*, **61**, 4, 833-845

24. Maciejewski I., Glowinski S., Krzyzynski T., 2014, Active control of a seat suspension with the system adaptation to varying load mass, *Mechatronics*, **24**, 8, 1242-1253

25. Maciejewski I., Zlobinski M., Krzyzynski T., Glowinski S., 2020, Vibration control of an active horizontal seat suspension with a permanent magnet synchronous motor, *Journal of Sound and Vibration*, **488**, 115655

26. Matak M., Solek P., 2013, Harvesting the vibration energy, *American Journal of Mechanical Engineering*, **7**, 438-442

27. Mescia L., Losito O., Prudenzano F., 2015, *Innovative Materials and Systems for Energy Harvesting Applications*, IGI Global: Hershey, PA, USA, 254-259, 271-272

28. Muscat A., Bhattacharya S., Zhu Y., 2022, Electromagnetic vibrational energy harvesters: A review, *Sensors*, **22**, 15, 5555

29. Naseri F., Farjah E., Ghanbari T., 2017, An efficient regenerative braking system based on battery/supercapacitor for electric, hybrid, and plug-in hybrid electric vehicles with BLDC motor, *IEEE Transactions on Vehicular Technology*, **66**, 5, 3724-3738

30. Onar O.C., Khaligh A., 2012, A novel integrated magnetic structure based DC/DC converter for hybrid battery/ultracapacitor energy storage systems, *IEEE Transactions on Smart Grid*, **3**, 1, 296-307

31. Priya S., Song H., Zhou Y., Varghese R., Chopra A., *et al.*, 2017, A review on piezoelectric energy harvesting, materials, methods, and circuits, *Energy Harvesting and Systems*, **4**, 3-39

32. Salman W., Qi L., Zhu X., Pan H., Zhang X., *et al.*, 2018, A high-efficiency energy regenerative shock absorber using helical gears for powering low-wattage electrical device of electric vehicles, *Energy*, **159**, 361-372

33. Song Z., Li J., Han X., Xu L., Lu L., *et al.*, 2014, Multi-objective optimization of a semi-active battery/supercapacitor energy storage system for electric vehicles, *Applied Energy*, **135**, 212-224

34. Sudevalayam S., Kulkarni P., 2011, Energy harvesting sensor nodes: Survey and implications, *IEEE Communications Surveys and Tutorials*, **13**, 443-461

35. Taut A., Pop O., Ceuca E., 2013, System for energy recovering with BLDC motor at deceleration momentum, *Proceedings of the 36th International Spring Seminar on Electronics Technology*, 299-304

36. Toshiyoshi H., Ju S., Honma H., Ji C.-H., Fujita H., 2019, MEMS vibrational energy harvesters, *Science and Technology of Advanced Materials*, **20**, 124-143

37. Ueno T., 2019, Magnetostrictive vibrational power generator for battery-free IoT application, *AIP Advances*, **9**, 035018

38. Wang X., 2020, Research on track vibration energy recovery system for vehicle operation based on network system, *Journal of Physics: Conference Series*, **1574**, 1, 012055

39. Yang Y.P., Liu J.J., Wang T.J., Kuo K.C., Hsu P.E., 2007, An electric gearshift with ultracapacitors for the power train of an electric vehicle with a directly driven wheel motor, *IEEE Transactions on Vehicular Technology*, **56**, 5, 2421-2431

40. Zhang R., Wang C., Zou P., Lin R., Ma L., *et al.*, 2022, Compositionally complex doping for zero-strain zero-cobalt layered cathodes, *Nature*, **610**, 67-73

41. Zhu Y., Moheimani S.O.R., Yuce M.R., 2010a, A 2-DOF MEMS ultrasonic energy harvester, *IEEE Sensors Journal*, **11**, 155-161

42. Zhu Y., Moheimani S., Yuce M., 2010b, Ultrasonic energy transmission and conversion using a 2-D MEMS resonator, *IEEE Electron Device Letters*, **31**, 374-376

Manuscript received November 15, 2023; accepted for print January 10, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 337-349, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185887

SIMULATION STUDIES AND EXPERIMENTAL RESEARCH OF OMNIDIRECTIONAL TRACKED VEHICLE[1]

Mateusz Fiedeń, Jacek Bałchanowski
Wrocław University of Science and Technology, Wrocław, Poland
e-mail: mateusz.fieden@pwr.edu.pl; jacek.balchanowski@pwr.edu.pl

The article focuses on the control of an omni-tracked vehicle in a symmetrical fully overlapping track system. The vehicle in question is equipped with four independently controlled tracks. The links of each crawler are equipped with a single rolling roller, fixed at an angle to the direction of the vehicle main axis. The authors propose mathematical description to determine the direction and speed of movement of a single pair of omni-tracks with oppositely arranged rolling rollers. Numerical tests were carried out, the results of which were compared with the mathematical model. The numerical studies were then subjected to experimental verification, using a full-scale prototype. A dynamic direction correction algorithm was also proposed with its effectiveness proved experimentally.

Keywords: kinematic, omnidirectional, omnitank, omnitracks, omnivehicle

1. Introduction

Omnidirectional vehicles and drives that enable this type of motion have been present in mechanics for more than 100 years (Grabowiecki, 1919). Wheeled robots equipped with Swedish wheels, or mecanum wheels, have been known for many years, so knowledge about them is extensive and mathematical descriptions are accurate (Bae and Kang, 2016; Taheri and Zhao, 2020; Yamada *et al.*, 2017). However, there is a certain group of omnidirectional vehicles that has appeared in scientific studies only recently – omnidirectional tracked vehicles. These vehicles are equipped with a variation of typical link-composed tracks, in which each link is equipped with at least one free rolling roller located at an appropriate angle with respect to the main axis of the track. An example of such a roller is shown in Fig. 1. Vehicles equipped with such tracks

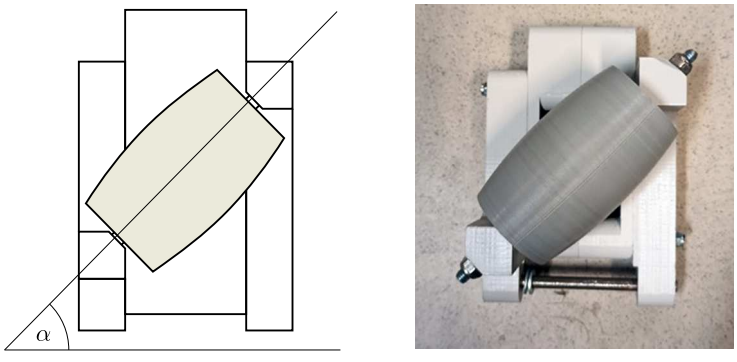


Fig. 1. Omnidirectional track link with a single rolling roller mounted at an angle α

can be divided according to the mutual orientation of the rollers and the mutual orientation of the entire track segments (Zhang *et al.*, 2018). There are also separate groups of omnidirectional

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

vehicles equipped with transverse active drive rollers (Takaduma *et al.*, 2008, 2018; Takaneetal, 2019).

Vehicles with non-parallel tracks have been present in the literature for more than 20 years. The article (Isoda *et al.*, 1999) presents the concept of a vehicle with four tracks. The tracks are arranged in a square plan. Each track has an independent drive. The rollers are located at $\alpha = 90°$ to the longitudinal axis of the track. Further testing of the vehicle is presented in the paper (Chen *et al.*, 2002), where dynamic analysis of the model, the proposed control system and the results of off-road runs of the prototype were presented. A similar concept was presented by Bruton (2023). Vehicle proposed by author is equipped with tracks with rollers fixed at $\alpha = 90°$. The track segments are arranged in the plan of an equilateral triangle. Vehicles with non-parallel tracks have one common feature: during translational movement, regardless of its direction, free rolling rollers are used to roll the vehicle body. For this reason, the potential of vehicles in question to overcome off-road obstacles, as well as to navigate difficult terrain, can be significantly reduced, although these vehicles are theoretically still tracked vehicles. A separate group is tracked vehicles with parallel tracks. Unlike the previously described group, there is no rotational movement of the free rolling rollers during movement in the main axis, which allows the features of a classic tracked vehicle to be maintained. The division of omnidirectional vehicles with parallel tracks is presented in Fig. 2.

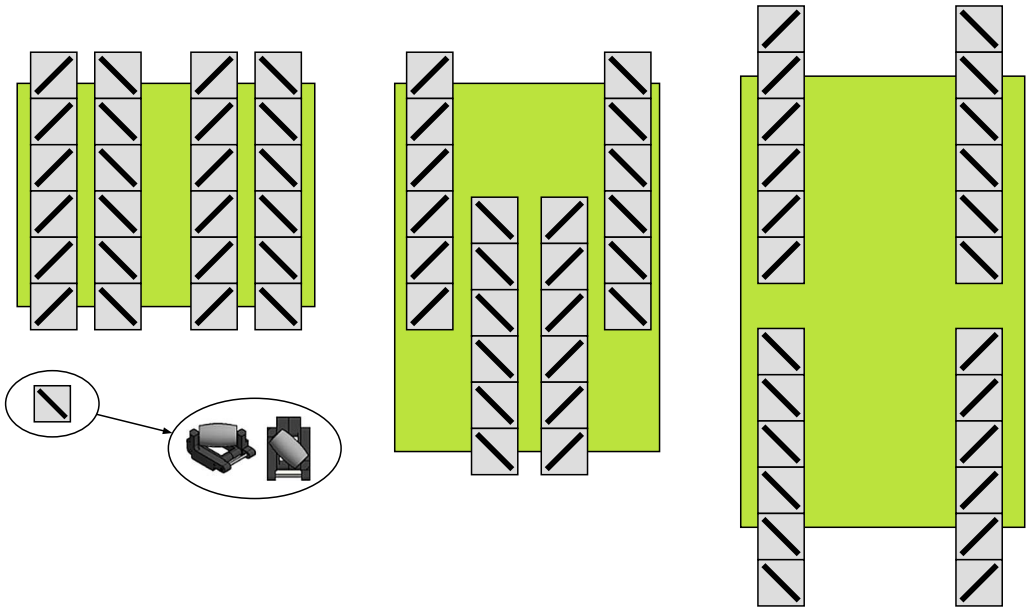


Fig. 2. Diagram of the arrangement of free rolling rollers in different types of omni-track vehicles. From left: fully overlapping, partially overlapping and non-overlapping tracks. In all examples, the distribution of rollers is longitudinally symmetrical

In the literature, the most widely described group of vehicles with parallel tracks are vehicles with fully non-overlapping tracks. The article (Zhang and Huang, 2015) presents theoretical considerations on the kinematics of vehicles with tracks with free rolling rollers in both parallel and non-parallel systems. It presents field tests of a prototype of the vehicle in question, including the current consumption of various drives during different types of movement and the relationship between the angular orientation of the robot and its ability to overcome terrain obstacles. In (Mortensen Ernits *et al.*, 2017), the authors present tests conducted on two built omni-track vehicles. The first of these had dimensions of $0.8\,\mathrm{m}\times1.2\,\mathrm{m}$, while the second was $2.5\,\mathrm{m}\times3.5\,\mathrm{m}$. The article describes a series of experimental tests that were carried out on the proposed demonstrators. The tests consisted of traversing preset trajectories that took into

account a change in the angular orientation of the body and movement in different directions. Theoretical considerations on the potential application of such vehicles are presented.

The work presented in (Fang *et al.*, 2020) includes numerical testing of a full-scale prototype of an omni-tracked vehicle. The equations of kinematics for this type of chassis, as well as analysis of the effect of the angular orientation of the rolling roll on the obtained motion speeds were presented. In the numerical tests, linear motion in different directions and motion folding of the prototype vehicle were simulated and described. In addition, the article provided experimental studies measuring the effect of movement direction and drives speed on the values of the current drawn by the drives. Tracked vehicles in a partially occurring system are described in (Zhang *et al.*, 2018), where simulation studies of such a vehicle are presented. The paper highlights the aspect of curvature of the motion trajectory when driving in a direction other than the main axis of the vehicle. This effect was confirmed in (Fang *et al.*, 2020; Fiedeń and Bałchanowski, 2021).

Vehicles with chassis in a fully overlapping track system are described in (Fiedeń and Bałchanowski, 2021). The authors presented the construction and testing of a lightweight prototype of the omni-tracked vehicle. The research included analysis of the trajectory of movement in the main axis and transverse axis, recorded on a bench equipped with a vision system. The paper proposes a method for counteracting the curvature of the trajectory of motion – the concept of static correction, which counteracts the unwanted rotation of the vehicle body by appropriately modifying the speed of individual drives. The results of tests confirm an improvement of the driving performance after applying the proposed correction algorithm.

The literature review revealed a significant research gap in the state of knowledge regarding omni-track vehicles. Many aspects of motion of such vehicles, for example, the compatibility of real models with theoretical models, the influence of design parameters on driving properties or adverse phenomena occurring during movement have been neglected or discussed and studied very narrowly (Fiedeń and Bałchanowski, 2022). The purpose of this paper is to discuss the unfavorable phenomenon of curvature of the trajectory of motion of an omni-track vehicle, as well as to present a proposed method to counteract this phenomenon.

2. Materials and methods

2.1. Design of kinematic model of an omnidirectional tracked vehicle and simulation research

Equations of kinematics have been proposed to determine the direction and speed of linear motion of a single omnidirectional track work. The starting point is the angular speed of drive wheels ω_r^1 and ω_r^2 with radius R. The individual vectors and the relationship between them are illustrated in Fig. 3.

The linear velocity of the track n (n = 1-4) denoted by v_{tr}^n, is calculated from the formula

$$v_{tr}^n = \omega_r^n R \tag{2.1}$$

The linear speed of the tracks can be decomposed into two components v_x^n and v_y^n. The component v_y^n is the same as the linear speed of the tracks, its direction is parallel to the track, and its value is

$$v_y^n = v_{tr}^n \tag{2.2}$$

The component v_x^n depends on the angular orientation of the free rolling roll, denoted by β, its direction is perpendicular to the track. For angles $\beta^1, \beta^3 = 135^\circ$ and $\beta^2, \beta^4 = 45^\circ$ it can be calculated from

$$v_x^n = \cot(\beta^n) v_{tr}^n \tag{2.3}$$

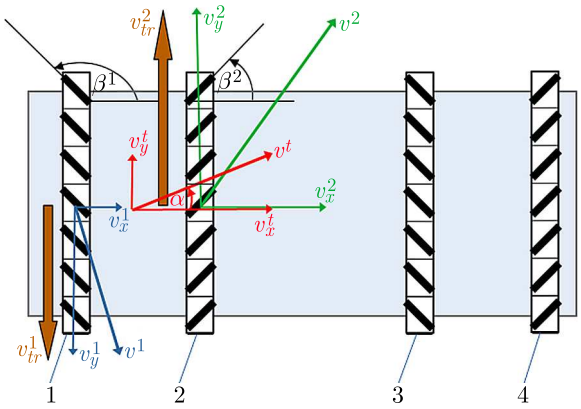


Fig. 3. Distribution of velocity vectors with designations

The velocity vector of a pair of tracks is denoted by v^t, and its angular orientation is denoted by α. The value of the velocity vector of the pair of tracks is

$$v^t = \sqrt{(v^t_x)^2 + (v^t_y)^2} \tag{2.4}$$

The values of the components, labeled v^t_x and v^t_y, depend on the values of the vectors v^n_x and v^n_y

$$v^t_x = \frac{v^1_x + v^2_x}{2} \qquad v^t_y = \frac{v^1_y + v^2_y}{2} \tag{2.5}$$

The angular orientation α is expressed by the formula

$$\alpha = \arctan 2(v^t_y, v^t_x) \tag{2.6}$$

A solid model of a vehicle equipped with four linear drives was proposed. Each drive moved a single beam with attached free rolling rollers. Two rolling rollers were attached to each beam. The rollers were fixed at an angle β^1, $\beta^3 = 135°$ and β^2, $\beta^4 = 45°$ to the main axis of the model, and the way they were fixed along with the whole model, is shown in Fig. 4. The distance between the rollers was 200 mm. The total weight of the model was 20 kg. The distance between adjacent beams was 0.1 m.

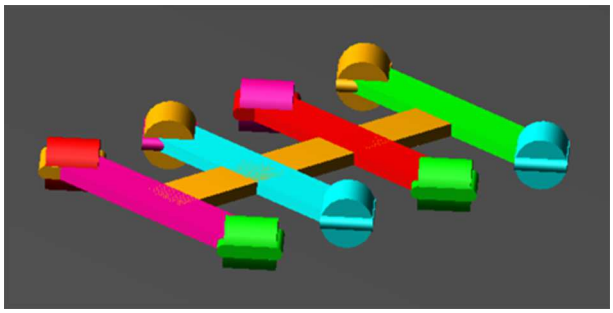


Fig. 4. A view of the simplified solid model of the omni-track vehicle

Dynamic contacts were formed between the rollers and the substrate. The contact parameters used for the simulation, selected based on (Engström *et al.*, 2010; Pasini, 2019), are shown in Table 1.

Based on the proposed equations, velocities were determined enabling the vehicle to move at selected angles. The displacements in time are given by a polynomial function with a smooth increment. The characteristics are shown in Fig. 5.

Table 1. Contact parameters between the substrate and the roller

Stiffness	20 N/mm	Static coefficient	1
Force exponent	2.2	Dynamic coefficient	1
Damping	0.001 Ns/mm	Stiction transition velocity	10 mm/s
Penetration depth	1 mm	Friction transition velocity	2000 mm/s

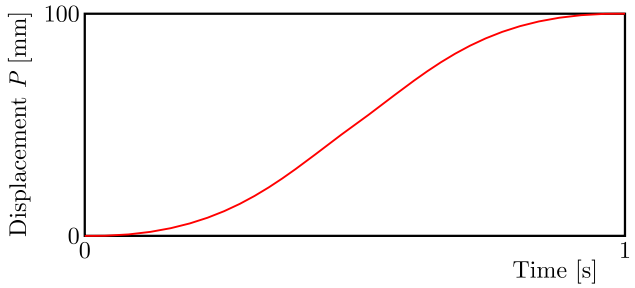


Fig. 5. Plot of the polynomial (Hexagon Adams STEP5 function) used as a control function in the drives of the proposed model

Table 2 shows the values of the speeds set for each drive, the expected orientation angles of the motion vector of the whole vehicle for these speeds, the expected speeds, and the orientation angle of the motion vector and speeds obtained by simulation. The t_s time of the simulation carried out was 1 s.

Table 2. Summary of set speeds of individual tracks, expected trajectory angles and obtained trajectory angles

ID	$v^1 = v^3$ [mm/s]	$v^2 = v^4$ [mm/s]	Expected trajectory angle [°]	Obtained trajectory angle [°]
a	−100	100	0.00	−0.19
b	−100	50	18.44	18.35
c	−100	0	45.00	45.01
d	−100	−50	71.56	71.56

Point P is located on the body of the robot. Figure 6 shows the obtained trajectory of the movement of point P during individual passes.

The simulations showed that the proposed model correctly reproduces the relationship between the rotational speeds of the drives and the direction of motion of the vehicle body. Therefore, it was reasonable to build a prototype with comparable design parameters for experimental verification of the obtained results.

2.2. Design of a prototype of an omnitracked vehicle and a research stand

To verify the correctness of numerical simulations, as well as the effectiveness of the proposed correction algorithm, a prototype omni-track vehicle was designed and manufactured. Then a measurement station equipped with a vision system was prepared.

The vehicle prototype consisted of four segments. Each segment was equipped with a crawler, one independent drive, drive wheel, tension wheel and road wheels. The kinematic diagram is shown in Fig. 7.

The projection of a single segment is shown in Fig. 8. Each segment consists of a single drive, which transmits torque to the drive wheel via a chain gear. A single track, consisting of 19 segments, is stretched between the drive and tension wheel. When moving on a flat surface at any time, a minimum of 6 rollers touch the ground. The weight of the vehicle is transferred

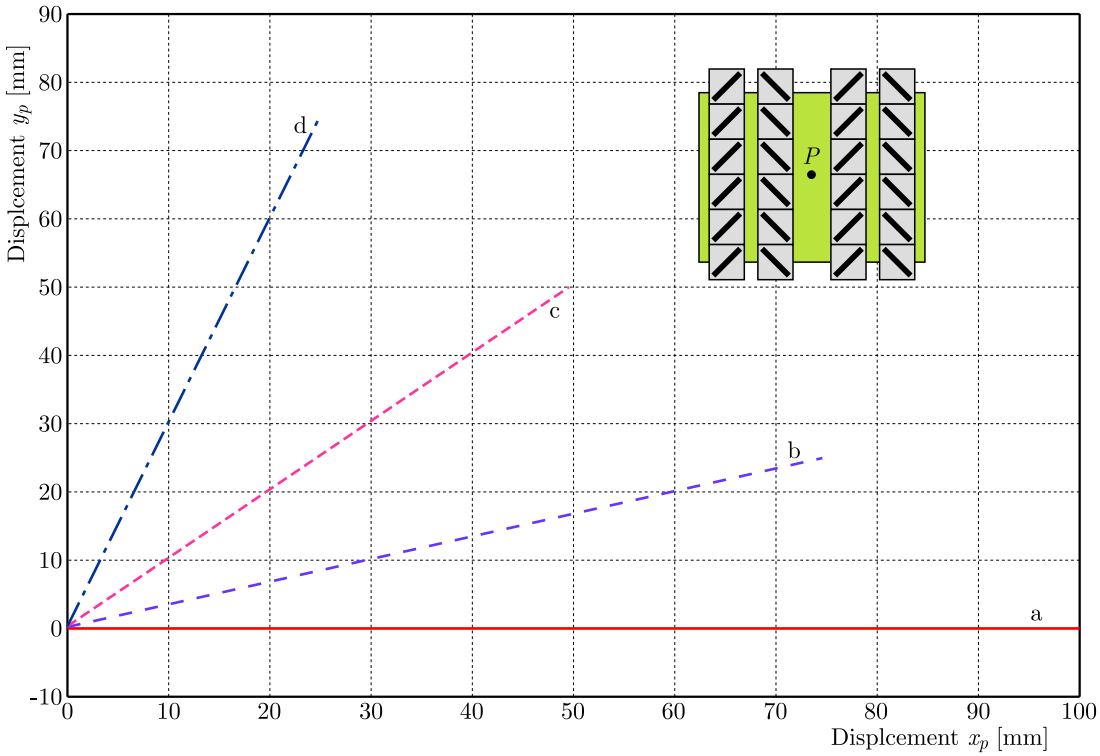


Fig. 6. Point P trajectories during individual runs $x_p y_p$

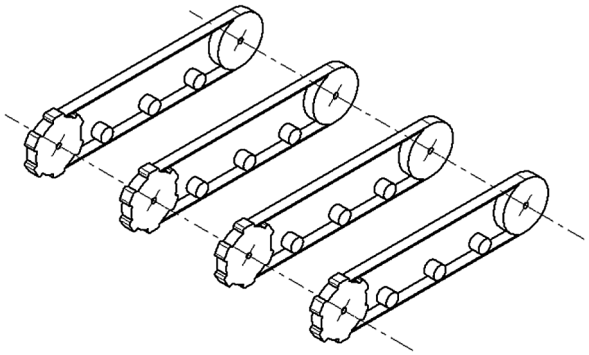


Fig. 7. Kinematic diagram of the prototype of an omni-track vehicle

to the ground via road wheels. This role is performed by road wheels and tension wheels, as well as three free wheels equipped with a spring-based compression system. A visualization of the prototype and a photo of the actual design are shown in Fig. 9. The basic design parameters of the robot are collected in Table 3.

Table 3. Basic design parameters of the robot

Roller diameter	50 mm	Drive wheel diameter	185 mm
Distance between drive and tension wheel	760 mm	Number of links in single track segment	19
Frame length	1050 mm	Frame width	960 mm
Weight	70 kg	Rated power of single drive	250 W

An AS5040 encoder with a resolution of 512 pulses per revolution was mounted on each drive, which, with the 9:16 gear used, allowed a theoretical resolution of less than 1 mm for measuring

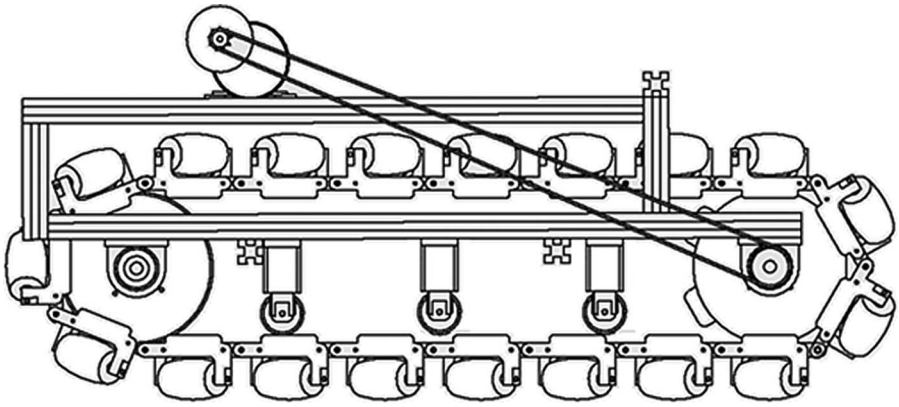


Fig. 8. Side view of a single track segment

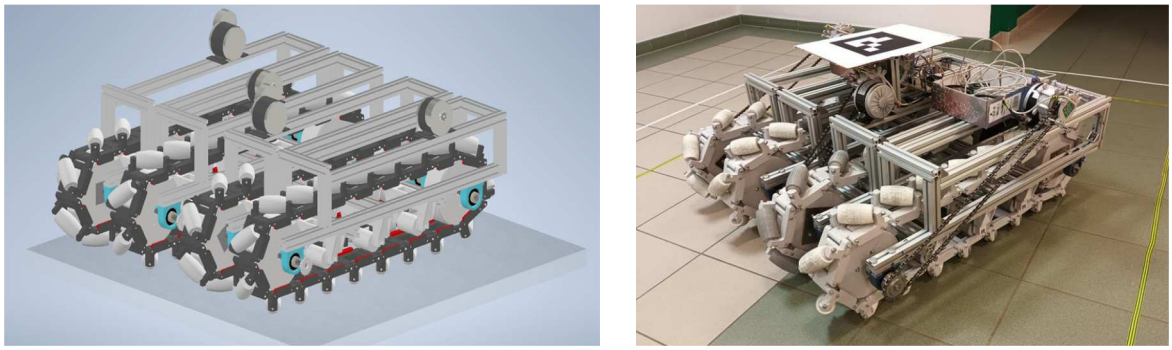


Fig. 9. Visualization of the prototype and a photo of the actual vehicle

linear motion of the track. An ATMega2560 microcontroller was used as the computational unit. Additional data on the vehicle angular orientation was provided by an NGIMU sensor. The control scheme of the vehicles is shown in Fig. 10. The robot can be controlled both by an operator, using RC apparatus, as well as by commands sent via UART.

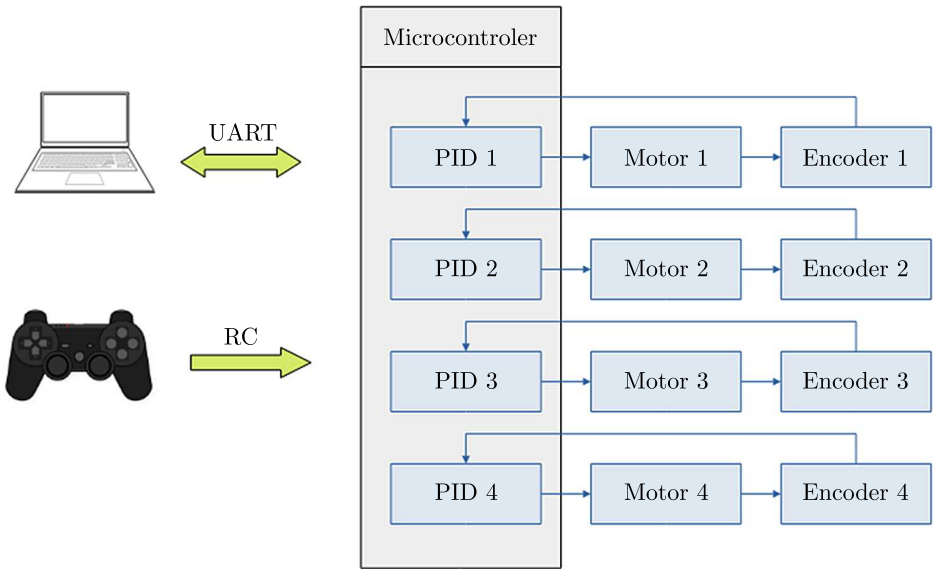


Fig. 10. General control scheme of the UART prototype

By equipping the prototype with an IMU sensor, it was possible to add an active direction compensation system. A schematic of the proposed system operation is shown in Fig. 11. The

variables i take values from 1 to 4, symbolizing successive drives. Variable φ_I is information about vehicle angular orientation at the time of takeoff, i.e. the yaw angle. This variable is transmitted once, before the start of movement. Variable φ_S is the stored initial angular orientation, which is to remain constant throughout the movement. Variable $\omega_i S$ are the set angular velocity values of individual drives. Variable $E\varphi$ is a difference between the actual and expected values of the vehicle angular orientation of the individual actuators. Analogically, $E\omega_i$ is a difference of angular velocity values. Variables $U\varphi$ and $U\omega_i$ are control signals, P_i describes values of the PWM signal that goes to the individual drives. Variables $\varphi(t)$ and $\omega(t)$ are the actual values of the actual angular orientation and angular velocity of the drives as measured by the IMU sensor and encoders. The $\omega_i S$ variable is the velocity reference value for the individual drives.

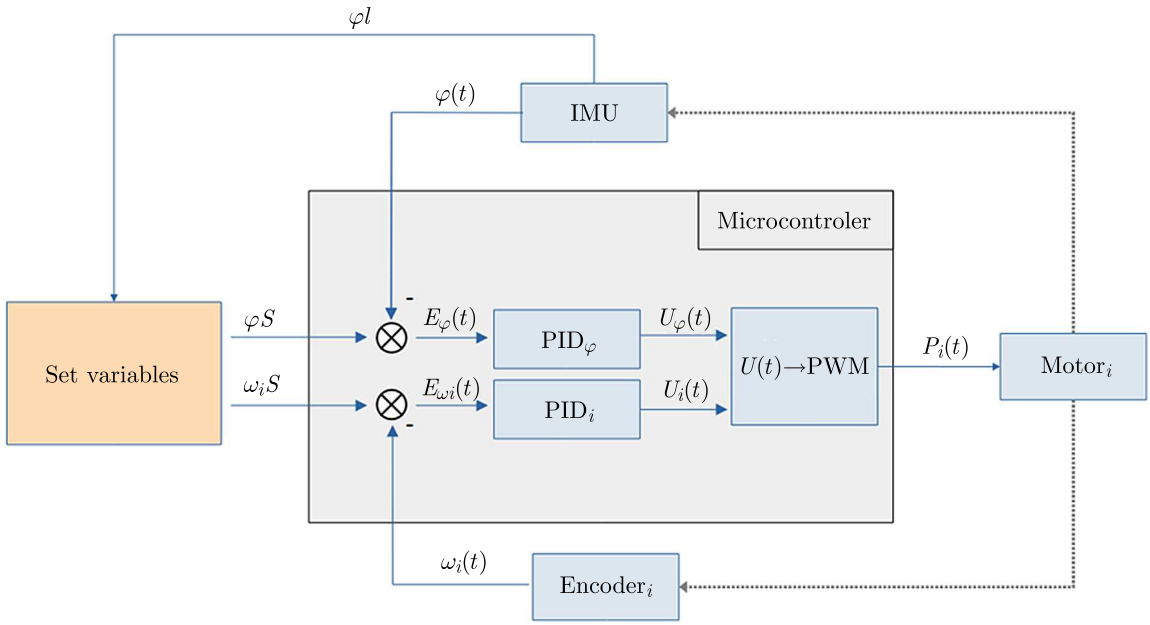


Fig. 11. Data flow diagram of the control system when using dynamic correction

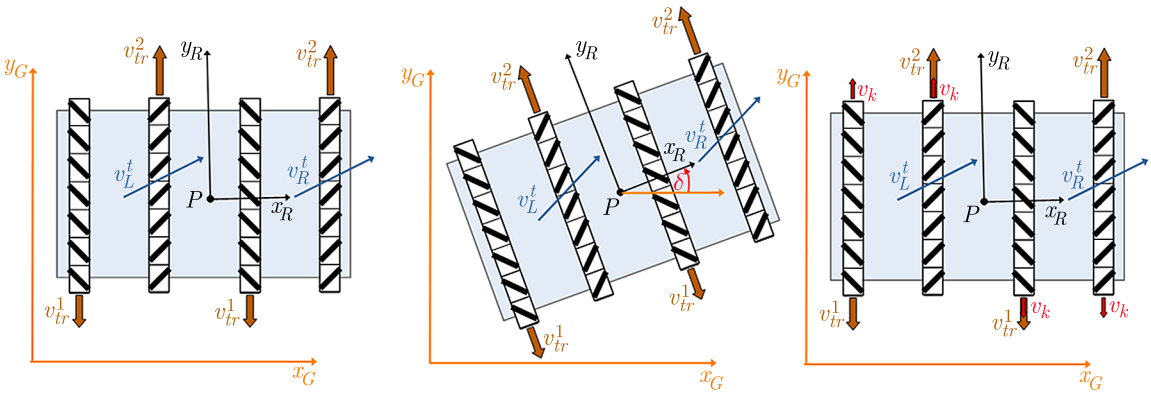


Fig. 12. Concept of dynamic correction of the movement direction. From the left: at the start of the movement, after losing the correct angular orientation of the body and after activating the correction

When moving with a fixed body orientation, the yaw angle should not change. Thanks to the data provided by the IMU sensor, information about the actual angle orientation, which determines the orientation of the vehicle on the plane, is fed into the system. Then a speed correction is introduced into the drive control system to counteract this unfavorable phenomenon. A diagram of how the correction works is presented in Fig. 12. The geometric center of omni-track vehicles is point P. Associated with this point is a coordinate system whose y-axis coincides

with the main axis of the vehicle. When there is a change in the angular orientation of the vehicle body (due to slippage, external forces or design defects), the value of the angle δ will change, which at the beginning is 0. The angle δ is between the coordinate system at the point P and the global coordinate system, denoted by X_G, Y_G. The value of the angle δ is converted into velocity values v_k, which are added to the set linear velocities of the tracks and force orientation correction.

In order to record the trajectory of actual motion of the prototype, a measurement station was prepared, consisting of a 3000 mm×2400 mm measurement field, an ELP-USB4KHDR01--MFV camera mounted at a height of $h = 4070$ mm above the measurement field, and a computer for image acquisition. The stand with the prototype is shown in Fig. 13.

Fig. 13. Measurement stand of the omnidirectional vehicle

The image distortion effect seen in the camera image was removed by performing camera calibration using a charUCO card. The calibration card was a charUCO board shown in Fig. 14. It is a typical calibration checkerboard enhanced with arUCO markers. A sample arUCO marker is shown in Fig. 15. Its operation is made possible by the arUCO module, which is a part of the OpenCV open source image analysis library (https://opencv.org/). This module makes it convenient to work with arUCO markers, making it possible to read their position and angular orientation relative to the camera.

After the camera calibration process, a test of the measurement station was performed. Six arUCO markers with known sizes, positions and relative angular orientation were applied to an A0 sheet. Then a series of images were recorded with the calibration sheet placed in different parts of the measurement field. Examplary images are shown in Fig. 16. Tests showed that the prepared measurement station allows recording the robot angular orientation φ_S to an accuracy of less than 1 degree and the position x_p, y_p to an accuracy of less than 0.04 m.

A single arUCO marker was placed on the body of the robot. This made it possible to read the actual position and angular orientation of the platform under test.

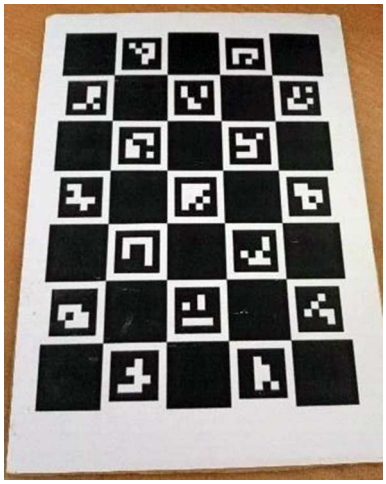


Fig. 14. CharUCO calibration card

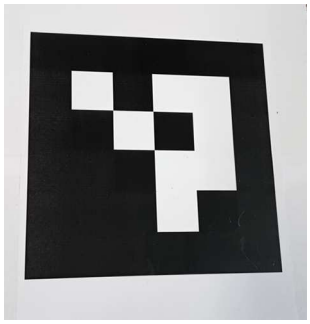


Fig. 15. arUCO marker

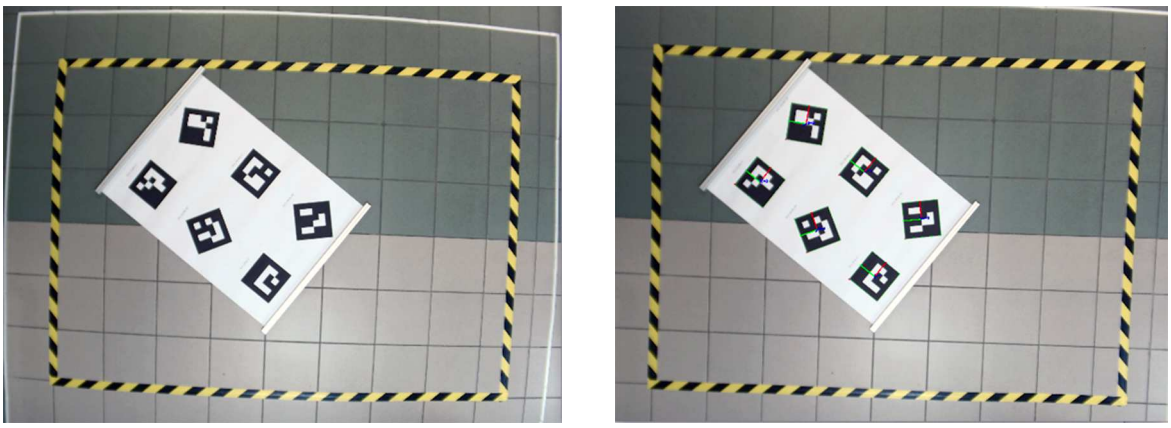


Fig. 16. Measurement field with calibration sheet before and after removal of distortion

3. Results

The prototype of the omni-tracked vehicle was used to conduct bench tests. These tests included conducting a series of test runs. During execution of the runs, the rotation of drive wheels and the angular orientation of the body were recorded by encoders and IMU sensor, and the actual position of the mobile platform was measured using the reference provided by vision system. The test runs were about 2 m in length. The measurements were compared with the results of numerical simulation. The effect of dynamic direction correction on the movement was also verified. In the type stage of the study, research of 5 ride types was conducted. Each tested drive type involved execution of 10 runs. The speed values for each drive were calculated based

on the proposed kinematic equations. The goal was to achieve movements at angles of 0, 18.44, 45, 71.56 and 90 degrees to the main axis of the platform. The theoretically calculated speed of platform movement, as well as the orientation angle of the velocity vector were compared with empirically measured values.

Table 4 collects the results of runs without correction as well as with the active drive correction system. The difference in angular deviation for runs with and without correction is shown. Standard deviations and variance were calculated for the results.

Table 4. Average angular orientation errors with respect to distance traveled for measurements that were made with and without the dynamic direction correction system. For each expected angle 10 passes were made. The highest and lowest scores were discarded

Trajectory angle φ [°]	Average orientation error δ [°/m]		Final orientation error variance σ^2		Standard deviation of final orientation error σ	
	no correction	with correction	no correction	with correction	no correction	with correction
18.44	2.60	0.81	0.05	0.22	0.22	0.47
45	1.30	0.65	0.29	0.19	0.54	0.43
71.56	3.02	0.04	0.67	0.25	0.82	0.50
90	3.79	−0.50	0.55	0.08	0.74	0.28

Statistical data shows repeatability of the measurements. Significant improvements in the driving performance can be seen when driving with the active correction system.

4. Summary, conclusions

The article deals with the control of an omni-tracked vehicle with parallel, fully overlapping tracks. A review of the literature shows the existing research gap in the issues of compensation for the unfavorable phenomenon: trajectory curvature during motion of a vehicle of this type. The equations of kinematics for controlling a single pair of parallel, fully overlapping tracks are presented. A solid model was prepared to simulate control of the presented vehicle. The simulation showed that the model with the assumed physical parameters behaved according to the proposed kinematic equations. Based on the simulation assumptions, a full-scale prototype of the omni-track vehicle was made. The kinematic scheme, design parameters and control system scheme were presented. An algorithm was proposed to dynamically compensate for the phenomenon trajectory curvature during motion. A number of test runs was carried out on a test stand equipped with a vision system. The obtained motion directions were compared with the expected values calculated on the basis of the equations of kinematics, and with the results of numerical simulation. Next, a series of test runs was carried out with the active direction correction system. These were compared with runs made without the active correction system.

The tests have shown that the proposed kinematics equations can be used to control the proposed solid model as well as actual prototype. When driving without the active correction system, there is a gradual drift of the vehicle angular orientation. As a result, the trajectory of motion becomes curved. This phenomenon is most easily noted when driving at an angle of 90, taking 3.8°/m. When using the dynamic direction correction algorithm, the phenomenon in question is significantly reduced. A several-fold improvement in the driving performance is apparent for all types of traffic studied. The smallest improvement (double) was noticed for the movement at an angle of 45 degrees. This is probably due to the way the drive is transmitted – the chain transmission has slack, which hinders the operation of the correction system. In the

remaining cases, at least a three-fold reduction in the angular orientation error was noticed. The future work will focus on detecting an active motion correction system during end-of-line motion. Tests are planned on various types of surfaces (grass, concrete, sand), as well as combined journeys, when the robot moves on several types of surfaces.

References

1. Bae J., Kang N., 2016, Design optimization of a mecanum wheel to reduce vertical vibrations by the consideration of equivalent stiffness, *Shock and Vibration*, **2016**
2. Bruton J., 2023, Triangle Tank Version 2, https://www.youtube.com/watch?v=rmvrlFp-qEU (online access 17.10.2023)
3. Chen P., Mitsutake S., Isoda T., Shi T., 2002, Omni-directional robot and adaptive control method for off-road running, *IEEE Transactions On Robotics And Automation*, **18**, 2
4. Engström J., Richloow E., Wickström A., 2010, Modeling of robotic hand for dynamic simulation, Bachelor Thesis, School of Industrial Engineering and Management (ITM), MMKB 2010:23
5. Fang Y., Zhang Y., Li N., Shang Y., 2020, Research on a medium-tracked omni-vehicle, *Mechanical Sciences*, **11**, 137-152
6. Fiedeń M., Bałchanowski J., 2021, A mobile robot with omnidirectional tracks – design and experimental research, *Applied Science*, **11**, 11778
7. Fiedeń M., Bałchanowski J., 2022, Testing the driving parameters of an omni-tracked robot (in Polish), *Prace Naukowe Politechniki Warszawskiej, Elektronika*, 255-263
8. Grabowiecki J., 1919, *Vehicle Wheel*, US patent 1305535
9. https://opencv.org/ (online access 17.10.2023)
10. Isoda T., Chen P., Mitsutake S., Toyota T., 1999, Roller-crawler type of omni-directional mobile robot for off-road running, *Transactions of the Japan Society of Mechanical Engineers, ed. C*, **65**, 636
11. Mortensen Ernits R., Hoppe N., Kuznetsov I., Uriarte C., Freitag M., 2017, A new omnidirectional track drive system for off-road vehicles, Proceedings of the XXII International Conference on "Material Handling, Constructions and Logistics"
12. Pasini D., 2019, Modelling of mobile vehicles for simulation and control, Master Degree Thesis, Politecnico di Torino
13. Tadakuma K., Tadakuma R., Nagatani K., Yoshida K., Peters S., Udengaard M., Iagnemma K., 2008, Crawler vehicle with circular cross-section unit to realize sideways motion, *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems IROS*, 2422-2428
14. Tadakuma K., Takane E., Fujita M., Nomura A., Komatsu H., Konyo M., Tadokoro S., 2018, Planar omnidirectional crawler mobile mechanism – development of actual mechanical prototype and basic experiments, *IEEE Robotics and Automation Letters*, **3**, 1
15. Taheri H., Zhao C., 2020, Omnidirectional mobile robots, mechanisms and navigation approaches, *Mechanism and Machine Theory*, **153**
16. Takane E., Tadakuma K., Shimzu T., Hayashi S., Watanabe M., Kagami S., Nagatani K., Konyo M., Tadakoro S., 2019, Basic performance of planar omnidirectional crawler during direction switching using disturbance degree of ground evaluation method, *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems IROS*, 2732-2739
17. Yamada N., Komura H., Endo G., Nabae H., Suzumor, K., 2017, Spiral mecanum wheel achieving omnidirectional locomotion in step-climbing, *Proceedings of the 2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 1285– 1290

18. Zhang Y., Huang T., 2015, Research on a tracked omnidirectional and cross-country vehicle, *Mechanism and Machine Theory*, **87**, 18-44

19. Zhang Y., Yang H., Fang Y., 2018, Design and motion analysis of a novel track platform, *Journal of Physics: Conference Series*, **1074**

Manuscript received October 30, 2023; accepted for print January 8, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 351-364, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185459

HIGHER ORDER NUMERICAL HOMOGENIZATION IN MODELING OF ASPHALT CONCRETE[1]

Marek Klimczak, Marta Oleksy
Cracow University of Technology, Cracow, Poland
corresponding author Marek Klimczak, e-mail: marek.klimczak@pk.edu.pl

In this paper, we present an enhanced version of the two-scale numerical homogenization with application to asphalt concrete modeling in the elastic range. We modified the method of effective material parameters tensor assessment for analysis based on the representative volume element (RVE). As the method was tested on asphalt concrete, we also present two possible approaches to geometrical modeling of its internal microstructure. Selected numerical tests were performed to verify the proposed approach. The main novelties of this study, i.e. higher order approximation at the macroscale and modification of boundary conditions at the level of RVE analysis, improved the efficiency of our methodology by error reduction. Practically, we obtained a reduction of NDOF up to 3 orders of magnitude (comparing to full-scale and homogenized analysis) that was accompanied with the introduced error of order of several percent (measured in L_2 norm).

Keywords: asphalt concrete, numerical homogenization, representative volume element

1. Introduction

Roads are systematically classified as "linear infrastructure objects". Practically, they exhibit a fully three-dimensional structure as multi-layered domains. In Poland, the roads with flexible pavement structures still remain the most popular among other types of pavements, i.e., rigid or semi-rigid ones. A flexible pavement structure consists of several asphalt layers and subbase layer(s) made of crushed rock resting on an improved subgrade. Due to specific groundwater or ground conditions, some additional layers (e.g., those made of geosynthetics) may be also applied.

The upper layers of the asphalt pavement structure play different roles in providing the demanded bearing capacity, durability and other parameters of the whole structure. Consequently, their internal structures are also different. It is mainly due to variety of asphalt mixture types that can be selected: asphalt concrete (AC), stone-mastic-asphalt (SMA), hot-rolled asphalt (HRA), reclaimed asphalt pavement (RAP), to mention only a few. The diversity within a specific mixture type can be obtained using extra additives or modifying the gradation curve of the aggregate mixture.

In this paper, we focus on the selected aspects of asphalt concrete modeling. In particular, the asphalt concrete microstructure and its impact on the effective macroscale parameters is studied. In Fig. 1, one can observe the limiting gradation curves of this asphalt mixture used for different pavement layers (GDDKiA, 2014). It can be noticed that the gradation is coarser for the bottom layers than for the upper ones. It is not only exclusively due to the fact that thickness of the subbase is greater than that of the wearing and binder course. Finer aggregate is used for the wearing course to resist visco-plastic deformations (ruts) occurring herein. Consequently, a coarser aggregate is used in the subbase in order to reduce the risk of structural deformations.

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

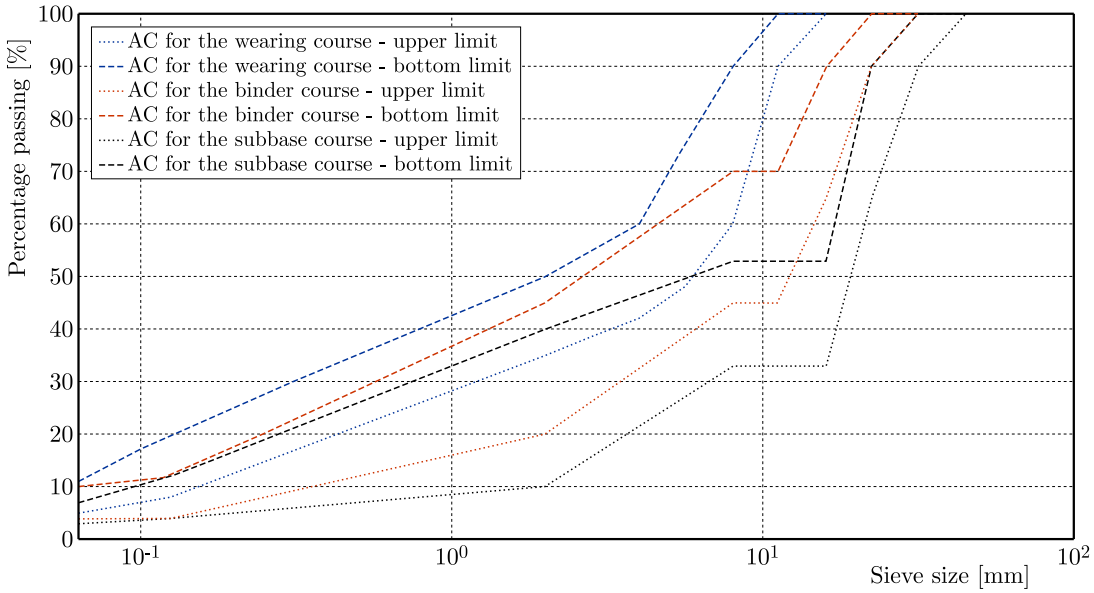


Fig. 1. AC gradation curves for different pavement layers plotted in a semilogarithmic scale (traffic category KR3-KR6)

As a material, asphalt concrete is a mixture of two main phases: aggregate particles and a mastic (which is, in turn, a mixture of the bitumen binder and mineral filler). The weight ratio of these two main phases is greater than 90/10. Volumetrically, several percent of air voids can be also distinguished. The description presented above is a rough approximation of the asphalt concrete recipe in its traditional form. There is a very active research field devoted to neat asphalt modifications, mastic modifications as well as the replacement of the natural aggregate mixture with different industrial wastes in the spirit of less-waste philosophy (Fakhari Tehrani *et al.*, 2013; Kim *et al.*, 2013; Schüller *et al.*, 2016; Ziaei-Rad *et al.*, 2012).

Aiming at the reliable numerical analysis of asphalt concrete, one needs to decide on the number of aspects. Let us list a three of them that are fundamental in our opinion:

- *Analysis scale* – from atomistic to the specimen/pavement structure scale (called in this paper as the macroscale). It is noticeable that the macroscale response is highly related to phenomena observed at the lower scales. In this paper, we present the framework for multiscale analysis that bridges the macroscale with the asphalt mixture scale (referred to as the microscale). We decided to keep this nomenclature for the sake of brevity. In the literature, the asphalt mixture scale is sometimes referred to as the mesoscale, whereas the microscale term is reserved for the scale of mastic with particles of dimensions of several μm. In this distinction, however, it would be difficult to describe the scale of asphalt mortar (with particles smaller than 2 mm). The sequence of consecutive analysis scales is shown in Fig. 2. Summing up, we model the specimen at the macroscale with its spatial dimensions kept but with the assumption on the homogeneity of the domain. Its effective parameters are numerically assessed on the basis of the microscale analysis. Precisely, we use the specific material parameters for the inclusions and for the matrix at the microscale and compute effective macroscale quantities.
- *Material model* – for every analysis scale addressed in this paper, one can easily identify the heterogeneity of its underlying scale. Regardless of the selected material model at the specific scale, the material response is evidently affected by lower scale phenomena. In the case of asphalt concrete, the constitutive equations describing the bitumen behavior are of particular interest. Linear (Aigner *et al.*, 2009; Collop *et al.*, 2003; Fakhari Tehrani *et al.*, 2013; Klimczak and Cecot, 2020a; Mo *et al.*, 2008; Woldekidan *et al.*, 2013; Ziaei-Rad *et al.*,

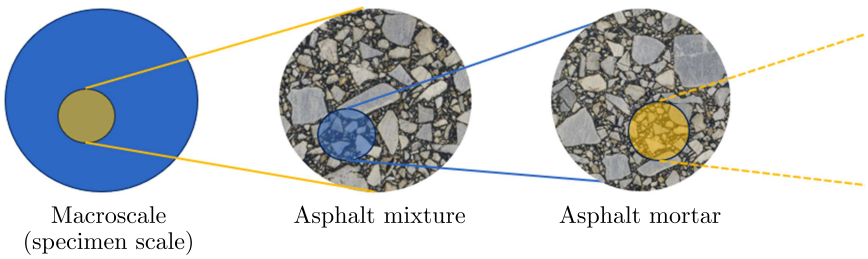


Fig. 2. Sequence of the selected analysis scales addressed in this study (yellow dotted line illustrates the occurrence of further underlying scales)

2012) and nonlinear viscoelasticity (Mitra *et al.*, 2012; Schüller *et al.*, 2016; Woldekidan *et al.*, 2013), viscoelastoplasticity (Aigner *et al.*, 2009; Collop *et al.*, 2003), viscoelasticity with damage (Kim *et al.*, 2013) and other theories are developed, to mention only a few. For the sake of brevity, we present in this study our results with the assumption on the elastic behavior of the binder phase. From the engineering point of view, this is a very strong assumption and can be understood as steady-state analysis at a specific temperature. The main scope of this paper is the application of the developed numerical method, however. Consequently, we apply both for the binder and aggregate phases the assumption on the linear elasticity in order to present the main findings of our research. Nonlinear analyses of materials can be found in our previous papers (Klimczak and Cecot, 2020b; Oleksy and Cecot, 2015).

- *Direct/multiscale analysis* – numerical analysis of the composite can be performed at a specific scale in a manifold manner. An assumption on the domain homogeneity can be made and the effective parameters are assessed phenomenologically/analytically or numerically. Consequently, the complexity of numerical analysis at this scale is very low, but the material response can be very smoothened. On the contrary, a very resource-consuming direct analysis can be performed with accounting for the internal structure of the composite (Fakhari Tehrani *et al.*, 2013; Mitra *et al.*, 2012; Mo *et al.*, 2008; Ziaei-Rad *et al.*, 2012). This type of analysis provides a deep insight into the lower scale phenomena, but it is time consuming and produces also an amount of data that can be hard to analyze and process. Somewhere in-between, there is a large group of multiscale methods. Generally, the macroscale analysis is performed at a low cost, but the microscale oscillations are accounted for performing some additional local analyses. The methods based on the concept of the representative volume element (RVE) are of particular significance nowadays.

In the field of asphalt concrete numerical modeling, one can distinguish a number of approaches to the above-mentioned aspects. A specific methodology is typically a trade-off between the complexity of the analyzed phenomena and the numerical cost of analysis. Below, we discuss several selected methodologies used in numerical modeling of asphalt concrete (or similar asphalt mixtures).

In (Collop *et al.*, 2003), the authors developed an elasto-visco-plastic model with damage for asphalt pavement layers. They modeled a multi-layered domains with the assumption on the homogeneity of every layer. The numerical effort of the analysis consisted in a time-stepping algorithm, whereas the geometry of the pavement structure was simple due to the assumed effective parameters for the whole layer. No multiscale analysis was necessary. In (Woldekidan *et al.*, 2013) and their other papers, the authors extended the scope of the analysis presented in (Collop *et al.*, 2003) and assessed the effective parameters of the asphalt concrete at different observation scales and for different material models. Such studies can be understood as a phenomenological assessment of the effective material parameters.

In (Mo *et al.*, 2008), the authors analyzed performance of porous asphalt concrete with a particular focus on the ravelling process (loss of wearing course aggregate particles). The internal structure of the domain was idealized significantly to facilitate the analysis. The authors used cylindrical (2D) or spherical (3D) objects to geometrically model the aggregate particles. They studied the interfacial zone between them and the bitumen binder using the viscoelastic material model. Since the analyzed specimens were idealized, it was possible to carry out a direct analysis. Nevertheless, it was an attempt to include in the numerical analysis the microscale phenomena. The numerical cost was reduced by geometry simplification.

Another approach to reliable modeling of asphalt concrete with a direct microstructure can be found e.g. in (Klimczak and Cecot, 2020a; Ziaei-Rad *et al.*, 2012). Therein, the authors analyze the specimens with the microstructures equivalent to the actual one. Namely, they generate synthetic microstructures possibly similar to the realistic ones. However, it is somehow arbitrary how to verify the similarity of these two types of microstructures beyond the visual inspection. The methods based on the Voronoi tessellation combined with the control of prescribed gradation curves are a typical approach. The shapes of inclusions are close to the realistic ones yet simplified. Consequently, direct transient analysis using e.g., viscoelasticity principles can be carried out at the microscale. Simplification of the aggregate particle shapes allows for a substantial reduction of the number of elements/nodes necessary for the numerical analysis. Typically, the asphalt mixture scale is used. If the contact phenomenon was also the analyzed problem, the aforementioned internal structural simplifications should refer to the respective scale. Oversimplified geometries would not cover the binder-aggregate interaction properly.

A very important and active research field is a multiscale analysis of composites. There is a wide spectrum of methods within this approach. For their comprehensive classification and review, we refer the reader e.g. to (Belytschko and de Borst, 2010; Fish, 2014; Kouznetsova *et al.*, 2002). Such a summary is beyond the scope of this paper. Instead, the applications of selected multiscale methods to the modeling of asphalt concrete and similar asphalt mixtures are briefly summarized below. The common feature of all these methods is the fact that they bridge the neighboring scales. The information derived from the lower analysis scale is transferred to the upper scale in order to facilitate computations at this level. Schematically, additional computations are necessary to incorporate the lower scale information at the upper scale. This cost, however, is justified in a manifold manner. Firstly, it is a way of making the analysis at some lower scales feasible. Secondly, the speed-up (very often due to possible parallel computing) is observed. A time-consuming direct analysis can be avoided. Thirdly, a number of neighboring scales can be analyzed simultaneously giving a deeper insight into the impact of the modifications introduced at the lower scale on the overall material response.

In (Aigner *et al.*, 2009), the authors used the concept of the localization tensor to study viscoelastic properties of asphalt concrete. Assuming spherical inclusions (aggregate particles) and using the Mori-Tanaka scheme (Mori and Tanaka, 1973) they obtained closed formulas for the effective material parameters at the macroscale. The idea of the representative volume element (RVE) was used to compute them locally. The RVE size should correspond with the size of inclusions to represent full information on the microstructure. Since asphalt concrete exhibits a random microstructure, this concept is more suitable than the unit-cell approach used for periodic domains.

In (Feyel and Chaboche, 2000; Guedes and Kikuchi, 1990; Kouznetsova *et al.*, 2002), the computational homogenization (typically associated with the finite element method) was developed and tested on various materials. Its standard version is also based on the RVE concept. In numerical analysis, two neighboring scales are specified. At the macroscale, one generates a coarse mesh and specifies a set of characteristic points (usually Gauss integration points) at this level. With each of such points, an RVE representing the local microstructure is defined. An iterative analysis is performed to transfer interchangeably the information from both scales.

First, the macroscale problem is solved. Deformation at Gauss points is used as a boundary condition for auxiliary boundary value problems solved within the RVE's corresponding with these points. Subsequently, the averaged quantities – see (Feyel and Chaboche, 2000; Guedes and Kikuchi, 1990; Kouznetsova *et al.*, 2002) for details – from this level are transferred to the macroscale and used for the next iterative solution. It should be underlined that no assumption on the material model at the macroscale is necessary in this approach, since the averaged strains and stresses are computed at the microscale level. In terms of the asphalt concrete modeling, the computational homogenization was used e.g., in (Kim *et al.*, 2013; Schüller *et al.*, 2016).

In (Wimmer *et al.*, 2016), a method of synthetic microstructure generation based on the Voronoi tessellation was presented to model the RVE. These local microstructures were later used for a randomly located set of RVE's. For each of them, periodic boundary conditions were used. Finally, the effective material parameter tensors were assessed using the statistics. The linear elastic model was used both for the matrix and the inclusions.

In (Schüller *et al.*, 2016), the effective macroscale stress using the generalized Maxwell-Zener viscoelastic model was computed on the basis of the RVE analysis. The authors also generated the Voronoi diagram-based microstructures replacing those obtained using the X-ray computed tomography (XRCT) as leading to the overkill mesh generation.

In (Kim *et al.*, 2013), the computational homogenization was used in order to bridge the effect of the cohesive zone occurrence at the microscale with damage observed at the macroscale. The macroscale stresses and strains were computed in terms of the microscale ones. Two numerical tests performed for idealized and real-like RVE microstructures illustrated the proposed framework.

Another approach to modeling of asphalt concrete was presented in (Klimczak and Cecot, 2020b). Therein, special shape functions accounting for the complex microstructure of the analyzed material were used for the solution approximation at the macroscale. This approach was successfully used for both the linear elastic and viscoelastic material models.

2. Methodology

2.1. Numerical homogenization

In this paper, we present the framework for higher order numerical homogenization of asphalt concrete. The numerical homogenization was developed e.g., in (Zohdi and Wriggers, 2005). Its idea is also based on the RVE concept. Namely, one performs a set of numerical tests for the identified RVE (Fig. 3).

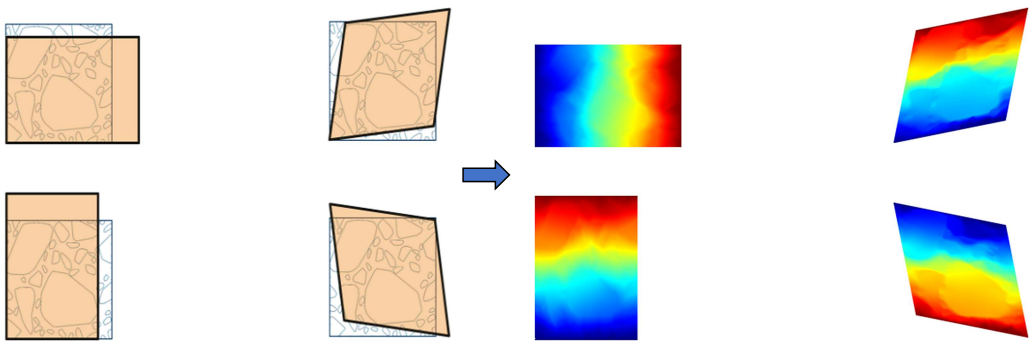


Fig. 3. A set of numerical tests (left) and their solutions (right)

Accounting for the microstructure observed at the RVE level, one solves a set of boundary value problems corresponding to real laboratory experiments (Fig. 3). Such an approach can be

understood as substitution of a similar procedure realized in a laboratory (phenomenological approach). In the numerical homogenization, average strain $\langle \varepsilon_{ij} \rangle$ and stress $\langle \sigma_{ij} \rangle$ components are computed as follows for these tests

$$\langle \varepsilon_{ij} \rangle = \frac{1}{V} \int_{\Omega} \varepsilon_{ij} \, d\Omega \qquad \langle \sigma_{ij} \rangle = \frac{1}{V} \int_{\Omega} \sigma_{ij} \, d\Omega \tag{2.1}$$

Assuming the constitutive law of the form

$$\langle \boldsymbol{\sigma} \rangle = \mathbf{C}_{\mathit{eff}} \langle \boldsymbol{\varepsilon} \rangle \tag{2.2}$$

one can obtain the effective tensor of material parameters $\mathbf{C}_{\mathit{eff}}$. Depending on the expected material behavior, various forms of this tensor can be adopted. In the case of asphalt concrete, the choice between isotropy or anisotropy seems reasonable. Such effective tensors of material parameters should be assessed for every RVE. The boundary value problem at the macroscale is finally solved using these effective tensors.

2.2. Analysis enhancements

In our study, we propose additional enhancements of the methodology presented in the previous Section. Firstly, we use the higher order approximation at the macroscale level (hierarchical shape functions used). Secondly, we modify the boundary conditions used for numerical tests performed for the specified RVE. This modification is based on the observation that the numerical tests are to reflect the behavior of the heterogeneous subdomain from the interior of the whole analyzed domain. Hence, the boundary conditions should account for the heterogeneity of the material along the boundaries. The idea is schematically presented on the example of the shear test in Fig. 4. Instead of the standard boundary conditions marked with red dashed lines, we use their modified versions marked with blue continuous lines.

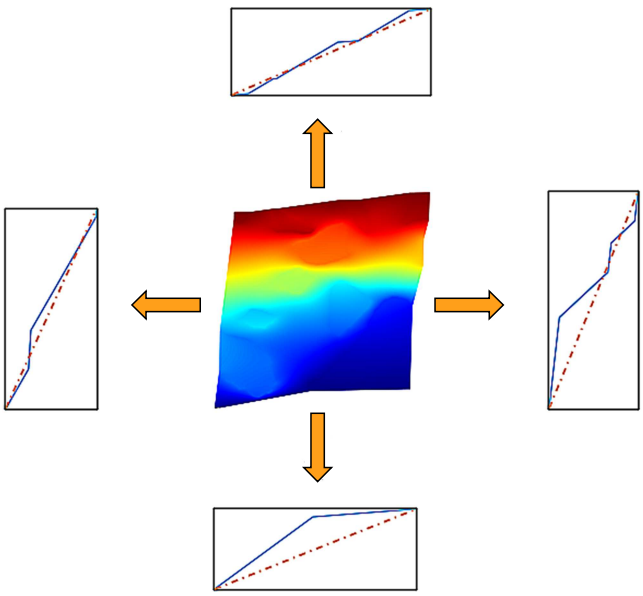


Fig. 4. Modified Dirichlet boundary conditions used for a specific RVE test

Modifications of the boundary conditions are performed in the spirit of the multiscale finite element method (Klimczak and Cecot, 2020b). Namely, we solve auxiliary boundary value problems along the edges of the RVE, where at least two components can be distinguished.

Practically, in the examples presented in this paper, we used 4 types of Dirichlet boundary conditions. They are schematically shown in Fig. 3 and represent tensile and shear tests in both

directions. We used the prescribed displacement of 5% for all four RVE tests. In the case of shear tests, it was the maximum value (c.f. Fig. 4).

On the basis of our experience with the multiscale finite element method, MsFEM (Klimczak and Cecot, 2020a,b), we were aware of the influence of heterogeneities occurring along the domain boundary on the homogenized solution. In terms of the MsFEM, we modify the standard shape functions in order to account for the domain (and also its boundary) heterogeneity. That method results in the assessment of effective macroscale stiffness matrices that contain information on the homogenized material properties.

In terms of the RVE-based homogenization methods, the influence of the heterogeneous boundary can be also accounted for using a buffer (homogeneous) zone with an experimentally adjusted width. This approach shares the similar observation that was the basis of our research. Namely, the subdomain occupied by the RVE (when the whole domain is subject to the load) does not exhibit the response being just a copy of the macroscale boundary conditions. Even in the case of the tensile test, the response within the domain is not linear due to its heterogeneity. Thus, it should be accounted for in the RVE analysis. Motivated by the MsFEM approach, we modify Dirichlet boundary conditions used for all 4 RVE tests. Considering the enforced displacements (see Fig. 3) as functions ψ, we propose below the method for their modifications that covers the microstructure heterogeneity along every RVE edge. Graphically, the idea is shown in Fig. 4 on the example of a shear test. Therein, one can observe the effect of our algorithm performance with respect to the standard boundary conditions used for the corresponding RVE test.

Given ψ, which is a standard scalar function describing Dirichlet boundary conditions, we look for its scalar-valued counterpart φ that is a discrete solution of the following Dirichlet boundary value problem with $\varphi \in C^0(\Omega)$

$$\begin{aligned} &\frac{d}{ds}(2\mu+\lambda)\frac{d\varphi}{ds} = 0 \qquad \forall s \in (0,l) \\ &\varphi = \psi \qquad \text{on} \quad \partial\Omega \end{aligned} \tag{2.3}$$

where μ and λ are Lamé constants, Ω is the analyzed RVE domain and l is length of the analyzed RVE edge.

Schematically, the proposed framework consists of the following steps:

- Generation of a set of RVE's.
- Solution of the auxiliary boundary value problems along the RVE edges with two phases present.
- Solutions of the microscale numerical tests with modified boundary conditions for every RVE, which leads to assessment of the effective parameters tensors $\mathbf{C}_{\mathit{eff}}$.
- Solutions of the macroscale problem with effective parameter tensors used at the integration points.

For the numerical tests presented in this paper, we used also some additional assumptions. In order to provide a reliable comparison between direct and multiscale approaches, we proceed as follows:

- A coarse mesh is generated at the macroscale level.
- Each of the coarse mesh elements is treated as the RVE i.e., a mesh refinement is performed within every RVE in order to account for the underlying microstructure.
- Effective parameters tensors $\mathbf{C}_{\mathit{eff}}$ are assessed for every RVE/coarse mesh element.
- The macroscale problem is solved using these effective tensors for integration.

For the sake of further comparison, for direct analysis we globally generate a fine mesh that is a union of fine meshes generated within RVE's/coarse mesh elements. In such a way, the asphalt concrete microstructure is accounted for with the same precision for both the direct and multiscale analysis. At the macroscale level, we use the higher order approximation to increase the accuracy of the solution.

3. Numerical results

In order to illustrate the efficiency of the proposed approach, we present two numerical tests: with a periodic AC synthetic microstructure and with a non-periodic one recognized from a high-quality image. For the sake of simplicity, we use dimensionless quantities for the tests.

3.1. Test 1: periodic AC microstructure

The periodic microstructure generated for this test was prepared according to the procedure presented in (Klimczak and Cecot, 2020b). Namely, a gradation curve for the aggregate is selected first. Then, a packing of spheres/circles algorithm is used to populate the oversampled domain with spheres of diameters corresponding to the gradation curve. Subsequently, only the centers of spheres/circles belonging to the analyzed domain (a subdomain of the oversampled domain) are left. They are copied in a 3x3 pattern and serve as seeds of Voronoi tessellation. Finally, only the microstructure cut out from the analyzed domain is left (see Fig. 5). Such an RVE local microstructure can be copied in both directions creating a periodic microstructure.

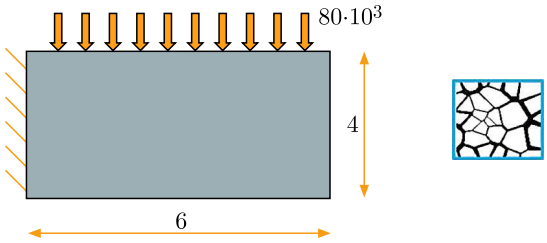


Fig. 5. Macroscale boundary value problem (left) and the periodic RVE (right)

In this test, we created an exemplary microstructure consisting of 24 RVE's shown in Fig. 5. Boundary conditions are also presented therein. The problem is analyzed on the assumption that both the aggregate particles and the binder are linear elastic. Additionally, the plane strain state is assumed. Young's moduli are equal to $80 \cdot 10^9$ (aggregate) and $10 \cdot 10^9$ (binder). Poisson's ratios are equal to 0.35 (aggregate) and 0.3 (binder). In Fig. 6, the domain microstructure obtained by copying the RVE in a 4×6 pattern and the corresponding fine mesh are shown. Since all RVE's exhibit the same microstructure, the tensor of effective parameters was assessed only once and used for every coarse mesh element.

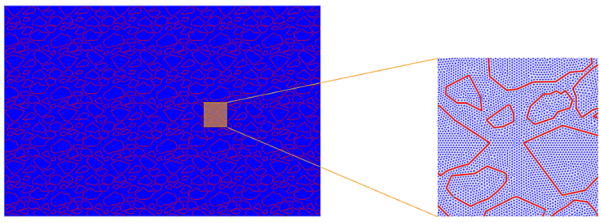


Fig. 6. Domain microstructure and the zoomed-in fine mesh

In Fig. 7, we present the comparison between the reference solution (top row, $p = 1$, NDOF = 500000) and the homogenized one (central row) obtained for the approximation order $p = 1$ at the macroscale. Additionally, the absolute error is presented in the bottom row. The coarse mesh consists of 24 rectangular elements.

It should be emphasized that the reference solution was obtained using approximately half a million degrees of freedom. The error convergence for this test is shown in Fig. 8. It can be observed that a substantial reduction of degrees of freedom was obtained in the test even for the linear approximation used at the macroscale level. Modifications of the boundary conditions

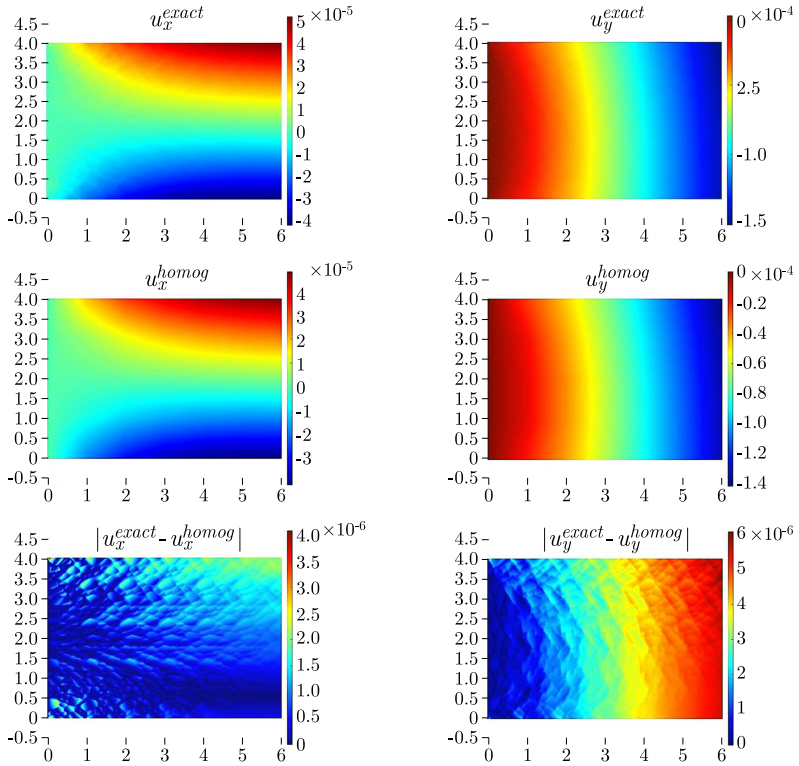


Fig. 7. Comparison between the reference and selected homogenized solutions – Test 1 (reference solution - top row, homogenized solution for the linear approximation at the macroscale – central row, absolute error – bottom row)

in the RVE tests increased the accuracy of the results. It can be seen that the convergence rate for the "modified BC's" solution (blue line) is faster than for the "standard BC's" solution (remaining lines).

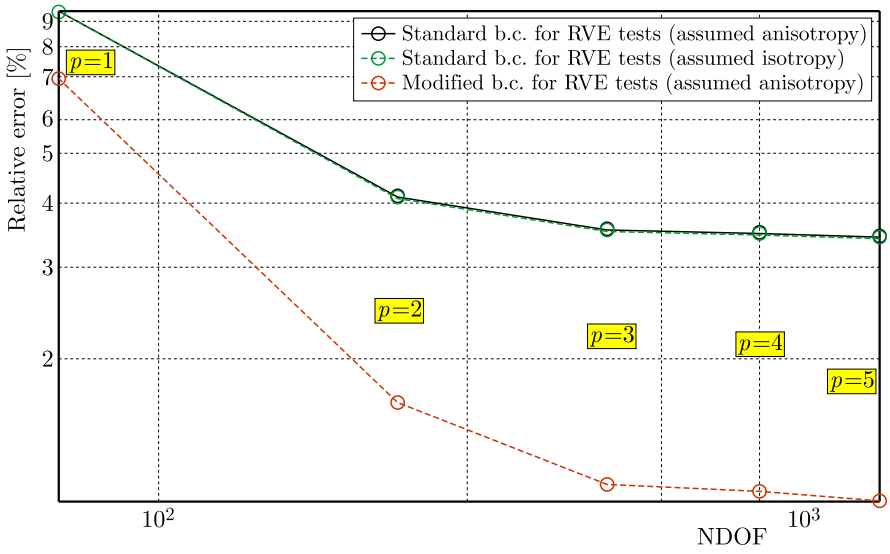


Fig. 8. Error p-convergence (logarithmic scale)

For the standard boundary conditions used in RVE tests, we additionally examined the influence of the assumption on anisotropy/isotropy of asphalt concrete. This effect was negligible (green and black lines).

3.2. Test 2: non-periodic AC microstructure

In this test, we generated the AC microstructure from a high-quality image of the real specimen. The scheme of image processing used for the microstructure recognition is presented in Fig. 9. The AC image shown therein is the actual one used in this test. It is also the case of the resulting microstructure geometry.

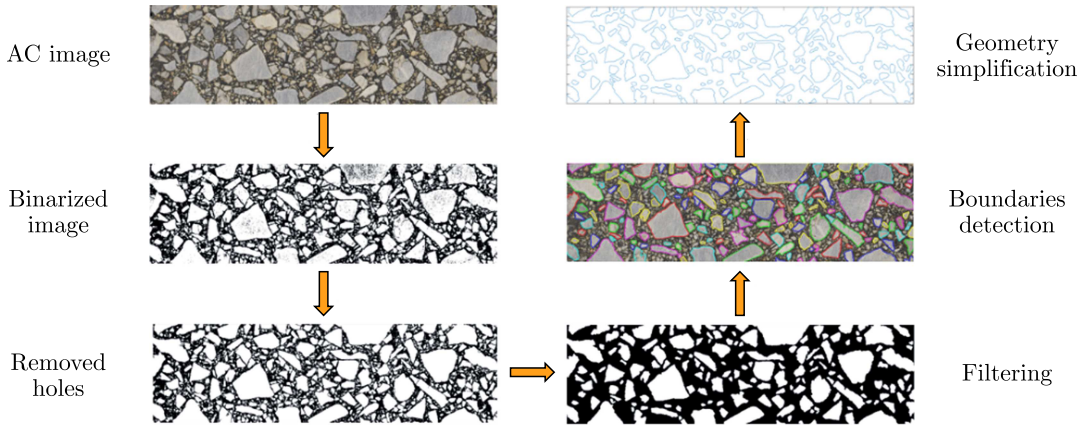


Fig. 9. Recognition of the AC microstructure using image processing

The first step in the image processing is image binarization. Small objects ("holes") within larger subdomains are removed subsequently. The next step is typically the filtering process that allows for further microstructure simplification. Namely, the objects with an area smaller than a threshold value are removed. Subsequently, the boundaries of aggregate particles are detected. At this level, the microstructure geometry can be further simplified due to possible boundaries processing. For the sake of this test, we used a moderate simplification in order to avoid excessively dense finite element mesh.

Material data for this test are the same as for the previous one, whereas the boundary conditions and domain dimensions are different (see Fig. 10).

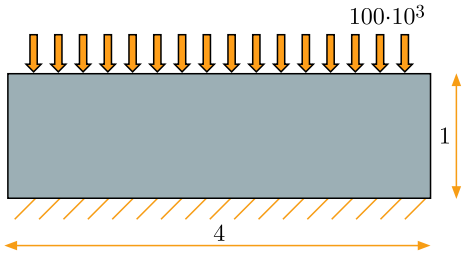


Fig. 10. Analyzed domain

Since the microstructure geometry is non-periodic, tensors of effective material parameters were assessed for each coarse mesh element independently. In this test, we also investigated the effect of the increasing approximation order at the macroscale level. In addition, we generated 2 coarse meshes consisting of 1×4 and 2×8 square elements, respectively. It is to present the necessity of careful RVE size selection.

In Fig. 11, the comparison between the reference and the homogenized solutions obtained for 2×8 coarse elements discretization at the macroscale is presented for the linear approximation used at both levels. In order to obtain the reference solution, the problem consisting of approximately 100000 degrees of freedom had to be solved.

Qualitatively, all of the homogenized solutions (these shown in Fig. 11 and those skipped for the sake of brevity) are of an acceptable form. Quantitative analysis can be performed using

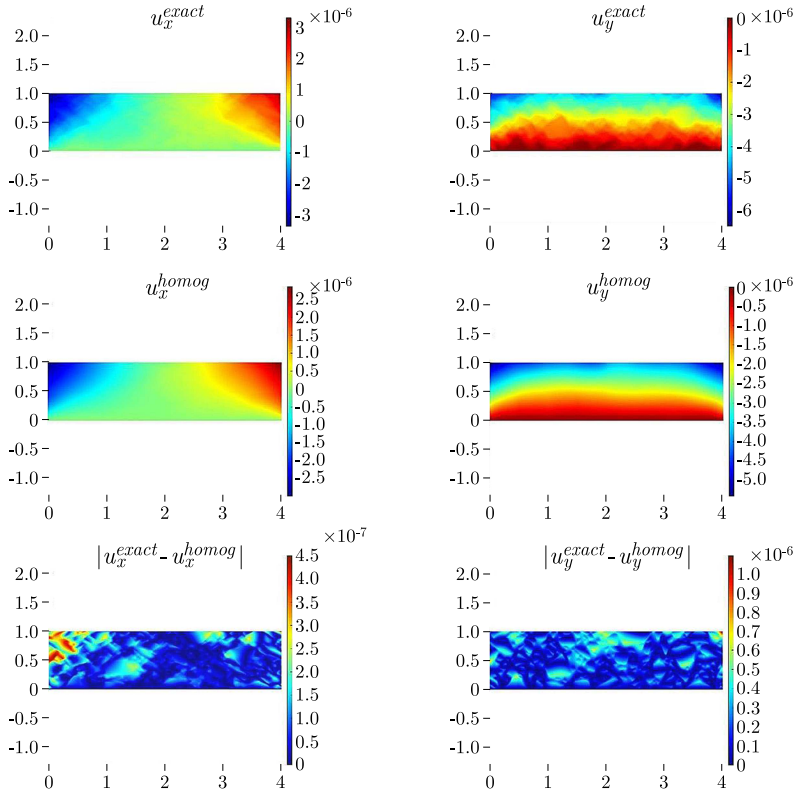


Fig. 11. Comparison between the reference and selected homogenized solutions – Test 2 (reference solution – top row, homogenized solution for the linear approximation at the macroscale – central row, absolute error – bottom row)

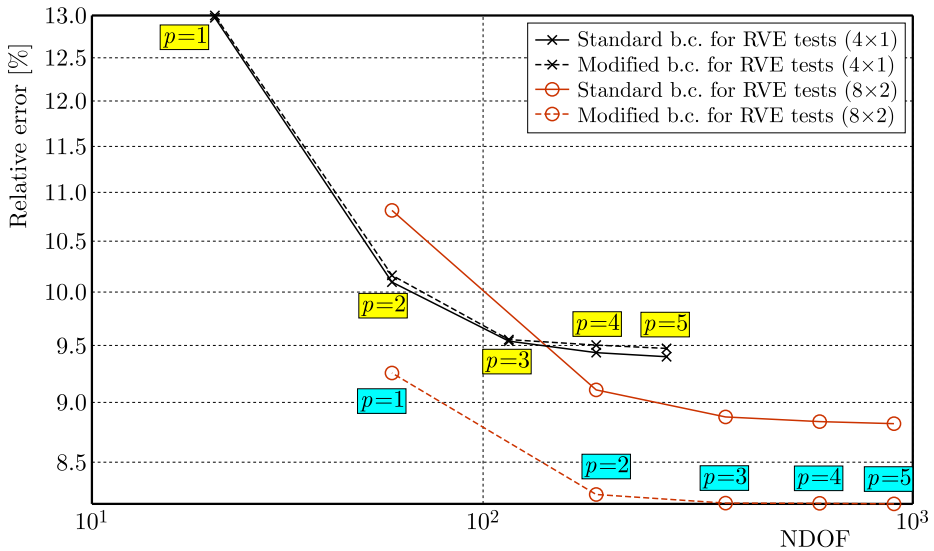


Fig. 12. Error p-convergence (logarithmic scale)

p-convergence for both coarse discretizations (shown in Fig. 12). We present in Fig. 12 the homogenized results obtained both for the standard and modified Dirichlet boundary conditions at the RVE level analysis. The difference between these curves for the coarser (1×4 elements) discretization is very small. The introduced modifications cause even a slight deterioration of the homogenized solution accuracy. The results for this discretization are shown in order to emphasize the necessity of careful selection of the RVE size. All of the approaches based on the

RVE concept require a distinct separation of scales. It refers to two cases: the ratio of the RVE size and the domain size as well as the ratio of the inclusion size to the RVE size. In practical applications, the coarser (1×4 elements) discretization would not be used due to violation of separation of the scales condition. For a more justified discretization (2×8 elements), the modification of Dirichlet boundary conditions at the RVE level improved the solution accuracy.

4. Discussion

Every newly developed numerical method should be effective in a sense of computational resources consumption. The numerical homogenization is evidently a reliable and efficient method. In the tests presented in the previous Section, it allowed for a substantial reduction of the number of degrees of freedom necessary in the analysis. Instead of a direct approach leading to a 500000-NDOF (Test 1) or 100000-NDOF (Test 2) problem, one can solve an equivalent 70-NDOF (Test 1, $p = 1$) or 20-NDOF (Test 2, $p = 1$, 1×4 finite elements) problem. Analyzing the introduced additional modeling error, approximately 7% and 13% for these two cases were measured in L_2 norm, respectively. In some initial tests, even such values are of an acceptable order. As it was demonstrated in Figs. 8 and 12 that the solution accuracy can be easily improved by a higher order approximation and modification of Dirichlet boundary conditions in the RVE analysis.

The numerical effectiveness of the proposed framework can be summarized as follows:

- RVE analyses are performed independently, hence this process can be parallelized.
- Assessment of the effective tensors of material parameters is performed only once, an increase of the approximation order at the macroscale does not require any updates, since standard shape functions are used.
- For a periodic microstructure, only one effective tensor of material parameters needs to be assessed.
- Modification of the boundary conditions for the RVE analysis is relatively cheap, since it is performed only along the boundaries.

5. Conclusions

In this paper, we presented the application of the newly developed version of numerical homogenization to linear elastic analysis of asphalt concrete. The main novelties of the proposed approach are as follows:

- Application of a higher order approximation at the macroscale level.
- Enhancement of the standard numerical homogenization due to modification of Dirichlet boundary conditions used for RVE analysis.
- Solution of two numerical tests with periodic and non-periodic asphalt concrete microstructures.

The obtained results confirmed the applicability of the developed method to linear elastic analyses of asphalt concrete. Precisely, we demonstrated that a reduction of the number of degrees of freedom by several orders of magnitude introduced the error at the acceptable level (maximum 13% was measured).

Our future research plan is to apply this approach to nonlinear analyses of asphalt concrete. In the limit of small displacements, we developed some alternative methods (Klimczak and Cecot, 2020a; Oleksy and Cecot, 2015). Therein, weak formulations of nonlinear problems were expressed in such a way that the terms corresponding to inelastic strains were added to the right--hand side (load vector). The stiffness matrix remained the same for all nonlinear iterations. In

terms of numerical homogenization, it means that the most time-consuming computations of $\mathbf{C}_{eff}$ would be performed only once, as for the linear elasticity.

References

1. Aigner E., Lackner R., Pichler Ch., 2009, Multiscale prediction of viscoelastic properties of asphalt concrete, *Journal of Materials in Civil Engineering*, **21**, 12, 771-780
2. Belytschko T., de Borst R., 2010, Multiscale methods in computational mechanics, *International Journal for Numerical Methods in Engineering*, **83**, 8-9, 939-939
3. Collop A.C., Scarpas A.T., Kasbergen C., de Bondt A., 2003, Development and finite element implementation of a stress dependent elasto-visco-plastic constitutive model with damage for asphalt, *Transportation Research Record: Journal of the Transportation Research Board*, **1832**, 96-104
4. Fakhari Tehrani F., Absi J., Allou F., Petit Ch., 2013, Heterogeneous numerical modeling of asphalt concrete through use of a biphasic approach: Porous matrix/inclusions, *Computational Materials Science*, **69**, 186-196
5. Feyel F., Chaboche L., 2000, FE2 multiscale approach for modelling the elasto-visco-plastic behaviour of long fibre SiC/Ti composite materials, *Computer Methods in Applied Mechanics and Engineering*, **183**, 309-330
6. Fish J., 2014, *Practical Multiscaling*, John Wiley & Sons, Ltd, Chichester
7. GDDKiA, 2014, Asphalt pavement structures on national roads WT-2, Part I (in Polish), Warsaw
8. Guedes J.M., Kikuchi N., 1990, Preprocessing and postprocessing for materials based on the homogenization method with adaptive finite element methods, *Computer Methods in Applied Mechanics and Engineering*, **83**, 143-198
9. Kim Y.R., Souza F.V., Teixeira J.E.S.L., 2013, A two-way coupled multiscale model for predicting damage-associated performance of asphaltic roadways, *Computational Mechanics*, **51**, 2, 187-201
10. Klimczak M., Cecot W., 2020a, Synthetic microstructure generation and multiscale analysis of asphalt concrete, *Applied Sciences*, **10**, 765
11. Klimczak M., Cecot W., 2020b, Towards asphalt concrete modeling by the multiscale finite element method, *Finite Elements in Analysis and Design*, **171**, 103367
12. Kouznetsova V., Geers M., Brekelmans W., 2002, Multi-scale constitutive modelling of heterogeneous materials with a gradient-enhanced computational homogenization scheme, *International Journal for Numerical Methods in Engineering*, **54**, 8, 1235-1260
13. Mitra K., Das A., Basu S., 2012, Mechanical behavior of asphalt mix: An experimental and numerical study, *Construction and Building Materials*, **27**, 1, 545-552
14. Mo L.T., Huurman M., Wu S.P., Molenaar A.A.A., 2008, 2D and 3D meso-scale finite element models for ravelling analysis of porous asphalt concrete, *Finite Elements in Analysis and Design*, **44**, 4, 186-196
15. Mori T., Tanaka K., 1973, Average stress in matrix and average elastic energy of materials with misfitting inclusions, *Acta Metallurgica*, **21**, 571-574
16. Oleksy M., Cecot W., 2015, Application of the fully automatic hp-adaptive FEM to elastic-plastic problems, *Computer Methods in Materials Science*, **15**, 204-212
17. Schüller T., Jänicke R., Steeb H., 2016, Nonlinear modeling and computational homogenization of asphalt concrete on the basis of XRCT scans, *Construction and Building Materials*, **109**, 96-108

18. Wimmer J., Stier B., Simon J.-W., Reese S., 2016, Computational homogenisation from a 3D finite element model of asphalt concrete – linear elastic computations, *Finite Elements in Analysis and Design*, **110**, 43-57

19. Woldekidan M., Huurman M., Pronk A., 2013, Linear and nonlinear viscoelastic analysis of bituminous mortar, *Transportation Research Record: Journal of the Transportation Research Board*, **2370**, 1, 53-62

20. Ziaei-Rad V., Nouri N., Ziaei-Rad S., Abtahi M., 2012, A numerical study on mechanical performance of asphalt mixture using a meso-scale finite element model, *Finite Elements in Analysis and Design*, **57**, 81-91

21. Zohdi T.I., Wriggers P. (Edit.), 2005, *An Introduction to Computational Mechanics*, Springer Berlin, Heidelberg

Manuscript received October 29, 2023; accepted for print January 22, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 365-376, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185888

THE NUMERICAL METHODS FOR SOLVING OF THE ONE-DIMENSIONAL ANOMALOUS REACTION-DIFFUSION EQUATION[1]

Marek Błasik

Silesian University of Technology, Department of Mathematics Applications and Methods for Artificial Intelligence, Poland
e-mail: marek.blasik@polsl.pl

This paper presents numerical methods for solving the one-dimensional fractional reaction--diffusion equation with the fractional Caputo derivative. The proposed methods are based on transformation of the fractional differential equation to an equivalent form of a integro--differential equation. The paper proposes an improvement of the existing implicit method, and a new explicit method. Stability and convergence tests of the methods were also conducted.

Keywords: fractional derivatives and integrals, integro-differential equations, numerical methods, anomalous diffusion

1. Introduction

Anomalous diffusion refers to a type of random motion or transport process that deviates from classical, normal, or Brownian diffusion behavior. In classical diffusion, particles move randomly and independently, following a Gaussian distribution. Anomalous diffusion is characterized by a non-Gaussian distribution (Metzler and Klafter, 2000, 2004;), and may involve mechanisms such as hindered motion or trapping.

The importance of anomalous diffusion in research lies in its prevalence in various natural and artificial systems as well as its implications for understanding complex physical (Solomon *et al.*, 1993; Weeks *et al.*, 1996; Kosztołowicz *et al.*, 2005a, 2005b) and environmental (Humphries *et al.*, 2010) processes.

The anomalous reaction-diffusion equation is a mathematical model that describes spatiotemporal evolution of a quantity such as concentration in a system where both diffusion and reaction processes are affected by anomalous behavior. This equation is commonly used to study how concentrations of substances change over time and space due to both diffusion and chemical reactions (Owolabi *et al.*, 2020; Haq *et al.*, 2021).

A number of numerical methods have been proposed to solve the anomalous reaction-diffusion equation, recent results devoted to this problem are contained in the papers (Coronel-Escamilla *et al.*, 2018; Liu *et al.*, 2015; Pradip and Prasad Goura, 2023; Saad and Gomez-Aguilar, 2018; Sandip and Srinivasan, 2023; Saxena *et al.*, 2015).

The article presents novel numerical techniques designed to solve the one-dimensional fractional reaction-diffusion equation with the fractional Caputo derivative. In particular, the paper introduces improvements to the existing implicit method and introduces an entirely new explicit method. Comprehensive tests to evaluate the stability (the algorithm presented in the paper is a novel approach) and convergence of these new methods are also presented.

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

2. Preliminaries

This Section includes definitions of the left-sided Riemann-Liouville integral and the left-sided Caputo derivative. Both definitions, along with the composition rule of the aforementioned operators, are taken from the monograph (Kilbas *et al.*, 2006) and written in the context of subdiffusion, i.e., for $\alpha \in (0,1]$.

Definition 1. *The left-sided Riemann-Liouville integral of order α, denoted as I^{α}_{0+}, is given by the following formula for* $\mathrm{Re}\,(\alpha) \in (0,1]$

$$I^{\alpha}_{0+} f(t) := \frac{1}{\Gamma(\alpha)} \int_0^t \frac{f(\tau)d\tau}{(t-\tau)^{1-\alpha}} \tag{2.1}$$

where Γ is the Euler gamma function.

Definition 2. *Let $Rem(\alpha) \in (0,1]$. The left-sided Caputo derivative of order α is given by the formula*

$${}^{C}D^{\alpha}_{0+} f(t) := \begin{cases} \dfrac{1}{\Gamma(1-\alpha)} \int\limits_0^t \dfrac{f'(\tau)}{(t-\tau)^{\alpha}}\, d\tau & \text{for} \quad 0 < \alpha < 1 \\ \dfrac{df(t)}{dt} & \text{for} \quad \alpha = 1 \end{cases} \tag{2.2}$$

Property 1. *Let function $f \in C^1(0,T)$. Then, the composition rule for the left-sided Riemann--Liouville integral and the left-sided Caputo derivative is given as follows*

$$I^{\alpha}_{0+}\, {}^{C}D^{\alpha}_{0+} f(t) = f(t) - f(0) \tag{2.3}$$

The definition of the Mittag-Leffler function that is used in the remainder of this article is taken from the monograph (Podlubny, 1999).

Definition 3. *Let $\gamma > 0$. The one-parameter Mittag-Leffler function is given as the following series*

$$E_{\gamma}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\gamma k + 1)} \tag{2.4}$$

Definition 4. *Let $\Pi = \{(x,t) : \ x \in [0,L];\ t \in [0,T]\}$ be a continuous region of solutions for the partial differential equation. Then the set $\overline{\Pi} = \{(x_i,t_j) \in \Pi : \ x_i = i\Delta x,\ i \in \{0,1,\ldots,m\},\ \Delta x = L/m;\ t_j = j\Delta t,\ j \in \{0,1,\ldots,n\};\ \Delta t = T/n\}$ we call the rectangular regular mesh described by the set of nodes.*

3. Mathematical formulation and numerical solution of the problem

Consider the one-dimensional anomalous reaction-diffusion equation

$${}^{C}D^{\alpha}_{0+,t} U(x,t) = D_{\alpha} \frac{\partial^2 U(x,t)}{\partial x^2} + Q_{\alpha}(x,t,U) \qquad 0 \leqslant x \leqslant L \qquad 0 \leqslant t \leqslant T \tag{3.1}$$

supplemented with the boundary conditions

$$U(0,t) = f(t) \qquad U(L,t) = g(t) \qquad 0 \leqslant t \leqslant T \tag{3.2}$$

and the initial condition

$$U(x,0) = h(x) \qquad 0 \leqslant x \leqslant L \tag{3.3}$$

3.1. Explicit numerical scheme

The implicit numerical method (Błasik, 2021) has some limitations, namely, it cannot solve an equation in which the source term is of the form $Q_\alpha(x,t,U)$. Therefore, in this Subsection, an explicit method is proposed that will work when the source term depends on the function U. A key role in the proposed approach is played by Property 1, which makes it possible to transform Eq. (3.1) into an equivalent integro-differential equation

$$U(x,t) = U(x,0) + \frac{D_\alpha}{\Gamma(\alpha)} \int_0^t \frac{1}{(t-\tau)^{1-\alpha}} \frac{\partial^2 U(x,\tau)}{\partial x^2}\, d\tau + \frac{1}{\Gamma(\alpha)} \int_0^t \frac{Q_\alpha(x,\tau,U)}{(t-\tau)^{1-\alpha}}\, d\tau \tag{3.4}$$

Further considerations will be carried out by taking into account the grid of nodes specified in Definition 4. For every node in the grid, we ascertain discrete representation of the integral kernel in the integrals mentioned in Eq. (3.4) on the right-hand side. To achieve this, we estimate the solution U by employing a constant function between two successive nodes in relation to the variable t (Diethelm, 2010) as follows $\overline{U}(x,t) = U(x,t_j)$ for $t_j \leqslant t \leqslant t_{j+1}$, $j = 0,\ldots,n-1$. Hence, we have

$$\frac{D_\alpha}{\Gamma(\alpha)} \int_0^{t_k} \frac{1}{(t_k-\tau)^{1-\alpha}} \frac{\partial^2 U(x,\tau)}{\partial x^2}\, d\tau \approx \frac{D_\alpha}{\Gamma(\alpha)} \int_0^{t_k} \frac{1}{(t_k-\tau)^{1-\alpha}} \frac{\partial^2 \overline{U}(x,\tau)}{\partial x^2}\, d\tau \tag{3.5}$$

From the additivity of the integral with respect to the integration interval and the approximation of the second order derivative of the function U with respect to the spatial variable by the differential quotient, we obtain

$$\begin{aligned}
&\frac{D_\alpha}{\Gamma(\alpha)} \int_0^{t_k} \frac{1}{(t_k-\tau)^{1-\alpha}} \frac{\partial^2 \overline{U}(x,\tau)}{\partial x^2}\, d\tau = \frac{D_\alpha}{\Gamma(\alpha)} \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \frac{1}{(t_k-\tau)^{1-\alpha}} \frac{\partial^2 U(x,t_j)}{\partial x^2}\, d\tau \\
&= \frac{D_\alpha}{\Gamma(\alpha)} \sum_{j=0}^{k-1} \frac{(t_k-t_j)^\alpha - (t_k-t_{j+1})^\alpha}{\alpha} \frac{\partial^2 U(x,t_j)}{\partial x^2} \\
&= \frac{D_\alpha \Delta t^\alpha}{\Gamma(\alpha+1)} \sum_{j=0}^{k-1} [(k-j)^\alpha - (k-j-1)^\alpha] \frac{\partial^2 U(x,t_j)}{\partial x^2} \\
&= \frac{D_\alpha \Delta t^\alpha}{\Gamma(\alpha+1)} \sum_{j=0}^{k-1} [(k-j)^\alpha - (k-j-1)^\alpha] \frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{(\Delta x)^2} \\
&= D_\alpha \sum_{j=0}^{k-1} \frac{r_{j,k}}{\Delta x^2} (U_{i-1,j} - 2U_{i,j} + U_{i+1,j})
\end{aligned}$$

where the discrete form of the kernel of the left-sided Riemann-Liouville integral is given by the formula

$$r_{j,k} = \frac{\Delta t^\alpha}{\Gamma(\alpha+1)} [(k-j)^\alpha - (k-j-1)^\alpha] \tag{3.6}$$

Repeating the same discretization for the second integral on the right-hand side of equation (3.4), we get

$$\frac{1}{\Gamma(\alpha)} \int_0^t \frac{Q_\alpha(x,\tau,U)d\tau}{(t-\tau)^{1-\alpha}} \approx \frac{1}{\Gamma(\alpha)} \int_0^t \frac{\overline{Q}_\alpha(x,\tau,U)}{(t-\tau)^{1-\alpha}}\, d\tau \tag{3.7}$$

and sequentially

$$\begin{aligned}\frac{1}{\Gamma(\alpha)}\int_0^t \frac{\overline{Q}_\alpha(x,\tau,U)}{(t-\tau)^{1-\alpha}}\,d\tau &= \frac{1}{\Gamma(\alpha)}\sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}} \frac{1}{(t_k-\tau)^{1-\alpha}} Q_\alpha(x,t_j,U)\,d\tau \\ &= \frac{1}{\Gamma(\alpha)}\sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}} \frac{(t_k-t_j)^\alpha-(t_k-t_{j+1})^\alpha}{\alpha} Q_\alpha(x,t_j,U)\,d\tau \\ &= \frac{\Delta t^\alpha}{\Gamma(\alpha+1)}\sum_{j=0}^{k-1}[(k-j)^\alpha-(k-j-1)^\alpha])Q_\alpha(x_i,t_j,U_{i,j}) = \sum_{j=0}^{k-1} r_{j,k}Q_\alpha(x,t_j,U)\end{aligned} \tag{3.8}$$

Finally, we get an explicit numerical scheme

$$U_{i,k} = U_{i,0} + D_\alpha\sum_{j=0}^{k-1}\frac{r_{j,k}}{\Delta x^2}(U_{i-1,j}-2U_{i,j}+U_{i+1,j}) + \sum_{j=0}^{k-1} r_{j,k}Q_{\alpha i,j} \tag{3.9}$$

3.2. Implicit numerical scheme

The numerical method proposed in this Subsection is a modification of the method presented in the paper (Błasik, 2021). The main change consists in a different way of discretizing the source term of the equation. The expression $\sum_{j=0}^{k} w_{j,k}Q_{\alpha i,j}$ has been replaced by $\sum_{j=0}^{k-1} r_{j,k}Q_{\alpha i,j}$, look at the last component of equation (3.10). We can write the improved implicit numerical scheme in the form

$$\begin{aligned}&-\frac{D_\alpha w_{k,k}}{(\Delta x)^2}U_{i-1,k} + \Big(1+\frac{2D_\alpha w_{k,k}}{(\Delta x)^2}\Big)U_{i,k} - \frac{D_\alpha w_{k,k}}{(\Delta x)^2}U_{i+1,k} \\ &\quad = U_{i,0} + \sum_{j=0}^{k-1}\frac{D_\alpha w_{j,k}}{(\Delta x)^2}(U_{i-1,j}-2U_{i,j}+U_{i+1,j}) + \sum_{j=0}^{k-1} r_{j,k}Q_{\alpha i,j}\end{aligned} \tag{3.10}$$

where

$$w_{j,k} := \frac{(\Delta t)^\alpha}{\Gamma(2+\alpha)}\begin{cases}(\alpha+1-k)k^\alpha+(k-1)^{\alpha+1} & j=0\\ (k-j+1)^{\alpha+1}-2(k-j)^{\alpha+1}+(k-j-1)^{\alpha+1} & 0<j<k\\ 1 & j=k\end{cases} \tag{3.11}$$

Note that now we do not need to know the value of the source term at the k-th time layer. So the scheme can be written as a system of $n-1$ algebraic equations in the matrix form

$$\mathbf{A}\mathbf{U}_k = \mathbf{B} \tag{3.12}$$

where matrices $\mathbf{A}$ and $\mathbf{B}$ are defined as

$$\mathbf{A} = \begin{bmatrix} 1+2a & -a & 0 & 0 & \cdots & 0 & 0 & 0 \\ -a & 1+2a & -a & 0 & \cdots & 0 & 0 & 0 \\ 0 & -a & 1+2a & -a & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -a & 1+2a & -a & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -a & 1+2a & -a \\ 0 & 0 & 0 & 0 & \cdots & 0 & -a & 1+2a \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_1 + aU_{0,k} \\ b_2 \\ b_3 \\ \vdots \\ b_i \\ \vdots \\ b_{m-2} \\ b_{m-1} + aU_{m,k} \end{bmatrix}$$

The elements of matrices $\mathbf{A}$ and $\mathbf{B}$ are defined by the formulas

$$a := \frac{D_\alpha w_{k,k}}{(\Delta x)^2}$$

$$b_i := U_{i,0} + \sum_{j=0}^{k-1} \frac{D_\alpha w_{j,k}}{\Delta x^2}(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}) + \sum_{j=0}^{k-1} r_{j,k} Q_{\alpha i,j}$$

4. Numerical examples

This Section presents numerical results which are compared with a closed exact solution. In the calculations, the following form of the analytical solution was adopted

$$U(x,t) = \frac{1}{2}\sin(x) E_\alpha(t^\alpha) \tag{4.1}$$

From the fact of invariance of the Mittag-Leffler function with respect to the left-sided Caputo derivative, the form of the source term is derived. Thus, we define the initial boundary value problem as

$${}^C D^\alpha_{0+,t} U(x,t) = \frac{\partial^2 U(x,t)}{\partial x^2} + 2U(x,t) \qquad 0 \leqslant x \leqslant \frac{\pi}{2} \qquad t \geqslant 0 \tag{4.2}$$

supplemented with the boundary conditions

$$U(0,t) = 0 \qquad U\Big(\frac{\pi}{2}, t\Big) = \frac{1}{2} E_\alpha t^\alpha \qquad t \geqslant 0 \tag{4.3}$$

and the initial condition

$$U(x,0) = \frac{1}{2}\sin x \qquad 0 \leqslant x \leqslant \frac{\pi}{2} \tag{4.4}$$

where the generalized diffusion coefficient D_α is equal to one. In the calculations, the following mesh parameters were assumed: $T = 0.1$, $L = \pi/2$, $m, n \in \{25, 50, 100, 200\}$, and the order of the left-sided Caputo derivative $\alpha \in \{0.25, 0.5, 0.75, 1\}$.

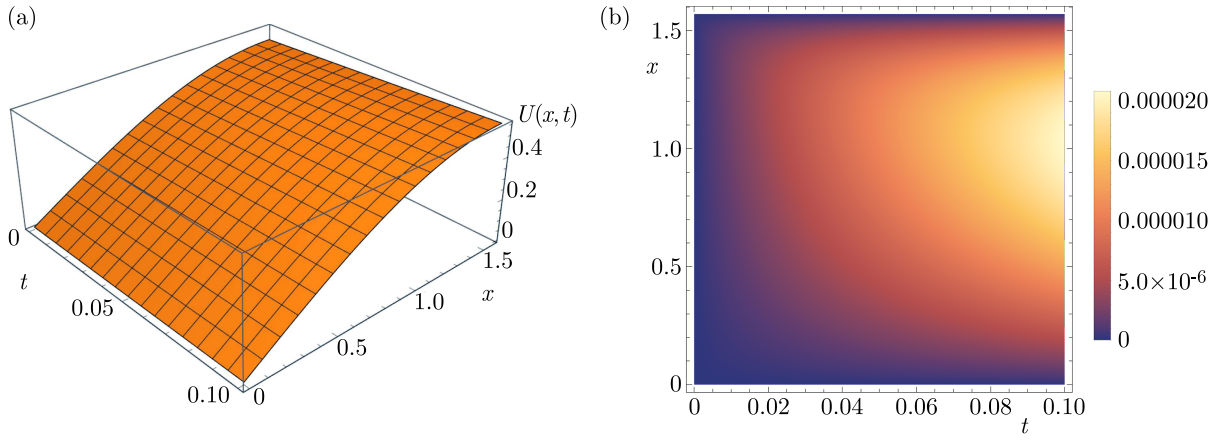


Fig. 1. The numerical solution of the initial-boundary value problem for $\alpha = 1$ (a). The absolute error generated by the numerical method for $\alpha = 1$ (b)

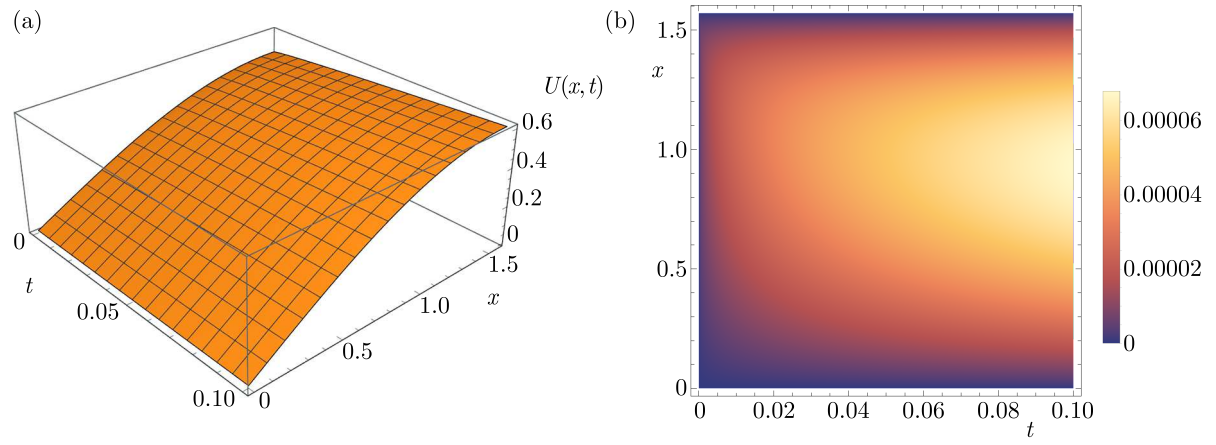


Fig. 2. The numerical solution of the initial-boundary value problem for $\alpha = 75$ (a). The absolute error generated by the numerical method for $\alpha = 75$ (b)

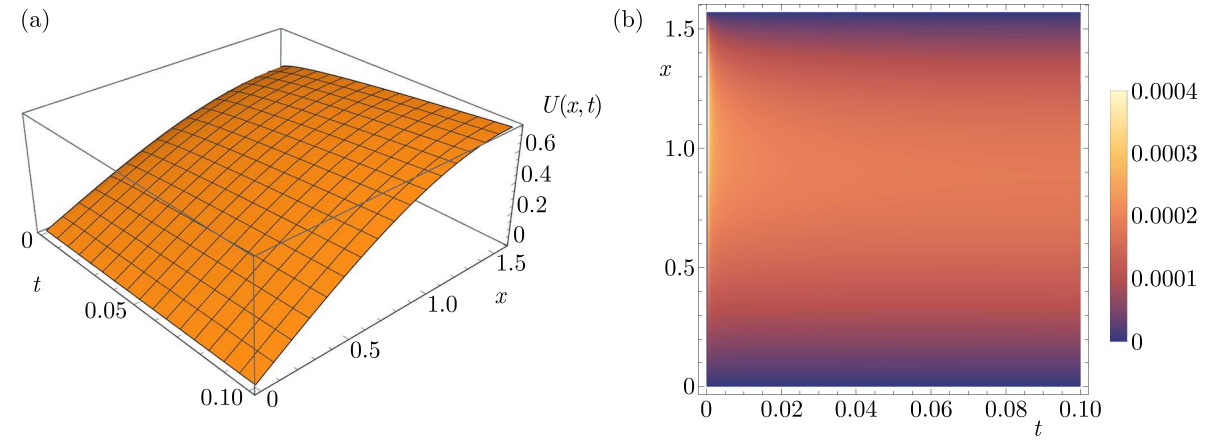


Fig. 3. The numerical solution of the initial-boundary value problem for $\alpha = 0.5$ (a). The absolute error generated by the numerical method for $\alpha = 0.5$ (b)

Figures 1-4, part (a), show the numerical solutions of Eq. (4.2)-(4.4) obtained by the implicit method for four different values of the order of the left-sided Caputo derivative. Part (b) presents the absolute error generated by the proposed method resulting from validation of the numerical scheme with exact solution Eq. (4.1). For large values of the order of the left-sided Caputo

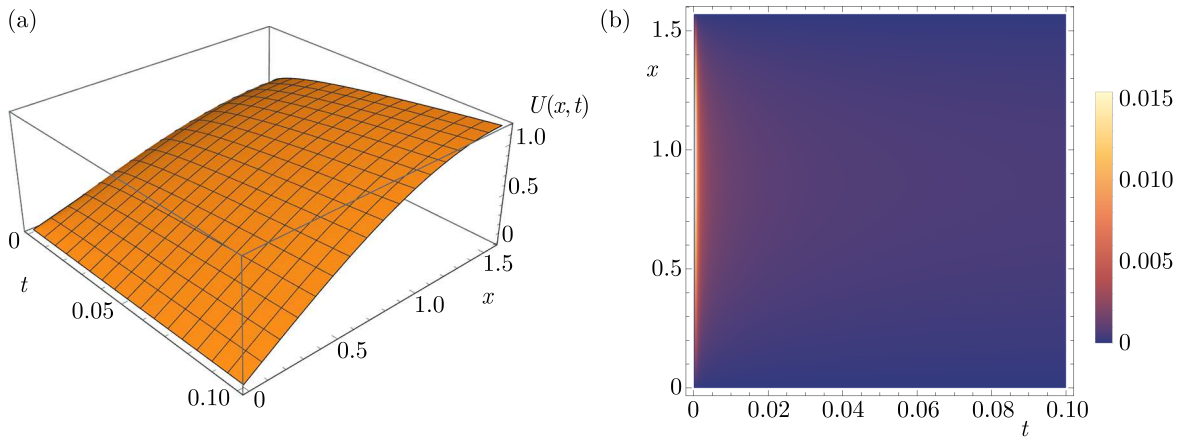


Fig. 4. The numerical solution of the initial-boundary value problem for $\alpha = 0.25$ (a). The absolute error generated by the numerical method for $\alpha = 0.25$ (b)

derivative: $\alpha = 1$ and $\alpha = 0.75$, we observe its smallest value near the boundary conditions, for $x = 0$ and $x = \pi/2$. We also notice that the error accumulates and reaches the maximum value for $t = 0.1$. In the case of $\alpha = 0.5$ and $\alpha = 0.25$, the numerical scheme generates the largest errors for small values of the variable t.

Tables 1-4 present the average absolute errors generated by the implicit numerical scheme for sixteen grid variants and four values of the order of the left-sided Caputo derivative. The results clearly show that the order of the derivative has a significant effect on the accuracy of the numerical method, and the average absolute error is negatively correlated with it.

Table 1. The average absolute error generated by the numerical method for $\alpha = 1$

n \ m	25	50	100	200
25	$4.801 \cdot 10^{-5}$	$5.237 \cdot 10^{-5}$	$5.376 \cdot 10^{-5}$	$5.424 \cdot 10^{-5}$
50	$2.191 \cdot 10^{-5}$	$2.57 \cdot 10^{-5}$	$2.681 \cdot 10^{-5}$	$2.715 \cdot 10^{-5}$
100	$8.81 \cdot 10^{-6}$	$1.232 \cdot 10^{-5}$	$1.328 \cdot 10^{-5}$	$1.356 \cdot 10^{-5}$
200	$2.247 \cdot 10^{-6}$	$5.609 \cdot 10^{-6}$	$6.505 \cdot 10^{-6}$	$6.749 \cdot 10^{-6}$

Table 2. The average absolute error generated by the numerical method for $\alpha = 0.75$

n \ m	25	50	100	200
25	$2.298 \cdot 10^{-4}$	$2.415 \cdot 10^{-4}$	$2.457 \cdot 10^{-4}$	$2.474 \cdot 10^{-4}$
50	$1.099 \cdot 10^{-4}$	$1.191 \cdot 10^{-4}$	$1.221 \cdot 10^{-4}$	$1.231 \cdot 10^{-4}$
100	$5.02 \cdot 10^{-5}$	$5.815 \cdot 10^{-5}$	$6.049 \cdot 10^{-5}$	$6.124 \cdot 10^{-5}$
200	$2.049 \cdot 10^{-5}$	$2.782 \cdot 10^{-5}$	$2.985 \cdot 10^{-5}$	$3.044 \cdot 10^{-5}$

Table 3. The average absolute error generated by the numerical method for $\alpha = 0.5$

n \ m	25	50	100	200
25	$9.967 \cdot 10^{-4}$	$1.031 \cdot 10^{-3}$	$1.045 \cdot 10^{-3}$	$1.051 \cdot 10^{-3}$
50	$4.809 \cdot 10^{-4}$	$5.047 \cdot 10^{-4}$	$5.134 \cdot 10^{-4}$	$5.169 \cdot 10^{-4}$
100	$2.271 \cdot 10^{-4}$	$2.458 \cdot 10^{-4}$	$2.518 \cdot 10^{-4}$	$2.54 \cdot 10^{-4}$
200	$1.027 \cdot 10^{-4}$	$1.188 \cdot 10^{-4}$	$1.235 \cdot 10^{-4}$	$1.25 \cdot 10^{-4}$

Table 4. The average absolute error generated by the numerical method for $\alpha = 0.25$

n \ m	25	50	100	200
25	$4.952 \cdot 10^{-3}$	$5.084 \cdot 10^{-3}$	$5.144 \cdot 10^{-3}$	$5.171 \cdot 10^{-3}$
50	$2.401 \cdot 10^{-3}$	$2.482 \cdot 10^{-3}$	$2.515 \cdot 10^{-3}$	$2.529 \cdot 10^{-3}$
100	$1.147 \cdot 10^{-3}$	$1.201 \cdot 10^{-3}$	$1.222 \cdot 10^{-3}$	$1.23 \cdot 10^{-3}$
200	$5.359 \cdot 10^{-4}$	$5.784 \cdot 10^{-4}$	$5.922 \cdot 10^{-4}$	$5.972 \cdot 10^{-4}$

4.1. Convergence analysis

The convergence order is a measure of how quickly the grid method approaches the solution to the problem as the number of mesh nodes increases. The experimental convergence order is determined empirically by analyzing the rate at which the error decreases with each increase in the number of nodes. To determine the experimental order of convergence of the implicit numerical method, we use the formula (Gu *et al.*, 2021)

$$\mathrm{EOC} = \log_2 \left(\frac{\frac{1}{(m+1)(n+1)} \sum_{j=0}^{n} \sum_{i=0}^{m} |U(x_i, t_j) - U_{i,j}^{m,n}|}{\frac{1}{(2m+1)(2n+1)} \sum_{j=0}^{2n} \sum_{i=0}^{2m} |U(x_i, t_j) - U_{i,j}^{2m,2n}|} \right) = \log_2 \Big(\frac{\Delta U^{m,n}}{\Delta U^{2m,2n}} \Big) \tag{4.5}$$

where $U(x_i, t_j)$ is the exact solution determined at node $(i\Delta x, j\Delta t)$, while $U_{i,j}^{m,n}$ represents the approximate solution calculated by the numerical method.

The data collected in Tables 5 and 6 clearly show that the experimental order of convergence tends to one very quickly, as m and n increases.

Table 5. Convergence order of the implicit numerical scheme for $\alpha \in \{1, 0.75\}$

n	m	$\alpha = 1$		$\alpha = 0.75$	
		$\Delta U^{m,n}$	EOC	$\Delta U^{m,n}$	EOC
25	25	$4.801 \cdot 10^{-5}$	–	$2.298 \cdot 10^{-4}$	–
50	50	$2.57 \cdot 10^{-5}$	0.902	$1.191 \cdot 10^{-4}$	0.948
100	100	$1.328 \cdot 10^{-5}$	0.953	$6.049 \cdot 10^{-5}$	0.977
200	200	$6.749 \cdot 10^{-6}$	0.977	$3.044 \cdot 10^{-5}$	0.991

Table 6. Convergence order of the implicit numerical scheme for $\alpha \in \{0.5, 0.25\}$

n	m	$\alpha = 1$		$\alpha = 0.75$	
		$\Delta U^{m,n}$	EOC	$\Delta U^{m,n}$	EOC
25	25	$9.967 \cdot 10^{-4}$	–	$4.952 \cdot 10^{-3}$	–
50	50	$5.047 \cdot 10^{-4}$	0.982	$2.482 \cdot 10^{-3}$	0.997
100	100	$2.518 \cdot 10^{-4}$	1.003	$1.222 \cdot 10^{-3}$	1.022
200	200	$1.25 \cdot 10^{-4}$	1.01	$5.972 \cdot 10^{-4}$	1.033

4.2. Stability analysis

During numerical tests, the explicit scheme defined by equation (3.9) showed features of conditional stability. For a certain ratio of the time and spatial step, it generated convergent solutions. After increasing the time step at a fixed spatial step, its divergence was observed. To determine the stability condition, an algorithm was proposed, which is described in this Section.

The stability condition of the numerical scheme is formulated for solution (4.1), which is monotonic in the region in which we solve the differential equation. Thus, considering any three

consecutive nodes with respect to the time or space variable, the value of the solution at the center node should not exceed the values obtained at neighboring nodes – this condition is written in the ninth line of the pseudocode. The algorithm can be described in several points:

1) we initiate the algorithm by specifying values: $\alpha = 1$, $t_{end} = 0.001$, $x_{start} = 0$, $x_{end} = \pi/2$, $m = 10$, $n = 10$,
2) we generate the solution by the explicit scheme and check the stability condition,
3) through the parameter $\lambda_t^2 = 0.(90)$, we modify the time step to obtain an unstable solution,
4) through the parameter $\lambda_t^1 = 0.99$, we slowly reduce the time step until the stability condition is fulfilled,
5) we write the spatial step and the largest possible time step that guarantee stability to the *result* list,
6) we reduce the spatial step with the parameter $\lambda_x = 0.93$ so that $x_{end} > x_{end}^*$, where $x_{end}^* = \pi/4$, and repeat steps 2-5.

The above-mentioned steps of the algorithm resulted in a set of points, which are presented in Fig. 5a. Then the approximation of points with the function $p_1 \Delta x^{p_2}$ was carried out. It should be noted that the fit of the function to the sets of points is almost perfect – the coefficient of determination after rounding to the fourth decimal place gives a value of one.

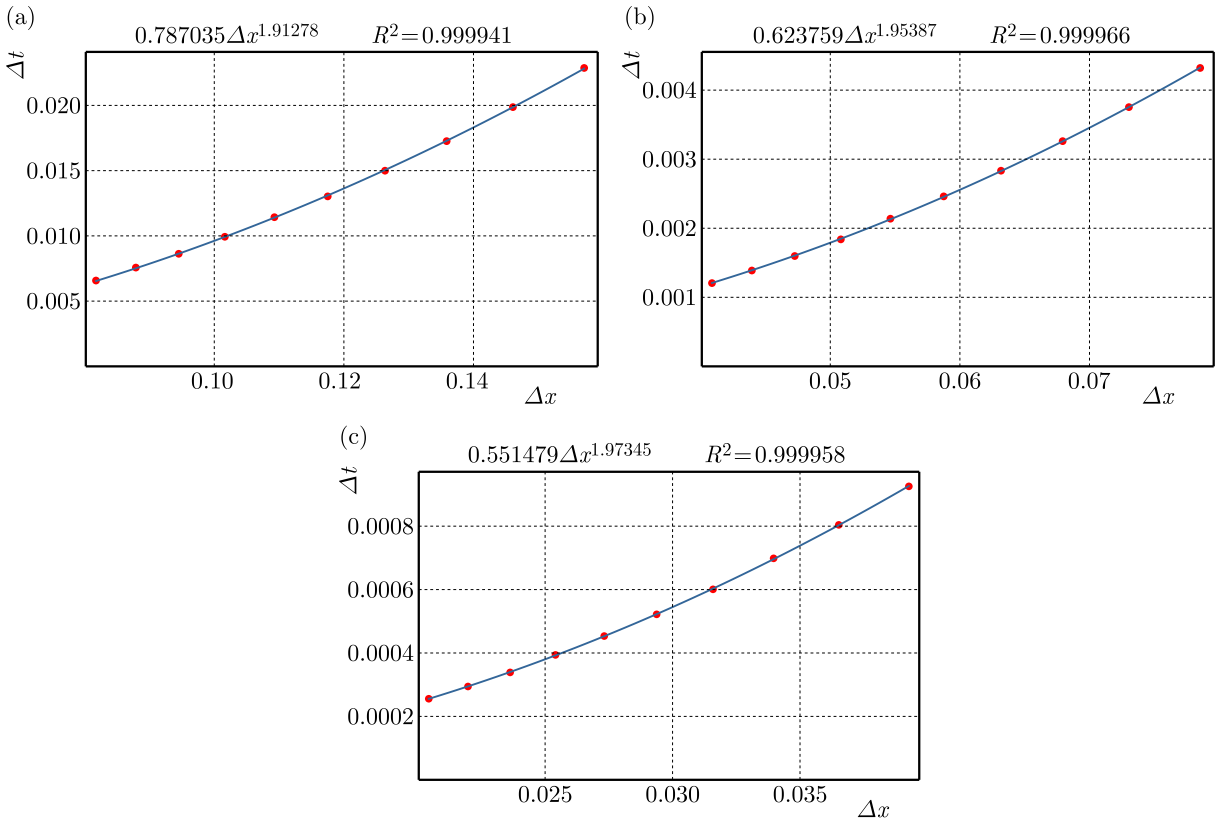


Fig. 5. Relationship between Δx and Δt for $\alpha = 1$ and (a) $m, n = 10$, (b) $m, n = 20$, (c) $m, n = 40$.

The obtained results are presented in Fig. 5 and Table 7. Analysis of the results leads to the following observations. For $\alpha = 1$ and increasing grid parameters m and n, $p_1 \to 0.5$ and $p_2 \to 2$, which is in accordance with the stability criterion for the explicit method for the classical diffusion equation where the relationship $\Delta t \leqslant 0.5 \Delta x^2$ occurs. Analysis of the results in the other cases of Table 7 leads to a more general condition in the form of $\Delta t \leqslant \delta_\alpha \Delta x^{2/\alpha}$.

Algorithm 1: Algorithm to test stability

Input : Initial values of: α, t_{end}, x_{start}, x_{end}, x^*_{end}, λ_x, λ^2_t, λ^1_t, m, n
Output: List of points: *result*

while $x_{end} > x^*_{end}$ **do**
  $sta \leftarrow 0$;
  $counter \leftarrow 1$;
  while $sta \neq 1$ **do**
    Implicit_scheme$(\alpha, t_{end}, x_{start}, x_{end}, m, n)$;
    $stop \leftarrow 0$;
    for $j \leftarrow 1$ **to** $n-1$ **do**
      for $i \leftarrow 1$ **to** $m-1$ **do**
        if $\left(\left|\frac{U_{i-1,j}+U_{i+1,j}}{2} - U_{i,j}\right| < \frac{1}{2}\left|U_{i+1,j} - U_{i-1,j}\right|\right)$ ***and*** $\left(\left|\frac{U_{i,j-1}+U_{i,j+1}}{2} - U_{i,j}\right| < \frac{1}{2}\left|U_{i,j+1} - U_{i,j-1}\right|\right)$ **then**
          $sta \leftarrow 1$;
        end
        else
          $sta \leftarrow 0$;
          $stop \leftarrow 1$;
          Break;
        end
      end
    end
    if $stop == 1$ **then**
      Break;
    end
  end
  if $sta == 1$ ***and*** $counter == 1$ **then**
    $t_{end} \leftarrow t_{end}/\lambda^2_t$;
    Break;
  end
  if $sta == 1$ **then**
    Append $[(x_{end} - x_{start})/m, t_{end}/n]$ to *result*;
  end
  else
    $t_{end} \leftarrow t_{end} \times \lambda^1_t$;
    $counter \leftarrow counter + 1$;
  end
  $x_{end} \leftarrow x_{end} \times \lambda_x$;
end

Table 7. Estimated values of the parameters p_1 and p_2 of the function $p_1 \Delta x^{p_2}$

n	m	$\alpha = 1$			$\alpha = 0.75$			$\alpha = 0.5$		
		p_1	p_2	R^2	p_1	p_2	R^2	p_1	p_2	R^2
10	10	0.787	1.913	1	0.457	2.518	1	0.172	3.736	1
20	20	0.624	1.954	1	0.347	2.588	1	0.127	3.873	1
40	40	0.551	1.973	1	0.307	2.631	1	0.102	3.926	1

5. Conclusions

In this paper, two numerical methods for solving the one-dimensional fractional reaction--diffusion equation were proposed. The explicit method in testing proved to be conditionally stable. The stability condition was determined by the algorithm proposed in the paper, its form $\Delta t \leqslant \delta_\alpha \Delta x^{2/\alpha}$ made the numerical scheme very time-consuming for small values of α. The order of convergence of the implicit method was also estimated, which was one.

References

1. Błasik M., 2021, The implicit numerical method for the one-dimensional anomalous subdiffusion equation with a nonlinear source term, *Bulletin of the Polish Academy of Sciences. Technical Sciences*, **69**, e138240
2. Coronel-Escamilla A., Gómez-Aguilar J.F.,Torres L., Escobar-Jimenéz R.F., 2018, A numerical solution for a variable-order reaction-diffusion model by using fractional derivatives with non-local and non-singular kernel, *Physica A*, **491**, 406-424
3. Diethelm K., 2010, *The Analysis of Fractional Differential Equations*, Springer-Verlag, Berlin
4. Gu X.M., Sun H.W., Zhao Y.L., Zheng X., 2021, An implicit difference scheme for time--fractional diffusion equations with a time-invariant type variable order, *Applied Mathematics Letters*, **120**, 107270
5. Haq S., Ali I., Sooppy Nisar K., 2021, A computational study of two-dimensional reaction--diffusion Brusselator system with applications in chemical processes, *Alexandria Engineering Journal*, **60**, 4381-4392
6. Humphries N.E., Queiroz N., Dyer J.R.M., Pade N.G., Musyl M.K., *et al.*, 2010, Environmental context explains Lévy and Brownian movement patterns of marine predators, *Nature*, **465**, 1066-1069
7. Kilbas A.A., Srivastava H.M., Trujillo J.J., 2006, *Theory and Applications of Fractional Differential Equations*, Elsevier, Amsterdam
8. Kosztołowicz T., Dworecki K., Mrówczyński S., 2005a, How to measure subdiffusion parameters, *Physical Review Letters*, **94**, 170602
9. Kosztołowicz T., Dworecki K., Mrówczyński S., 2005b, Measuring subdiffusion parameters, *Physical Review E*, **71**, 041105
10. Liu Y., Du Y., Li H., Li J., He S., 2015, A two-grid mixed finite element method for a nonlinear fourth-order reaction-diffusion problem with time-fractional derivative, *Computers and Mathematics with Applications*, **70**, 2474-2492
11. Metzler R., Klafter J., 2000, The random walk's guide to anomalous diffusion: a fractional dynamics approach, *Physics Reports*, **339**, 1-77
12. Metzler R., Klafter J., 2004, The restaurant at the end of the random walk: Recent developments in the description of anomalous transport by fractional dynamics, *Journal of Physics A: Mathematical and General*, **37**, 161-208
13. Owolabi K.M., Atangana A., Akgul A., 2020, Modelling and analysis of fractal-fractional partial differential equations: Application to reaction-diffusion model, *Alexandria Engineering Journal*, **59**, 2477-2490
14. Podlubny I., 1999, *Fractional Differential Equations*, Academic Press, San Diego
15. Pradip R., Prasad Goura V.M.K., 2023, An efficient numerical scheme and its stability analysis for a time-fractional reaction diffusion model, *Journal of Computational and Applied Mathematics*, **422**, 114918

16. Saad K.M., Gómez-Aguilar J.F., 2018, Analysis of reaction-diffusion system via a new fractional derivative with non-singular kernel, *Physica A*, **509**, 703-716

17. Sandip M., Srinivasan N., 2023, Analytical and numerical solutions of time-fractional advection--diffusion-reaction equation, *Applied Numerical Mathematics*, **185**, 549-570

18. Saxena R.K., Mathai A.M., Haubold H.J., 2015, Computational solutions of unified fractional reaction-diffusion equations with composite fractional time derivative, *Communications in Nonlinear Science and Numerical Simulation*, **27**, 1-11

19. Solomon T.H., Weeks E.R., Swinney H.L., 1993, Observations of anomalous diffusion and Lévy flights in a 2-dimensional rotating flow, *Physical Review Letters*, **71**, 3975-3979

20. Weeks E.R., Urbach J.S., Swinney L., 1996, Anomalous diffusion in asymmetric random walks with a quasi-geostrophic flow example, *Physica D: Nonlinear Phenomena*, **97**, 291-310

Manuscript received December 27, 2023; accepted for print February 19, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 377-388, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185320

A COMPARISON OF ROBUST AND RELIABILITY BASED DESIGN OPTIMIZATION[1]

Paweł Zabojszcza, Urszula Radoń
Kielce University of Technology, Faculty of Civil Engineering and Architecture, Kielce, Poland
e-mail: pawelzab@tu.kielce.pl; zmbur@tu.kielce.pl

This article compares two optimization methods considering random variations in design parameters. One is reliability-based design optimization, which depends on the availability of the joint probability density function. A more practical alternative is robust optimization, which does not require the estimation of failure probability. It accounts for the random response of the structure through definitions of objective functions and constraints, incorporating mean values and response variances. An important element of the algorithm involves approximating unknown responses of the structures and employing efficient statistical moment estimation methods. The kriging method was used in this paper. Additionally, the article evaluates two experimental plan techniques: the classical random sampling plan and the OLH plan.

Keywords: deterministic optimization, robust optimization, reliability based design optimization, first order reliability method

1. Introduction

Currently, most Finite Element Method (FEM) structural design programs, popular among engineers, also include modules based on the deterministic optimization formulation. The result of optimization is a structure that is characterized by optimal features due to criteria adopted as a measure of their quality. Two factors clearly determine usefulness of the solution obtained this way. One of them is the adequacy of the numerical model itself, which must well reflect the actual physical phenomenon. Failure to meet this condition leads to serious mistakes and, consequently, bad decisions. The second factor is proper formulation of the optimization task. Inappropriate selection of the objective function, design constraints and, above all, calculation methodology may make the optimal solution completely useless.

Analysis of the influence of the random nature of parameters describing the modeled phenomenon is extremely important in the process of optimal design. Solutions that work for nominal parameter values may turn out to be unacceptable when random imperfections are taken into account. These imperfections may concern inevitable dispersion of material parameters, dimensions and external influences. The results of deterministic optimization, while maintaining previously defined coefficients of variation, may turn out to be completely useless. Striving for finding a solution that is not sensitive to imperfections in model parameters or external influences which are difficult to control, we have two options. The first one is robust optimization (Doltsinis *et al.*, 2005; Chen *et al.*, 2000; Li *et al.*, 2006; Hwang *et al.*, 2001; Sbaraglia *et al.*, 2018; Stocki, 2010). The second is optimization based on the so-called reliability based design optimization RBDO (Lopez and Beck, 2012; Aoues and Chateauneuf, 2010; Beck *et al.*, 2015; Kuschel and Rackwitz, 1997; Youn and Choi, 2004; Streicher and Rackwitz, 2002). If ensuring a high level of safety is the most important requirement for the designed structure, it is

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

worth choosing RBDO. In the RBDO framework, design constraints are formulated using failure probabilities. The applicability of RBDO is strongly conditioned by the availability of the joint probability density function of random variables describing the problem. The reliability of the estimated failure probability values depends on a precise stochastic model. A formulation of non-deterministic optimization that better adapts to design realities is robust optimization. The goal of robust optimization should be to simultaneously minimize the mean value and standard deviation of the objective function. Unlike RBDO optimization, this formulation does not require the estimation of failure probabilities. The random nature of the structure response is taken into account through the definition of the objective function and constraints, containing mean values and variances. The computational complexity of this approach is related to the use of effective methods for estimating statistical moments.

The aim of the analyzed work was to compare the numerical effectiveness of robust and RBDO optimization. The Costrel module of the Strurel computing environment (http://www.Strurel.de) was used for RBDO calculations. In the Costrel module, calculations are carried out in accordance with the idea of single-level methods. The aim of these methods is to eliminate the internal loop associated with reliability analysis by expanding the set of decision variables and replacing reliability constraints with optimality criteria for design point search tasks. Calculations related to "robust" optimization were performed using Numpress Explore software (http://www.numpress.ippt.pan.pl/). Appropriate approximation of the objective function and constraints is crucial for the effectiveness and convergence of the analyzes performed. The work uses the kriging method in its approximation version along with an experimental plan based on the concept of the optimal Latin hypercube and random sampling (Simpson *et al.*, 2001; Liefvendahl and Stocki, 2006; Zabojszcza and Radoń, 2022).

2. Deterministic optimization

In the currently dominant design practice, a building should not only be safe, but also optimal. The behaviour of a building under a given load is closely related to strength parameters of the materials used and stiffness of the structure. The designer decides whether the response of the structure is satisfactory, which depends on the assumptions and requirements introduced.

The need to take into account the variability or uncertainty of design parameters is suggested in most proposed design and construction standards (Standards and Eurocodes). Strict adherence to standard instructions is the simplest course of action, and such a treatment of the problem is called the deterministic approach.

A typical formulation of the deterministic optimization problem can be expressed as follows: find values of the variables $\mathbf{X}_d$, minimizing $f(\mathbf{X}_d)$ with constrains

$$\begin{aligned} &g_i(\mathbf{X}_d) \geqslant 0 \qquad i = 1, \ldots, k_g \qquad \text{– unequal constraints} \\ &h_i(\mathbf{X}_d) = 0 \qquad i = 1, \ldots, k_g \qquad \text{– equality constraints} \\ &X_{dj}^{\,l} \leqslant X_{dj} \leqslant X_{dj}^{\,u} \qquad j = 1, \ldots, n_d \qquad \text{– simple constraints} \end{aligned} \tag{2.1}$$

where: $\mathbf{X}_d$ – design variables, $f(\mathbf{X}_d)$ – objective function.

In the above formulation, both design variables, as well as all parameters defining the structure model, as well as objective and constraint functions, are deterministic, i.e., they are represented by one nominal value. The dominant methods of solving the task are linear or nonlinear programming methods. The optimal solution is most often searched for in an iterative manner. The most popular algorithms include: gradient algorithms, such as the conjugate gradient method, the sequential quadratic programming method, and the sequential linear programming method. An interesting comparison of various methods used in optimization was made in (Schittkowski *et al.*, 1994).

3. Reliability Based Design Optimization (RBDO)

The formulation of the RBDO consists in minimizing the objective function under probabilistic constraints. This formulation is written as: find $\mathbf{d}$, $\boldsymbol{\mu}_x$, minimize $f(\mathbf{d}, \boldsymbol{\mu}_X, \boldsymbol{\mu}_P)$ with constraints

$$\begin{aligned} &p[g_i(\mathbf{d}, \mathbf{X}, \mathbf{P}) \leqslant 0] - \Phi(-\beta_i^t) \leqslant 0 \qquad i = 1, \ldots, k_g \\ &d_j^l \leqslant d_j \leqslant d_j^u \qquad j = 1, \ldots, n_d \\ &\mu_{x_r}{}^l \leqslant \mu_{x_r} \leqslant \mu_{x_r}{}^u \qquad r = 1, \ldots, n_x \end{aligned} \tag{3.1}$$

where: $p_f^i = p[g_i(\mathbf{d}, \mathbf{X}, \mathbf{P}) \leqslant 0]$ – failure probability corresponding to the i-th limit function $g_i(\cdot)$, $\Phi(\cdot)$ – cumulative distribution function of the standard normal distribution, $\mathbf{X}$, $\mathbf{P}$ – vectors of random variables with expected values, respectively $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_P$, β_i^t, $i = 1, \ldots, k_g$ – minimum reliability indices established by the designer. Variables $\mathbf{d}$ describe deterministic values.

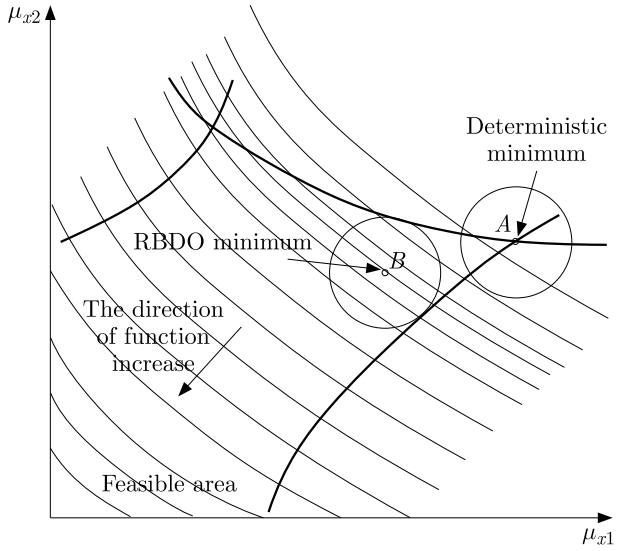


Fig. 1. Comparison of the optimal solution with deterministic optimization – point A and reliability based design optimization – point B

The idea of reliability optimization is presented in Fig. 1. In the case of a hypothetical optimization problem with two design variables and three constraints, the solution to this task in the deterministic version is point A. At the optimal point located on the border of the feasible area, two constraints are active. Let us assume that the design variables are not deterministic quantities but are characterized by a certain dispersion, and the coordinates of point A create a vector of expected values of the appropriate random variables. In such a case, most of the possible realizations of these variables will fall within some limited region around the deterministic optimum. For simplicity of presentation in Fig. 1, this area is marked as a circle centered at point A. It is easy to observe that a large part of the implementation of design variables is outside the feasible area. To ensure the required level of reliability, the circle surrounding point A should be moved inside the feasible area so that its new center B determines the solution guaranteeing higher reliability. This operation, of course, leads to an increase in the value of the objective function. How far solution B must be from the boundary of the permissible area is determined by the assumed safety margin.

In this formulation, an important element of the algorithm is calculation of the probability of meeting the design constraints. Numerical methods for determining the probability of failure and reliability indicators have been the subject of many publications in the field of various structure analyses. The articles (Mochocki and Radoń, 2019; Mochocki *et al.*, 2020) concern the reliability

analysis of lattice towers using a systems approach. The articles analyzed the impact of the wind load probability distribution and the type of connections in towers on their reliability. The paper (Kubicka and Radoń, 2018) is devoted to a unique design situation, which is undoubtedly the occurrence of fire in building structures. The authors examine the change in reliability indicators during fire using the example of lattice trusses. The article (Dudzik and Potrzeszcz-Sut, 2021) presents two approaches to the analysis of structural reliability. The primary research method was the First Order Reliability Method (FORM). The second analysis proposed a hybrid approach enabling the introduction of explicit forms of the limit state function into the reliability program. Neural networks and the proprietary MES module were used to create descriptions of this formula. The reliability of single-layer steel domes using FORM and Monte Carlo was the main subject of papers (Zabojszcza and Radoń, 2019, 2020; Radoń *et al.*, 2021).

In this paper, reliability-based design optimization is calculated using Costrel module of Strurel software (http://www.Strurel.de). The optimization task was solved by a constrained sequential quadratic programming procedure. The optimization scheme is a first order scheme. It requires twice-differentiability of the state function and uses only first-order approximations for failure probabilities (FORM). In Costrel, two search algorithms, Joint5 and NLPQL, respectively, are implemented. Both are multi-constraint optimizers with special driving routines.

4. Robust optimization

The possibility of using reliability optimization in design practice depends on the availability of the joint probability density function of the structure and load parameters. Unfortunately, due to the lack of appropriate statistical data, the use of this formulation becomes impossible. A formulation that better adapts to design realities is robust optimization. This formulation does not require an estimate of the failure probability. The random nature of the structure response is taken into account through the definitions of the objective function and constraints, which include mean values and response variances. The typical robust optimization formulation is written as: find d, $\boldsymbol{\mu}_x$, minimize $\{E[f(\mathbf{d},\mathbf{X},\mathbf{P})], \sigma[f(\mathbf{d},\mathbf{X},\mathbf{P})]\}$ with constraints

$$\begin{aligned}
&E[g_i(\mathbf{d},\mathbf{X},\mathbf{P})] - \widetilde{\beta}_i \sigma[g_i(\mathbf{d},\mathbf{X},\mathbf{P})] \geqslant 0 \qquad i = 1,\ldots,k_g \\
&\sigma[c_k(\mathbf{d},\mathbf{X},\mathbf{P})] \leqslant \sigma_k^u \qquad k = 1,\ldots,k_c \\
&d_j^l \leqslant d_j \leqslant d_j^u \qquad j = 1,\ldots,n_d \\
&{\mu_{x_r}}^l \leqslant \mu_{x_r} \leqslant {\mu_{x_r}}^u \qquad r = 1,\ldots,n_x
\end{aligned} \tag{4.1}$$

where: $\mathbf{d}$ – deterministic design variables, $\mathbf{X}$, $\mathbf{P}$ – vectors of random variables with expected values of μ_x, μ_p, f – objective function, g_i – functions of constraints, c_k – functions, the standard deviations of which must not exceed the allowable values σ_k^u, $\widetilde{\beta}_i > 0$ – coefficients corresponding to the constraints $g_i \geqslant 0$ which represent the safety margin with which these constraints must be met.

The robust optimization task is a multi-objective optimization task. In addition to the average value of the objective function, its dispersion is also minimized. The task can be modified to the following scalar optimization task: find values of variables $\mathbf{d}$, $\boldsymbol{\mu}_x$, that minimize

$$\widetilde{f} = \frac{1-\gamma}{\mu^*} E[f(\mathbf{d},\mathbf{X},\mathbf{P})] + \frac{\gamma}{\sigma^*} \sigma[f(\mathbf{d},\mathbf{X},\mathbf{P})] \tag{4.2}$$

with constraints (4.1).

The weighting factor $\gamma \in [0,1]$ defines the importance of each criterion. Values μ^* and σ^* are normalizing constants.

The computational complexity of the task requires the use of appropriate approximations of the unknown responses of the structure, as well as the use of effective methods for estimating statistical moments. In the paper, the kriging algorithm with optimal Latin hypercubes and random sampling was used. Additionally, in order to verify the correctness of the obtained results, calculations were performed using the second order method. The calculations were conducted using Numpress Explore software (http://www.numpress.ippt.pan.pl).

5. Numerical results and discussion

5.1. Geometry

In this paper, a steel single-storey frame with dimensions $h = 600\,\text{cm}$ and $L = 2h = 1200\,\text{cm}$ (Fig. 2) is analysed. The columns were originally modeled using square tubes with dimensions $D = 26\,\text{cm}$ and $d = 18\,\text{cm}$, Young's modulus $E = 210\,\text{GPa}$, Poisson's ratio $\nu = 0.3$, yield strength $f_y = 235\,\text{MPa}$. The stiffness of the beam is very high compared to the stiffness of the columns. In further calculations we assume $EI_b = \infty$. The structure is loaded with a horizontal force $P = 120\,\text{kN}$. The initial column mass is $f_{M1} = 1658\,\text{kg}$.

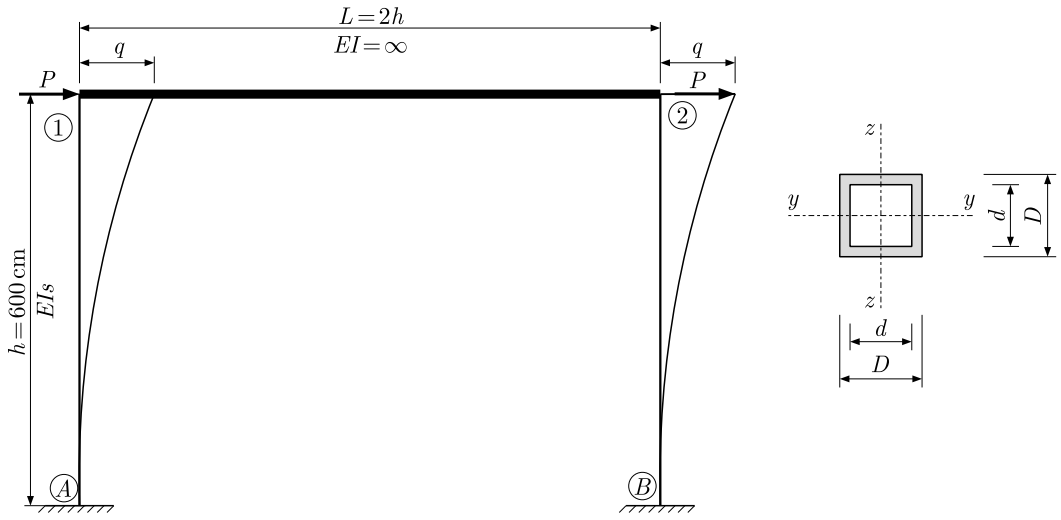


Fig. 2. Frame geometry and load

For the structure presented above, a series of analyzes is performed in subsequent stages. The first stage aims to verify the condition of the basic structure by performing a reliability analysis in Numpress Explore using the FORM method. The subsequent stages are related to the optimization of the structure. Deterministic optimization (in the Numpress Explore), reliability based design optimization (in the Costrel) and robust optimization (in the Numpress Explore) are performed successively.

5.2. Static analysis

When using the displacement method in terms of the first-order theory, the frame is once geometrically indeterminate. In the basic diagram of the displacement method, only one translational displacement is active, i.e. the horizontal displacement of the beam q. From the sum of projections on the X axis, we get

$$\Sigma X = P + P + T_{1A} + T_{2B} = 0 \qquad 2P = -T_{1A} - T_{2B} \tag{5.1}$$

After using the transformation formulas of the displacement method, the horizontal displacement of the frame is

$$\frac{24EJ}{h^3}q = 2P \quad \rightarrow \quad q = \frac{2Ph^3}{24EJ} = \frac{2Ph^3 \cdot 12}{24E(D^4 - d^4)} = \frac{Ph^3}{E(D^4 - d^4)} \tag{5.2}$$

In the example, in order to compare two optimization methods that take into account the random nature of design parameters, we only analyze the serviceability limit state. It expresses the difference between the permissible displacement and the displacement obtained as a result of calculations.

5.3. Reliability analysis

The reliability analysis of the structure was carried out using the FORM method. For the example under consideration, random variables were assumed as: D – the external dimension of the cross-section, d – the internal dimension of the cross-section, E – Young's modulus and P – force. Random variables are not correlated. The mean values of random variables and the coefficient of variation are listed in Table 1.

Table 1. Description of random variables

Random variables X_i	Mean values	Standard deviation	Coefficient of variation
D	26 cm	0.26 cm	1 %
d	18 cm	0.18 cm	1 %
E	21 000 kN/cm^2	630 kN/cm^2	3 %
P	120 kN	3.6 kN	3 %

The limit function is the limitation of the permissible horizontal displacement q_d of the node (SLS)

$$f_{SLS}(x) = q_d - q = 4 - \frac{Ph^3}{E(D^4 - d^4)} \tag{5.3}$$

where: q – horizontal displacement of the frame bolt, q_d – maximum horizontal displacement equal to $L/150 = 4$ cm.

The reliability index is $\beta^{SLS} = 1.909$ and probability of failure $p_f = 2.812\text{E-}02$.

5.4. Deterministic optimization

To emphasize the advisability of using the uncertainty of design parameters, in the first stage we performed deterministic optimization along with the assessment of structure reliability. In this optimization method, we look for optimal cross-section dimensions, using the classic deterministic optimization algorithm. The objective function is the mass of the single column

$$f_c = \min(D^2 - d^2)h\rho = \min(\text{mass}) \tag{5.4}$$

where: $\rho = 0.00785$ kg/cm^3 – steel density, h [cm] – column height.

Simple bounds are described in Table 2. They are the upper and lower limits of the searched design variables.

For this case 1% tolerance of the cross-sectional dimensions of the square tubes has been assumed. Inequality limits are formulated as conditions for not exceeding the permissible frame displacement

$$f_{SLS}(x) = q_d - q = 4 - \frac{Ph^3}{E(D^4 - d^4)} \tag{5.5}$$

Table 2. Simple constraints of the design variables

Design variable	Lower limit	Upper limit
D	25 cm	27 cm
d	17 cm	19 cm

Additionally, the feasible area is shown in Fig. 3. The vertical lines (green dotted) and horizontal lines (blue dashed) represent the simple constraints (for D and d) used in the considered example. The red line marks the limitation of the permissible horizontal node displacement of the frame. The permissible area is the result of individual restrictions and is marked in grey.

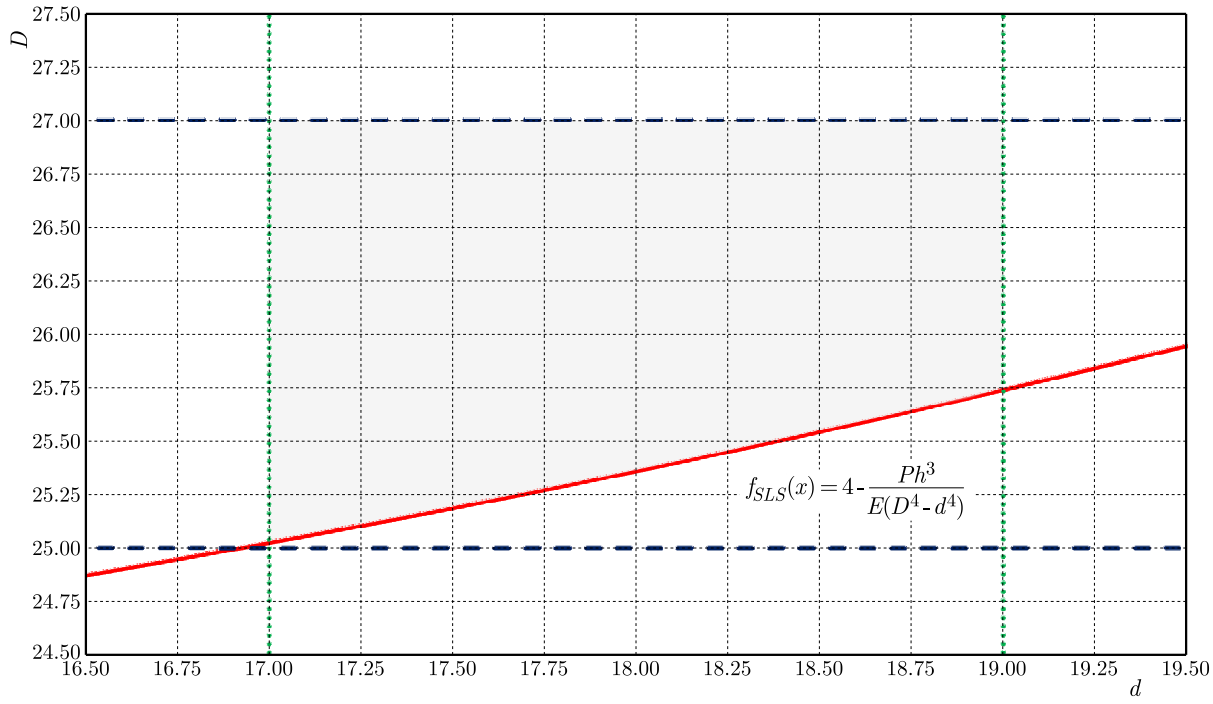


Fig. 3. Feasible area

The resulting cross-sectional dimensions are summarized in Table 3. The value of the objective function is 1420 kg.

Table 3. Values of the design variables obtained in deterministic optimization

Design variable	Optimal value
D	25.74 cm
d	19.00 cm

The probability of failure and the reliability index have also been verified, $p^{SLS} = 0.5$, $\beta^{SLS} = 0.0004$.

5.5. Reliability based design optimization

In the next approach, reliability based design optimization was used.

The task of RBDO takes the form: find μ_D, μ_d, minimizing $f_c = (D^2 - d^2)h\rho = \min(\text{mass})$ with constrains

$$p\Big(4 - \frac{Ph^3}{E(D^4 - d^4)}\Big) - \Phi(-1.8) \leqslant 0 \qquad 25 \leqslant \mu_D \leqslant 27 \qquad 17 \leqslant \mu_d \leqslant 19 \tag{5.6}$$

In the case of reliability optimization, it is necessary to assume a limit reliability index (failure probability). In the case under consideration, the limit was set at $\beta = 1.8$. After performing reliability optimization, the values of width and height of the cross-section were obtained as: $D = 26.34\,\text{cm}$ and $d = 19.00\,\text{cm}$. The probability of failure and the reliability index for RBDO approach were $\beta^{SGU} = 1.8$. $p_f^{SGU} = 3.6\text{E-}02$.

The weight of the optimized structure was $f_c = 1567\,\text{kg}$.

5.6. Robust optimization

The objective function is mass of the structure, but assuming that it takes into account the weighting factor γ, it determines the meaning of each criterion. Design variables are the expected values of the external and internal dimensions of the cross-section: μ_D, μ_d. The value of the coefficient of variation was set at 1%.

The robust optimization task takes the form: find: μ_D, μ_d, minimizing $f_C = (1-\gamma)E(\text{mass}) + \gamma\sigma(\text{mass})$ with constrains

$$E\Big(4 - \frac{Ph^3}{E(D^4 - d^4)}\Big) - \widetilde{\beta}_i\sigma\Big(4 - \frac{Ph^3}{E(D^4 - d^4)}\Big) \geqslant 0$$
$$25 \leqslant \mu_D \leqslant 27 \qquad 17 \leqslant \mu_d \leqslant 19 \tag{5.7}$$

where $\gamma \in [0, 1]$ – weighting factor determines the importance of each criterion.

Structural optimization was performed using the kriging response surface. Experiments are generated according to the plan of optimal Latin cubes and random sampling (Fig. 4). The parameters are $\gamma = 0.5$, $\widetilde{\beta^{SGU}} = 2.0$.

The values of the design variables are summarized in Table 4.

Table 4. Values of the random variables obtained in robust optimization

Design variable	OLH sampling	Random sampling
D [cm]	26.33	26.29
d [cm]	18.91	18.89

An increase in the cross-section height and an increase in the weight of the structure result in a significant change in the value of the reliability index and the probability of failure, which in this case are, respectively, for OLH and random sampling $\beta_R^{\text{OLH}} = 1.868$ and $\beta_R^{\text{RS}} = 1.775$, while the mass of structure is $f_{f_R}^{\text{OLH}} = 1581\,\text{kg}$ and $f_{f_R}^{\text{RS}} = 1577\,\text{kg}$.

5.7. Summary

Table 5 presents a summary of the results, including cross-sectional dimensions, structure weight, reliability index and failure probability. An additional aspect involves comparing the necessary number of iterations and the calculation time for each case.

The optimization results (Table 5) show that the best optimized design in terms of weight (a change of almost 240 kg compared to the initial value) does not meet the safety requirements. For deterministic optimization, the reliability index tended to 0. The results of Robust optimization, both in the case of generating experiments using the Optima Latin Hypercube (OLH) and Random Sampling (RS) plans, gave similar results. However, the use of the assumed better and much more effective method of generating experimental points (OLH) allowed the result to be provided much faster than in the case of RS. The Latin hypercube concept ensures an even distribution of points in the experiment plan (Fig. 4). This avoids clustering in certain areas and leaving other areas unexplored. The formation of such clusters has a particularly negative impact on the operation of the starting point selection procedure. Obtaining the final result obviously

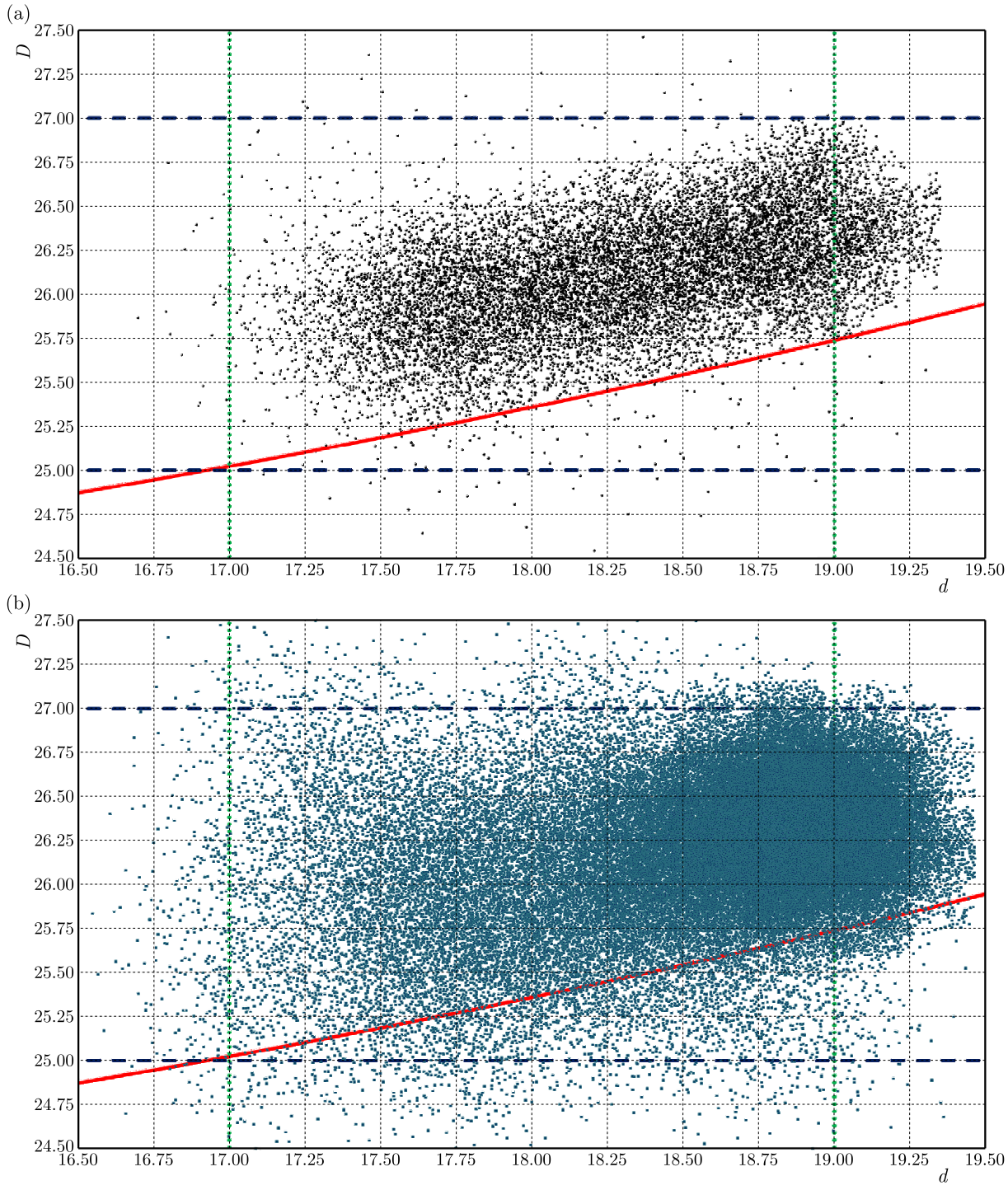


Fig. 4. An example of implementing a random sample using: (a) OLH, (b) random sampling

depends on the choice of the starting point for the iterative process. In the analyzed task, we see this in the example of comparing computation times using OLH and random sampling. For calculations using OLH it is 49 seconds, while for random sampling it is 8 hours 49 minutes and 49 seconds. The weight of the optimized structure is lower by approximately 80 kg (less than 5%) while still having a satisfactory reliability index. A similar Reliability Based Design Optimization analysis resulted in a slightly better optimized design, assuming a similar reliability index (at the level of 1.8). However, the obtained result is within the assumed limit of the permissible

area. Adopting such a solution may have a negative impact on possible additional unforeseen aspects of the analysis, i.e. inaccurate adoption of analysis parameters (standard deviation of the adopted variables, availability of the joint probability density function, etc.).

Table 5. Summary of results for individual analyses

Variable	Initial Values	Determi-nistic	Robust (Kriging)		RBDO
			OLH sampling	Random sampling	
D [cm]	26.00	25.74	26.38	26.29	26.34
d [cm]	18.00	19.00	18.97	18.89	19
Mass [kg]	1658	1420	1581	1577	1567
Reliability index	1.909	0.0004	1.868	1.775	$\beta = 1.8$
p_f	0.0281	0.5	0.0309	0.0380	0.036
No. Iterations	–	–	6	9	5
Time of calculation	–	1 s	49 s	8 h 49 min 49 s	0.2 s

Figure 5 shows the results of design optimization for the four analyses performed. The results of deterministic optimization and RBDO were on the border of the acceptable range. Only the Robust optimization results were within the acceptable range.

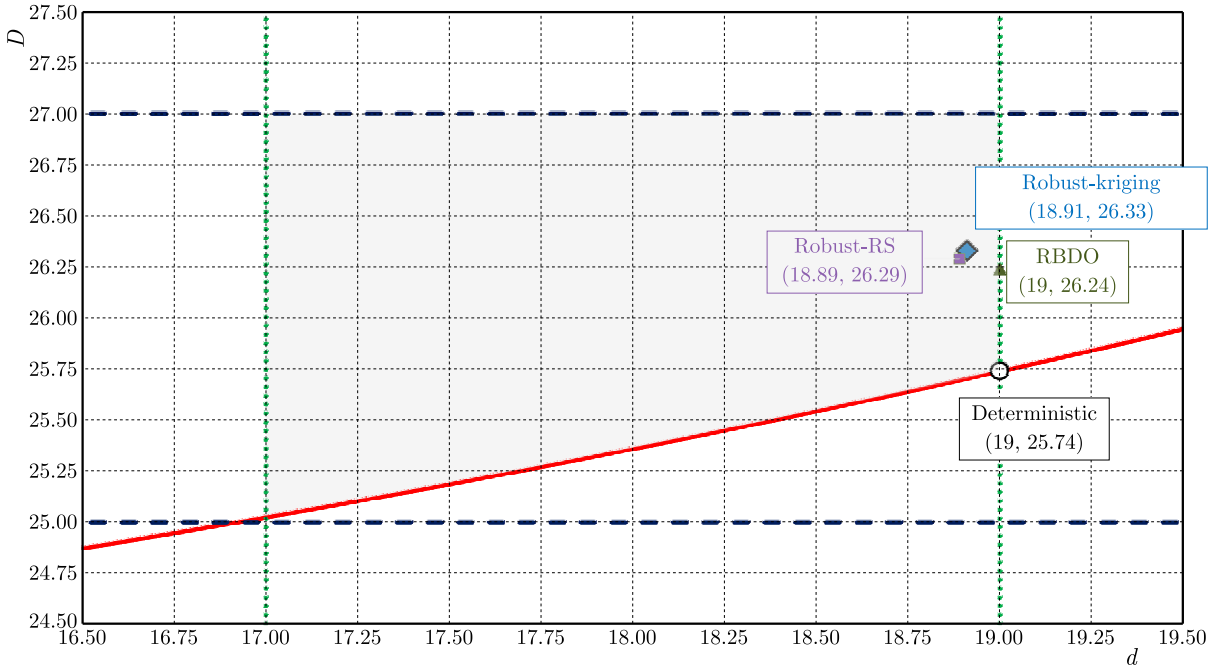


Fig. 5. Results for individual analyses with the feasible area

6. Conclusions

In the traditional deterministic approach, the random nature of the design variables and other parameters involved in the optimization formulation is accounted for by partial safety factors, which are typically calibrated to be applicable to the widest range of design tasks.

In order to find a solution that is insensitive to imperfections of model parameters or external influences that are difficult to control, we have two options. The first one is robust optimization. The second one is optimization based on the reliability of the so-called RBDO.

If guaranteeing a high level of safety is the most important requirement for the designed structure, it is worth choosing RBDO. Within RBDO, design constraints are formulated using failure probabilities. The applicability of RBDO is strongly dependent on the availability of the joint probability density function. The reliability of the estimated failure probability values depends on the precise stochastic model.

A formulation of non-deterministic optimization that better adapts to design realities is robust optimization. Unlike RBDO optimization, this formulation does not require estimation of failure probabilities. The random nature of the structure response is taken into account by defining the objective function and constraints, including mean values and variances. The computational complexity of this approach is related to the use of effective methods of estimating statistical moments.

References

1. Aoues Y., Chateauneuf A., 2010, Benchmark study of numerical methods for reliability-based design optimization, *Structural and Multidisciplinary Optimization*, **41**, 2, 277-294
2. Beck A.T., Gomes W.J.S., Lopez R.H., Miguel L.F.F., 2015, A comparison between robust and risk-based optimization under uncertainty, *Structural and Multidisciplinary Optimization*, **52**, 3, 479-492
3. Chen W., Fu W., Biggers S.B., Latour R.A., 2000, An affordable approach for robust design of thick laminated composite structure, *Optimization and Engineering*, **1**, 3, 305-322
4. Doltsinis I., Kang Z., Cheng G., 2005, Robust design of non-linear structures using optimization methods, *Computer Methods Applied Mechanics and Engineering*, **194**, 12-16, 1179–1795
5. Dudzik A., Potrzeszcz-Sut B., 2021, Hybrid approach to the first order reliability method in the reliability analysis of a spatial structure, *Applied Science*, **11**, 2, 648
6. Hwang K.-H., Lee K.-W., Park G.-J., 2001, Robust optimization of an automobile rearview mirror for vibration reduction, *Structural and Multidisciplinary Optimization*, **21**, 4, 300-308
7. Kubicka K., Radoń U., 2018, Influence of randomness of buckling coefficient on the reliability index's value under fire conditions, *Archives of Civil Engineering*, **64**, 3, 173-179
8. Kuschel N., Rackwitz R., 1997, Two basic problems in reliability-based structural optimization, *Mathematical Methods of Operational Research*, **46**, 3, 309-333
9. Li Y.Q., Cui Z.S., Ruan X.Y., Zhang D.J., 2006, CAE-based six sigma robust optimization for deep-drawing process of sheet metal, *The International Journal of Advanced Manufacturing Technology*, **30**, 631-637
10. Liefvendahl M., Stocki R., 2006, A study on algorithms for optimization of Latin hypercubes, *Journal of Statistical Planning and Inference*, **136**, 9, 3231-3247
11. Lopez R.H., Beck A.T., 2012, Reliability-based design optimization strategies based on FORM: a review, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, **34**, 4, 506-514
12. Mochocki W., Obara P., Radoń U., 2020, Impact of the wind load probability distribution and connection types on the reliability index of truss towers, *Journal of Theoretical and Applied Mechanics*, **58**, 2, 403-414
13. Mochocki W., Radoń U., 2019, Analysis of basic failure scenarios of a truss tower in a probabilistic approach, *Applied Sciences*, **9**, 13, 1-17
14. Numpress computer system, http://www.numpress.ippt.pan.pl/ [Accessed: 27.09.2023]
15. Radoń U., Szaniec W., Zabojszcza P., 2021, Probabilistic approach to limit states of a steel dome, *Materials*, **14**, 19, 5528

16. Sbaraglia F., Farokhi H., Aliabadi F.M.H., 2018, Robust and reliability-based design optimization of a composite floor beam, *Key Engineering Materials*, **774**, 486-491

17. Schittkowski K., Zillober C., Zotemantel R., 1994, Numerical comparison of nonlinear programming algorithms for structural optimization, *Structural Optimization*, **7**, 1-19

18. Simpson T.W., Mauery T.M., Korte J.J., Mistree F., 2001, Kriging models for global approximation in simulation-based multidisciplinary design optimization, *AIAA Journal*, **39**, 12, 2233-2241

19. Stocki R., 2010, Reliability analysis and resistance optimization of complex structures and technological processes (in Polish), *PRACE IPPT*, **2**

20. Streicher H., Rackwitz R., 2002, Structural optimization – a one level approach, [In:] *AMAS Workshop on Reliability-Based Design and Optimization – RBO'02*, Jendo S., Doliński K., Kleiber M., (Eds.)

21. Strurel computer system http://www.Strurel.de [Accessed: 27.09.2023]

22. Youn B. D., Choi K. K., 2004, A new response surface methodology for reliability-based design optimization, *Computers and Structures*, **82**, 2-3, 241-256

23. Zabojszcza P., Radoń U., 2019, The impact of node location imperfections on the reliability of single-layer steel domes, *Applied Sciences*, **9**, 2742

24. Zabojszcza P., Radoń U., 2020, Stability analysis of the single-layer dome in probabilistic description by the Monte Carlo method, *Journal of Theoretical and Applied Mechanics*, **58**, 2, 425-436

25. Zabojszcza P., Radoń U., 2022, Optimization of steel roof framing taking into account the random nature of design parameters, *Materials*, **15**, 14, 5017

Manuscript received October 26, 2023; accepted for print December 9, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 389-401, Warsaw 2024
https://doi.org/10.15632/jtam-pl/186178

NUMERICAL MODELLING OF THE LASER HIGH-TEMPERATURE HYPERTHERMIA USING THE DUAL-PHASE LAG EQUATION[1]

Mikołaj Stryczyński, Ewa Majchrzak

Silesian University of Technology, Department of Computational Mechanics and Engineering, Gliwice, Poland
corresponding author Mikołaj Stryczyński, e-mail: mikolaj.stryczynski@polsl.pl

In the paper, thermal processes occurring in a soft tissue subjected to laser irradiation are analyzed. The bioheat transfer in an axisymmetric domain is described by a dual-phase lag equation, which takes into account temperature-dependent thermophysical parameters of the tissue. The source term in this equation is related to laser irradiation, and is determined by solving the optical diffusion equation. It is assumed that the optical parameters depend on the Arrhenius integral, which is a measure of the degree of tissue destruction. In the model, the process of evaporation of water contained in the tissue is also considered.

Keywords: bioheat transfer, hyperthermia, optical diffusion equation, dual-phase lag equation, finite difference method

1. Introduction

Oncological hyperthermia involves raising the patient's body temperature in controlled conditions. It is used to support conventional therapies, such as chemotherapy or radiotherapy. The main goal of the procedure is to induce a state of increased patient temperature, activating the immune system to eliminate cancer cells from the body. This leads to an increase in the number of leukocytes and initiates a natural intervention against the tumor (Foster *et al.*, 2020).

Thermal ablation is a procedure aimed at destroying the tumor under the influence of high temperature. Heat is delivered directly to the tumor using needles or probes, without the need to surgically open the patient's body (Barnoon and Bakhshandehfard, 2021). The procedure is often performed in combination with laparoscopic and ultrasound methods, which allow for precise localization of the tumor (Giglio *et al.*, 2020) and is performed under anesthesia to relieve pain. During thermal ablation, probes applied to the tumor are heated to a temperature ranging from 65°C to 85°C for 10 to 15 minutes. Among various heating techniques, the laser-induced hyperthermia stands out for its precision and non-invasiveness.

Destruction of cancer tissue by laser irradiation is used, among others, in removal of oncological lesions within the liver using laparoscopic methods (Ellebrecht *et al.*, 2018). This process can be modelled using a dual-phase lag equation, which includes a source component that takes into account the interaction of laser with biological tissue. To determine this component, an appropriate mathematical model that describes light propagation in biological tissues must be selected (Ashley *et al.*, 1995; Dombrovsky and Baillis, 2010; Jacques and Pogue, 2008). One of such models is the radiative transport equation but, in some cases, it is possible to approximate the radiative transport equation with the optical diffusion equation, e.g. (Dombrovsky *et al.*, 2012; Jaunich *et al.*, 2008). Considering that in soft tissues, scattering dominates over absorption for wavelengths from 650 nm to 1300 nm, the optical diffusion equation was used in this study. Many articles in the literature are devoted to modelling laser interactions with biological

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

tissues. Most of them concern the use of the Pennes equation in combination with the radiative transport equation or the optical diffusion equation, e.g. (Kim and Guo, 2007; Kim *et al.*, 1996). Papers using the dual-phase lag equation appear relatively rarely (Majchrzak *et al.*, 2019; Zhou *et al.*, 2009). Moreover, constant thermophysical and optical parameters of biological tissues are commonly assumed.

In this paper, the dual-phase lag equation combined with the optical diffusion equation is used to model the interaction of laser with biological tissues. Additionally, the temperature-dependent tissue thermophysical parameters and optical tissue parameters changing with the Arrhenius integral are taken into account.

2. Mathematical model

An axisymmetric fragment of the liver subjected to laser irradiation is considered, as shown in Fig. 1. The dual–phase lag equation is based on the following relationship between the heat flux and temperature gradient (Tzou, 1995)

$$\mathbf{q}(r,z,t+\tau_q) = -\lambda(T)\,\mathrm{grad}\,T(r,z,t+\tau_T) \tag{2.1}$$

where τ_q represents the delay in the appearance of heat flux and its associated conduction through the medium, τ_T is the delay in the appearance of temperature gradient caused by heat conduction through structures of a small scale or size, λ is the thermal conductivity coefficient, T denotes temperature, $\mathbf{q}$ is the heat flux, r, z represent geometrical coordinates, and t is a time. This relationship is called generalized Fourier's law, because for time delays equal to zero ($\tau_T = \tau_q = 0$), one obtains the classical Fourier law.

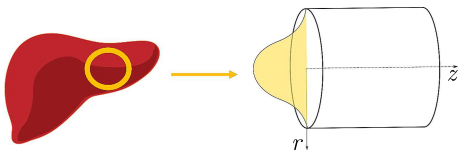


Fig. 1. An axisymmetric fragment of the liver

The functions $T(r,z,t+\tau_T)$ and $\mathbf{q}(r,z,t+\tau_q)$ are expanded into a Taylor series with an accuracy to the first derivatives

$$\mathbf{q}(r,z,t) + \tau_q \frac{\partial \mathbf{q}(r,z,t)}{\partial t} = \lambda(T)\,\mathrm{grad}\,T(r,z,t) + \tau_T \lambda(T) \frac{\partial[\,\mathrm{grad}\,T(r,z,t)]}{\partial t} \tag{2.2}$$

As known, the Fourier equation has the following form

$$c(T)\rho(T)\frac{\partial T(r,z,t)}{\partial t} = -\,\mathrm{div}\,\mathbf{q}(r,z,t) + Q(r,z,t) \tag{2.3}$$

where $c(T)$ is the specific heat of tissue, $\rho(T)$ is mass density and $Q(r,z,t)$ is the source function.

Basing on Eqs. (2.2) and (2.3), after appropriate transformations, the final form of the dual-phase lag is obtained (the arguments are omitted for simplicity) (Majchrzak and Stryczyński, 2022)

$$C(T)\frac{\partial T}{\partial t} + \tau_q \frac{\partial}{\partial t}\left[C(T)\frac{\partial T}{\partial t}\right] = \mathrm{div}\,[\lambda(T)\,\mathrm{grad}\,T] + \tau_T\,\mathrm{div}\left[\lambda(T)\frac{\partial(\,\mathrm{grad}\,T)}{\partial t}\right] + Q + \tau_q \frac{\partial Q}{\partial t} \tag{2.4}$$

where $C(T) = c(T)\rho(T)$ is the volumetric thermal capacity, and

$$Q = w(\psi)c_b(T_a - T) + Q_{met}(\psi) + Q_{ext} \tag{2.5}$$

while $w(\psi)$ is the blood perfusion rate, c_b is the specific heat of blood, T_a is the arterial temperature, $Q_{met}(\psi)$ is the metabolic heat source. Q_{ext} is the source function related to laser irradiation, and ψ is the so-called Arrhenius integral (Niemz, 2007)

$$\psi = \psi(r,z,t^f) = P\int_0^{t^f} \exp\Big(-\frac{E}{RT(r,z,t)}\Big)\,dt \tag{2.6}$$

where P is the pre-exponential factor, E is the activation energy, R is the universal gas constant, and $[0,t^f]$ is the time interval under consideration.

It should be emphasized that the calculation of the Arrhenius integral allows us to estimate the degree of destruction of biological tissue. Thus, a value of damage integral $\psi(r,z,t^f)=1$ corresponds to a 63% probability of cell death at a specific point (r,z), while $\psi(r,z,t^f)=4.6$ corresponds to 99% probability of cell death at this point. The value $\psi(r,z,t^f)=1$ is treated as extremely important because, from this moment, the tissue coagulation begins (Niemz, 2007).

Because

$$\frac{\partial}{\partial t}\Big[C(T)\frac{\partial T}{\partial t}\Big] = \frac{\partial C(T)}{\partial t}\frac{\partial T}{\partial t} + C(T)\frac{\partial^2 T}{\partial t^2} = \frac{dC(T)}{dT}\Big(\frac{\partial T}{\partial t}\Big)^2 + C(T)\frac{\partial^2 T}{\partial t^2} \tag{2.7}$$

thus Eq. (2.4) can be written in the form

$$\begin{aligned} &C(T)\Big(\frac{\partial T}{\partial t} + \tau_q\frac{\partial^2 T}{\partial t^2}\Big) + \tau_q\frac{dC(T)}{dT}\Big(\frac{\partial T}{\partial t}\Big)^2 \\ &\quad = \operatorname{div}\Big[\lambda(T)\operatorname{grad}T\Big] + \tau_T\operatorname{div}\Big[\lambda(T)\frac{\partial(\operatorname{grad}T)}{\partial t}\Big] + Q + \tau_q\frac{\partial Q}{\partial t} \end{aligned} \tag{2.8}$$

When the biological tissue reaches a temperature of approximately 100°C then in the mathematical model of the heating process, the phenomenon of water evaporation within the tissue should be taken into account. In this case, an additional source term related to the evaporation is introduced. This term is denoted as $Q_{evap}(T)$, and takes the following form (Yang *et al.*, 2007; Mochnacki and Majchrzak, 2007)

$$Q_{evap}(T) = L\frac{\partial W}{\partial t} = L\frac{dW}{dT}\frac{\partial T}{\partial t} \tag{2.9}$$

where L is the latent heat of water vaporization and W is the water volumetric fraction in the tissue domain. Thus

$$Q_{evap}(T) + \tau_q\frac{\partial Q_{evap}(T)}{\partial t} = L\frac{dW}{dT}\frac{\partial T}{\partial t} + \tau_q L\Big[\frac{d^2W}{dT^2}\Big(\frac{\partial T}{\partial t}\Big)^2 + \frac{dW}{dT}\frac{\partial^2 T}{\partial t^2}\Big] \tag{2.10}$$

Introducing dependence (2.10) into the dual-phase lag equation (2.8), one obtains

$$\begin{aligned} &\Big[C(T) - L\frac{dW}{dT}\Big]\frac{\partial T}{\partial t} + \tau_q\Big[C(T) - L\frac{dW}{dT}\Big]\frac{\partial^2 T}{\partial t^2} + \tau_q\Big[\frac{dC(T)}{dT} - L\frac{d^2W}{dT^2}\Big]\Big(\frac{\partial T}{\partial t}\Big)^2 \\ &\quad = \operatorname{div}\Big[\lambda(T)\operatorname{grad}T\Big] + \tau_T\operatorname{div}\Big[\lambda(T)\frac{\partial(\operatorname{grad}T)}{\partial t}\Big] + Q + \tau_q\frac{\partial Q}{\partial t} \end{aligned} \tag{2.11}$$

or

$$\begin{aligned} &\widehat{C}(T)\Big(\frac{\partial T}{\partial t} + \tau_q\frac{\partial^2 T}{\partial t^2}\Big) + \tau_q\frac{d\widehat{C}(T)}{dT}\Big(\frac{\partial T}{\partial t}\Big)^2 \\ &\quad = \operatorname{div}\Big[\lambda(T)\operatorname{grad}T\Big] + \tau_T\operatorname{div}\Big[\lambda(T)\frac{\partial(\operatorname{grad}T)}{\partial t}\Big] + Q + \tau_q\frac{\partial Q}{\partial t} \end{aligned} \tag{2.12}$$

where

$$\widehat{C} = C(T) - L\frac{dW}{dT} \tag{2.13}$$

is the effective volumetric specific heat (substitute thermal capacity).

Equation (2.12) is supplemented by boundary condition (Majchrzak and Stryczyński, 2022)

$$(r,z) \in \Gamma \cup \Gamma_0 : \qquad -\lambda(T)\Big(\mathbf{n}\cdot \operatorname{grad} T + \tau_T \mathbf{n}\cdot \operatorname{grad}\frac{\partial T}{\partial t}\Big) = 0 \tag{2.14}$$

where $\mathbf{n}$ is the normal outward vector, Γ_0 is the axis of the cylinder, and Γ is the outer surface of the cylinder.

The initial conditions are also known

$$t = 0 \qquad T = T_p \qquad \frac{\partial T}{\partial t} = \frac{Q(T_p)}{\widehat{C}(T_p)} \tag{2.15}$$

where T_p is the initial temperature of tissue.

As mentioned earlier, in soft tissues, the scattering dominates over absorption for wavelengths from 650 to 1300 nm, and then the source function Q_{ext} related to laser irradiation appearing in Eq. (2.12) (c.f. formula (2.5)) is of the form (Jasiński *et al.*, 2016)

$$Q_{ext}(r,z,t) = \mu_a \phi(r,z) p(t) \tag{2.16}$$

where μ_a is the absorption coefficient, $\phi(r,z)$ is the total light fluence rate and $p(t)$ is the function equal to 1 when laser is *on* and equal to 0 when laser is *off*.

The total light fluence $\phi(r,z)$ is the sum of collimated part $\phi(r,z)$ and diffuse part $\phi_d(r,z)$ (Abraham and Sparrow, 2007; Jasiński *et al.*, 2016)

$$\phi(r,z) = \phi_c(r,z) + \phi_d(r,z) \tag{2.17}$$

In the case of soft tissues, in order to determine the diffuse fluence rate, the steady-state optical diffusion equation (Dombrovsky *et al.*, 2012; Kim *et al.*, 2007) should be solved

$$(r,z) \in \Omega : \qquad \operatorname{div}[D \operatorname{grad}\phi_d(r,z)] - \mu_a \phi_d(r,z) + \mu_s' \phi_c(r,z) = 0 \tag{2.18}$$

where

$$D = \frac{1}{3[\mu_a + (1-g)\mu_s]} \tag{2.19}$$

and $\mu_s' = (1-g)\mu_s$ is the effective scattering coefficient (μ_s is the scattering coefficient, g is the anisotropy factor).

Equation (2.18) is supplemented by the boundary conditions

$$-D\mathbf{n}\cdot \operatorname{grad}\phi_d(r,z) = \begin{cases} \dfrac{\phi_d(r,z)}{2} & \text{for} \quad (r,z) \in \Gamma \\ 0 & \text{for} \quad (r,z) \in \Gamma_0 \end{cases} \tag{2.20}$$

The collimated fluence rate is given as (Zhou *et al.*, 2009)

$$\phi_c(r,z) = I_0 \exp(-\mu_t' z) \exp\Big(-\frac{r^2}{r_D^2}\Big) \tag{2.21}$$

where I_0 is the surface irradiance of laser, r_D is the radius of laser beam, and μ_t' is the attenuation coefficient defined as

$$\mu_t' = \mu_a + \mu_s' \tag{2.22}$$

It should be pointed out that the optical parameters depend on the degree of tissue damage described by the Arrhenius integral, and take the form (Fasano *et al.*, 2010)

$$\begin{aligned}
&\mu_a = \mu_a(\psi) = \exp(-\psi)\mu_{a,n} + [1-\exp(-\psi)]\mu_{a,c} \\
&\mu_s = \mu_s(\psi) = \exp(-\psi)\mu_{s,n} + [1-\exp(-\psi)]\mu_{s,c} \\
&g = g(\psi) = \exp(-\psi)g_n + [1-\exp(-\psi)]g_c
\end{aligned} \tag{2.23}$$

where the indexes n and c represent the tissue in its natural and coagulated state.

Summing up, at first, equation (2.18) supplemented by boundary conditions (2.20) and next dual-phase lag equation (2.12) with boundary condition (2.14) and initial conditions (2.15), should be solved.

3. Method of solution

First, optical diffusion equation (2.18) is considered. For a cylindrical co-ordinate system, we have

$$\operatorname{div}(D\operatorname{grad}) = \frac{1}{r}\frac{\partial}{\partial r}\Big(rD\frac{\partial\phi_d}{\partial r}\Big) + \frac{\partial}{\partial z}\Big(D\frac{\partial\phi_d}{\partial z}\Big) \tag{3.1}$$

After determining the appropriate derivatives, one obtains

$$\operatorname{div}(D\operatorname{grad}) = \frac{1}{r}D\frac{\partial\phi_d}{\partial r} + D\Big(\frac{\partial^2\phi_d}{\partial r^2} + \frac{\partial^2\phi_d}{\partial z^2}\Big) + \frac{\partial D}{\partial r}\frac{\partial\phi_d}{\partial r} + \frac{\partial D}{\partial z}\frac{\partial\phi_d}{\partial z} \tag{3.2}$$

Since the diffusion coefficient D (Eq. (2.19)) depends on the parameters μ_a, μ_s and g, the parameters μ_a, μ_s, g depend on the Arrhenius integral (Eq. (2.23)), and the Arrhenius integral depends on temperature (Eq. (2.6)), the derivatives $\partial D/\partial r$ and $\partial D/\partial z$ are calculated using the chain rule

$$\begin{aligned}
\frac{\partial D}{\partial r} &= \frac{\partial D}{\partial\mu_a}\frac{\partial\mu_a}{\partial\psi}\frac{\partial\psi}{\partial T}\frac{\partial T}{\partial r} + \frac{\partial D}{\partial\mu_s}\frac{\partial\mu_s}{\partial\psi}\frac{\partial\psi}{\partial T}\frac{\partial T}{\partial r} + \frac{\partial D}{\partial g}\frac{\partial g}{\partial\psi}\frac{\partial\psi}{\partial T}\frac{\partial T}{\partial r} \\
&= \Big(\frac{\partial D}{\partial\mu_a}\frac{\partial\mu_a}{\partial\psi} + \frac{\partial D}{\partial\mu_s}\frac{\partial\mu_s}{\partial\psi} + \frac{\partial D}{\partial g}\frac{\partial g}{\partial\psi}\Big)\frac{\partial\psi}{\partial T}\frac{\partial T}{\partial r} = P\frac{\partial T}{\partial r}
\end{aligned} \tag{3.3}$$

and

$$\frac{\partial D}{\partial z} = \Big(\frac{\partial D}{\partial\mu_a}\frac{\partial\mu_a}{\partial\psi} + \frac{\partial D}{\partial\mu_s}\frac{\partial\mu_s}{\partial\psi} + \frac{\partial D}{\partial g}\frac{\partial g}{\partial\psi}\Big)\frac{\partial\psi}{\partial T}\frac{\partial T}{\partial z} = P\frac{\partial T}{\partial z} \tag{3.4}$$

where

$$P = \Big(\frac{\partial D}{\partial\mu_a}\frac{\partial\mu_a}{\partial\psi} + \frac{\partial D}{\partial\mu_s}\frac{\partial\mu_s}{\partial\psi} + \frac{\partial D}{\partial g}\frac{\partial g}{\partial\psi}\Big)\frac{\partial\psi}{\partial T} \tag{3.5}$$

Finally, Eq. (2.18) can be written in the form

$$\frac{1}{r}D\frac{\partial\phi_d}{\partial r} + D\Big(\frac{\partial^2\phi_d}{\partial r^2} + \frac{\partial^2\phi_d}{\partial z^2}\Big) + P\Big(\frac{\partial T}{\partial r}\frac{\partial\phi_d}{\partial r} + \frac{\partial T}{\partial z}\frac{\partial\phi_d}{\partial z}\Big) - \mu_a\phi_d + \mu_s'\phi_c = 0 \tag{3.6}$$

Optical diffusion equation (3.6) is solved using the finite difference method (FDM). The differential grid is shown in Fig. 2. For the internal nodes (i,j), where $i = 1,2,\ldots,m-1$ and $j = 1,2,\ldots,n-1$, the following FDM approximation of this equation is proposed

$$\begin{aligned}
&\frac{1}{r_{i,j}}D^f_{i,j}\frac{\phi^f_{d,i+1,j} - \phi^f_{d,i-1,j}}{2h} + D^f_{i,j}\frac{\phi^f_{d,i-1,j} - 2\phi^f_{d,i,j} + \phi^f_{d,i+1,j}}{h^2} \\
&\quad + D^f_{i,j}\frac{\phi^f_{d,i,j-1} - 2\phi^f_{d,i,j} + \phi^f_{d,i,j+1}}{h^2} + P_{i,j}\frac{T^f_{i+1,j} - T^f_{i-1,j}}{2h}\frac{\phi^f_{d,i+1,j} - \phi^f_{d,i-1,j}}{2h} \\
&\quad + P_{i,j}\frac{T^f_{i,j+1} - T^f_{i,j-1}}{2h}\frac{\phi^f_{d,i,j+1} - \phi^f_{d,i,j-1}}{2h} - \mu^f_{a,i,j}\phi^f_{d,i,j} + \mu'^f_{s,i,j}\phi_{c,i,j} = 0
\end{aligned} \tag{3.7}$$

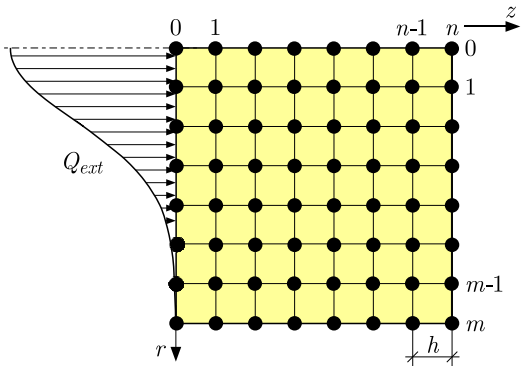


Fig. 2. Differential mesh

or

$$\begin{aligned}\phi^f_{d,i,j} = {} & \frac{1}{r_{i,j}} D^f_{i,j} \frac{\phi^f_{d,i+1,j} - \phi^f_{d,i-1,j}}{2hG^f_{i,j}} + D^f_{i,j} \frac{\phi^f_{d,i-1,j} + \phi^f_{d,i+1,j} + \phi^f_{d,i,j-1} + \phi^f_{d,i,j+1}}{h^2 G^f_{i,j}} \\ & + P^f_{i,j} \frac{T^f_{i+1,j} - T^f_{i-1,j}}{2hG^f_{i,j}} \frac{\phi^f_{d,i+1,j} - \phi^f_{d,i-1,j}}{2h} \\ & + P^f_{i,j} \frac{T^f_{i,j+1} - T^f_{i,j-1}}{2hG^f_{i,j}} \frac{\phi^f_{d,i+1,j} - \phi^f_{d,i-1,j}}{2h} + \frac{\mu'^f_{s,i,j}}{G^f_{i,j}} \phi_{c,i,j}\end{aligned} \tag{3.8}$$

where

$$G^f_{i,j} = 4\frac{D^f_{i,j}}{h^2} + \mu^f_{a,i,j} \tag{3.9}$$

Boundary conditions (2.20) are approximated in a similar way, and then:
— for $j = 0$, $i = 1, 2, \ldots, m-1$

$$\phi^f_{d,i,j} = \frac{2D^f_{i,j}}{2D^f_{i,j} + h} \phi^f_{d,i,j+1} \tag{3.10}$$

— for $j = n$, $i = 1, 2, \ldots, m-1$

$$\phi^f_{d,i,j} = \frac{2D^f_{i,j}}{2D^f_{i,j} + h} \phi^f_{d,i,j-1} \tag{3.11}$$

— for $i = 0$, $j = 1, 2, \ldots, n-1$

$$\phi^f_{d,i,j} = \phi^f_{d,i+1,j} \tag{3.12}$$

— for $i = m$, $j = 1, 2, \ldots, n-1$

$$\phi^f_{d,i,j} = \frac{2D^f_{i,j}}{2D^f_{i,j} + h} \phi^f_{d,i-1,j} \tag{3.13}$$

It should be pointed out that the f index means that the optical diffusion equation due to variable optical parameters must be solved at each time step. To summarize the algorithm, at each time step, the values of optical parameters dependent on the Arrhenius integral (Eqs. (2.23)) are calculated. The collimated part of the light fluence is determined (Eq. (2.21)) and the diffuse

part is determined by solving the system of Eqs. (3.8)-(3.13). This system was solved using the iterative Gaussian method. Then, the source component Q_{ext} is calculated from formula (2.16).

Now, the method of solving the dual-phase lag equation will be presented. Equation (2.12) can be written in the form (c.f. formula (2.5))

$$\begin{aligned}\widehat{C}(T)\Big(\frac{\partial T}{\partial t}+\tau_q\frac{\partial^2 T}{\partial t^2}\Big)+\tau_q\frac{d\widehat{C}(T)}{dT}\Big(\frac{\partial T}{\partial t}\Big)^2 &= \operatorname{div}\Big[\lambda(T)\operatorname{grad}T\Big]\\ &+\tau_T\operatorname{div}\Big[\lambda(T)\frac{\partial(\operatorname{grad}T)}{\partial t}\Big]+w(\psi)c_b(T_a-T)+Q_{met}(\psi)+Q_{ext}\\ &+\tau_q\Big(\frac{dw(\psi)}{d\psi}\frac{\partial\psi}{\partial t}c_b(T_a-T)-w(\psi)c_b\frac{\partial T}{\partial t}+\frac{dQ_{met}(\psi)}{d\psi}\frac{\partial\psi}{\partial t}+\frac{\partial Q_{ext}}{\partial t}\Big)\end{aligned} \tag{3.14}$$

or

$$\begin{aligned}&[\widehat{C}(T)+\tau_q w(\psi)c_b]\frac{\partial T}{\partial t}+\tau_q\widehat{C}(T)\frac{\partial^2 T}{\partial t^2}+\tau_q\frac{d\widehat{C}(T)}{dT}\Big(\frac{\partial T}{\partial t}\Big)^2\\ &\quad=\operatorname{div}[\lambda(T)\operatorname{grad}T]+\tau_T\operatorname{div}\Big[\lambda(T)\frac{\partial(\operatorname{grad}T)}{\partial t}\Big]+w(\psi)c_b(T_a-T)\\ &\quad+Q_{met}(\psi)+Q_{ext}+\tau_q\Big(\frac{dw(\psi)}{d\psi}\frac{\partial\psi}{\partial t}c_b(T_a-T)+\frac{dQ_{met}(\psi)}{d\psi}\frac{\partial\psi}{\partial t}+\frac{\partial Q_{ext}}{\partial t}\Big)\end{aligned} \tag{3.15}$$

where (Majchrzak and Stryczyński, 2022)

$$\begin{aligned}&\operatorname{div}[\lambda(T)\operatorname{grad}T]+\tau_T\operatorname{div}\Big[\lambda(T)\operatorname{grad}\frac{\partial T}{\partial t}\Big]=\frac{1}{r}\lambda(T)\Big[\frac{\partial T}{\partial r}+\tau_T\frac{\partial}{\partial r}\Big(\frac{\partial T}{\partial t}\Big)\Big]\\ &+\frac{d\lambda(T)}{dT}\frac{\partial T}{\partial r}\Big[\frac{\partial T}{\partial r}+\tau_T\frac{\partial}{\partial r}\Big(\frac{\partial T}{\partial t}\Big)\Big]+\lambda(T)\Big[\frac{\partial^2 T}{\partial r^2}+\tau_T\frac{\partial^2}{\partial r^2}\Big(\frac{\partial T}{\partial t}\Big)\Big]\\ &+\frac{d\lambda(T)}{dT}\frac{\partial T}{\partial z}\Big[\frac{\partial T}{\partial z}+\tau_T\frac{\partial}{\partial z}\Big(\frac{\partial T}{\partial t}\Big)\Big]+\lambda(T)\Big[\frac{\partial^2 T}{\partial z^2}+\tau_T\frac{\partial^2}{\partial z^2}\Big(\frac{\partial T}{\partial t}\Big)\Big]\end{aligned} \tag{3.16}$$

To solve Eq. (3.15), the implicit scheme of the finite difference method is used. For internal node (i,j), $i=1,2,\ldots,m-1$, $j=1,2,\ldots,n-1$ and transition $t^f\to t^{f+1}$, the following approximation of operator (3.16) is obtained (Majchrzak and Stryczyński, 2022)

$$\begin{aligned}&\operatorname{div}[\lambda(T)\operatorname{grad}T]_{i,j}^{f+1}+\tau_T\operatorname{div}\Big[\lambda(T)\operatorname{grad}\frac{\partial T}{\partial t}\Big]_{i,j}^{f+1}=A_{i,j}^f\Big(1-\frac{h}{2r_{i,j}}\Big)T_{i-1,j}^{f+1}\\ &+A_{i,j}^f\Big(1+\frac{h}{2r_{i,j}}\Big)T_{i+1,j}^{f+1}+A_{i,j}^f(T_{i,j-1}^{f+1}+T_{i,j+1}^{f+1})-4A_{i,j}^f T_{i,j}^{f+1}+B_{i,j}^f\end{aligned} \tag{3.17}$$

where

$$A_{i,j}^f=\frac{\lambda_{i,j}^f(\Delta t+\tau_T)}{h^2\Delta t} \tag{3.18}$$

and

$$\begin{aligned}B_{i,j}^f=&-\frac{\lambda_{i,j}^f\tau_T}{h^2\Delta t}(T_{i-1,j}^f+T_{i+1,j}^f+T_{i,j-1}^f+T_{i,j+1}^f-4T_{i,j}^f)-\frac{\lambda_{i,j}^f\tau_T}{2hr_{i,j}\Delta t}(T_{i+1,j}^f-T_{i-1,j}^f)\\ &+\frac{1}{4h^2\Delta t}\Big(\frac{d\lambda(T)}{dT}\Big)_{i,j}^f[(\Delta t+\tau_T)(T_{i+1,j}^f-T_{i-1,j}^f)^2-\tau_T(T_{i+1,j}^f-T_{i-1,j}^f)(T_{i+1,j}^{f-1}-T_{i-1,j}^{f-1})]\\ &+\frac{1}{4h^2\Delta t}\Big(\frac{d\lambda(T)}{dT}\Big)_{i,j}^f[(\Delta t+\tau_T)(T_{i,j+1}^f-T_{i,j-1}^f)^2-\tau_T(T_{i,j+1}^f-T_{i,j-1}^f)(T_{i,j+1}^{f-1}-T_{i,j-1}^{f-1})]\end{aligned} \tag{3.19}$$

while Δt is the time step.

The approximation of the left-hand side of Eq. (3.15) is the following

$$\begin{aligned}&\Big\{[\widehat{C}(T)+\tau_q w(\psi)c_b]\frac{\partial T}{\partial t}+\tau_q\widehat{C}(T)\frac{\partial^2 T}{\partial t^2}+\tau_q\frac{d\widehat{C}(T)}{dT}\Big(\frac{\partial T}{\partial t}\Big)^2\Big\}_{i,j}^{f+1}\\&\quad=(\widehat{C}_{i,j}^f+w_{i,j}^f c_b)\frac{T_{i,j}^{f+1}-T_{i,j}^f}{\Delta t}+\tau_q\widehat{C}_{i,j}^f\frac{T_{i,j}^{f+1}-2T_{i,j}^f+T_{i,j}^{f-1}}{(\Delta t)^2}\\&\quad+\tau_q\Big(\frac{d\widehat{C}(T)}{dT}\Big)_{i,j}^f\Big(\frac{T_{i,j}^f-T_{i,j}^{f-1}}{\Delta t}\Big)^2\end{aligned}\tag{3.20}$$

Thus, one obtains finally the approximation of Eq. (3.15) in the form

$$\begin{aligned}&(\widehat{C}_{i,j}^f+\tau_q w_{i,j}^f c_b)\frac{T_{i,j}^{f+1}-T_{i,j}^f}{\Delta t}+\tau_q\widehat{C}_{i,j}^f\frac{T_{i,j}^{f+1}-2T_{i,j}^f+T_{i,j}^{f-1}}{(\Delta t)^2}\\&\quad+\tau_q\Big(\frac{d\widehat{C}(T)}{dT}\Big)_{i,j}^f\Big(\frac{T_{i,j}^f-T_{i,j}^{f-1}}{\Delta t}\Big)^2=A_{i,j}^f\Big(1-\frac{h}{2r_{i,j}}\Big)T_{i-1,j}^{f+1}+A_{i,j}^f\Big(1+\frac{h}{2r_{i,j}}\Big)T_{i+1,j}^{f+1}\\&\quad+A_{i,j}^f(T_{i,j-1}^{f+1}+T_{i,j+1}^{f+1})-4A_{i,j}^f T_{i,j}^{f+1}-c_b\Big[w_{i,j}^f+\tau_q\Big(\frac{dw(\psi)}{d\psi}\frac{\partial\psi}{\partial t}\Big)_{i,j}^f\Big]T_{i,j}^{f+1}+D_{i,j}^f\end{aligned}\tag{3.21}$$

where

$$\begin{aligned}D_{i,j}^f&=B_{i,j}^f+w_{i,j}^f c_b T_a+(Q_{met})_{i,j}^f+(Q_{ext})_{i,j}^f\\&\quad+\tau_q\Big[\Big(\frac{dw(\psi)}{d\psi}\frac{\partial\psi}{\partial t}\Big)_{i,j}^f c_b T_a+\Big(\frac{dQ_{met}(\psi)}{d\psi}\frac{\partial\psi}{\partial t}\Big)_{i,j}^f+\Big(\frac{\partial Q_{ext}}{\partial t}\Big)_{i,j}^f\Big]\end{aligned}\tag{3.22}$$

From Eq. (3.21), it results

$$T_{i,j}^{f+1}=\frac{A_{i,j}^f}{F_{i,j}^f}\Big(1-\frac{h}{2r_{i,j}}\Big)T_{i-1,j}^{f+1}+\frac{A_{i,j}^f}{F_{i,j}^f}\Big(1+\frac{h}{2r_{i,j}}\Big)T_{i+1,j}^{f+1}+\frac{A_{i,j}^f}{F_{i,j}^f}(T_{i,j-1}^{f+1}+T_{i,j+1}^{f+1})+\frac{E_{i,j}^f}{F_{i,j}^f}\tag{3.23}$$

where

$$\begin{aligned}E_{i,j}^f&=D_{i,j}^f+\frac{(\widehat{C}_{i,j}^f+\tau_q w_{i,j}^f c_b)\Delta t+2\tau_q\widehat{C}_{i,j}^f}{(\Delta t)^2}T_{i,j}^f-\frac{\widehat{C}_{i,j}^f\tau_q}{(\Delta t)^2}T_{i,j}^{f-1}-\tau_q\Big(\frac{d\widehat{C}(T)}{dT}\Big)_{i,j}^f\Big(\frac{T_{i,j}^f-T_{i,j}^{f-1}}{\Delta t}\Big)^2\\F_{i,j}^f&=\frac{(\widehat{C}_{i,j}^f+\tau_q w_{i,j}^f c_b)\Delta t+\tau_q\widehat{C}_{i,j}^f}{(\Delta t)^2}+4A_{i,j}^f+c_b\Big[w_{i,j}^f+\tau_q\Big(\frac{dw(\psi)}{d\psi}\frac{\partial\psi}{\partial t}\Big)_{i,j}^f\Big]\end{aligned}\tag{3.24}$$

Boundary condition (2.14) should also be approximated (Majchrzak and Stryczyński, 2022).

At each time step, the system of equations (3.23) is solved using the Gauss-Seidl iterative method.

4. Results of computations

An axisymmetric fragment of the biological tissue (liver) of dimensions $R = 0.02\,\mathrm{m}$ and $Z = 0.02\,\mathrm{m}$ is considered (Fig. 2).

To solve modified dual-phase lag Eq. (2.12), the temperature dependence of water content is needed. Based upon experiments that the measured water content as a function of temperature, the following dependence is used (Yang *et al.*, 2007)

$$W(T)=0.778\begin{cases}1-\exp\Big(\dfrac{T-106}{3.42}\Big) & \text{for} \quad T\leqslant 103^\circ\mathrm{C}\\ S(T) & \text{for} \quad 103^\circ\mathrm{C}\leqslant T\leqslant 104^\circ\mathrm{C}\\ \exp\Big(\dfrac{-(T-80)}{34.37}\Big) & \text{for} \quad T\geqslant 104^\circ\mathrm{C}\end{cases}\tag{4.1}$$

where $S(T)$ is a cubic C^1 spline between two exponential functions, and has the following form

$$S(T) = -4.05821416 \cdot 10^4 + 1.18204602 \cdot 10^3 T - 11.4752357 T^2 + 3.71298243 \cdot 10^{-2} T^3 \quad (4.2)$$

In Fig. 3, the course of function $W(T)$ is shown. As can be seen, the water content of the soft tissue is approximately 77.8% by volume and remains almost constant until the phase transition temperature is reached. As the temperature increases further, the water content gradually decreases down to approximately 20% of the volume at 130°C.

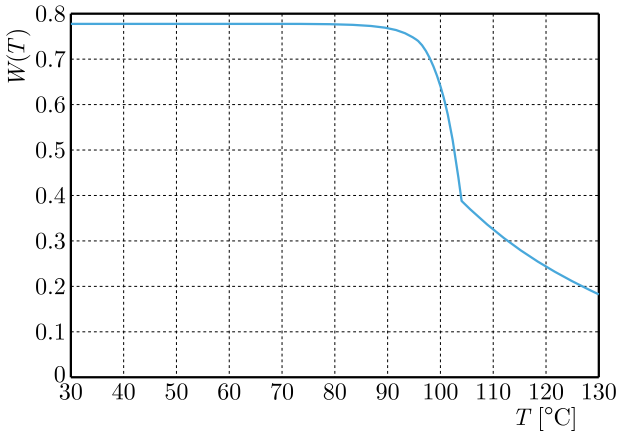


Fig. 3. The water content in soft tissue as a function of temperature

The temperature-dependent thermal conductivity and volumetric specific heat of the liver tissue are taken from Lopresto *et al.* (2019)

$$\lambda(T) = \begin{cases} 0.5075 + 5.6261 \cdot 10^{-51} T^{25.296} & \text{for} \quad T \leqslant 99^\circ\text{C} \\ S_\lambda(T) & \text{for} \quad 99^\circ\text{C} \leqslant T \leqslant 101^\circ\text{C} \\ 55.44 - 0.99701 T + 4.4988 \cdot 10^{-3} T^2 & \text{for} \quad T \geqslant 101^\circ\text{C} \end{cases}$$

$$C(T) = \begin{cases} 3.3012 + \dfrac{3.6186}{100 - T} & \text{for} \quad T \leqslant 99^\circ\text{C} \\ S_C(T) & \text{for} \quad 99^\circ\text{C} \leqslant T \leqslant 101^\circ\text{C} \\ 90.808 - 1.5491 T + 6.6664 \cdot 10^{-3} T^2 & \text{for} \quad T \geqslant 101^\circ\text{C} \end{cases} \quad (4.3)$$

where $S_\lambda(T)$ and $S_C(T)$ are the cubic C^1 splines.

Tissue destruction significantly affects the blood flow process. Abraham and Sparrow (2007) presented the following relationship between the blood perfusion rate and the degree of tissue damage

$$w(\psi) = \begin{cases} (1 + 25\psi - 260\psi^2) w_{b0} & \text{for} \quad 0 \leqslant \psi \leqslant 0.1 \\ (1 - \psi) w_{b0} & \text{for} \quad 0.1 \leqslant \psi \leqslant 1 \\ 0 & \text{for} \quad \psi \geqslant 1 \end{cases} \quad (4.4)$$

where $w_{b0} = 0.5\,\text{kg}/(\text{m}^3\text{s})$ is the baseline value of the blood perfusion rate.

A similar relationship is assumed for the metabolic term (Abraham and Sparrow, 2007)

$$Q_{met}(\psi) = \begin{cases} (1 + 25\psi - 260\psi^2) Q_{m0} & \text{for} \quad 0 \leqslant \psi \leqslant 0.1 \\ (1 - \psi) Q_{m0} & \text{for} \quad 0.1 \leqslant \psi \leqslant 1 \\ 0 & \text{for} \quad \psi \geqslant 1 \end{cases} \quad (4.5)$$

where $Q_{m0} = 245\,\text{W}/\text{m}^3$ is the baseline value of the metabolic heat source.

The remaining data used at the stage of numerical computations are collected below: initial temperature $T_p = 37°\text{C}$, relaxation time $\tau_q = 4\,\text{s}$, thermalization time $\tau_T = 2\,\text{s}$, optical parameters: $\mu_{a,n} = 195\,1/\text{m}$, $\mu_{a,c} = 13\,1/\text{m}$, $\mu_{s,n} = 4350\,1/\text{m}$, $\mu_{s,c} = 30590\,1/\text{m}$, $g_n = 0.0931$ and $g_c = 0.09165$ (Fasano *et al.*, 2010), specific heat of blood $c_b = 3770\,\text{J/(kgK)}$, arterial temperature $T_a = 37°\text{C}$, radius of laser beam $r_D = 0.001\,\text{m}$. In the Arrhenius integral (2.6): $P = 7.39{\cdot}10^{37}\,1/\text{s}$, $E = 2.58 \cdot 10^5\,\text{J/mol}$, $R = 8.314\,\text{J/(mol K)}$ (Szasz *et al.*, 2011).

The computations were performed for the difference mesh $m = n = 100$ nodes using the time step $\Delta t = 0.0005\,\text{s}$ until reaching the observation time, which was set to 150 seconds.

Before proceeding to the detailed analysis, calculations were performed for both constant and Arrhenius integral-dependent optical parameters in order to investigate the difference in temperature profiles. It was assumed that the laser operated for 120 s with power $I_0 = 1.33 \cdot 10^5\,\text{W/m}^2$.

A comparison of the results for constant and variable optical parameters is shown in Fig. 4 in the form of temperature profiles at the point with coordinates $(0.0002\,\text{m}, 0.0002\,\text{m})$. As can be seen, in the case of variable optical parameters, a much lower temperature was achieved. This is due to significant changes in the values of tissue scattering and absorption coefficients in the coagulated and natural state. The temperature difference is significant. It can, therefore, be concluded that in the case of high-temperature hyperthermia, it is important to take into account the optical parameters of the tissue that change with the Arrhenius integral. Otherwise, overestimated temperatures may be obtained, which may lead to incorrect predictions of tissue damage.

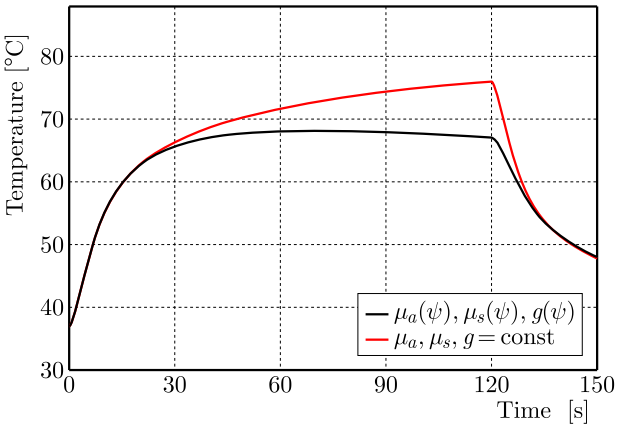


Fig. 4. Temperature history for constant and variable optical parameters, $I_0 = 1.33 \cdot 10^5\,\text{W/m}^2$, exposure time 120 s – point $(0.0002\,\text{m}, 0.0002\,\text{m})$

Then, calculations were performed for three different laser irradiation powers: $I_0 = 1.33 \cdot 10^5\,\text{W/m}^2$, $I_0 = 2 \cdot 10^5\,\text{W/m}^2$ and $I_0 = 2.5 \cdot 10^5\,\text{W/m}^2$ with an exposure time of 120 s. The temperature profiles marked with a dashed line in Fig. 5 correspond to the model using the function $W(T)$, while the solid line refers to the model without the function taking into account the percentage of water content in the tissue. As observed, at low laser powers, there are no significant differences between the obtained temperatures. However, the discrepancies increase as I_0 increases. The greatest differences occur at the moment when the maximum temperature is reached. Taking into account the process of water evaporation gives lower temperatures, which is related to the release of the latent heat of evaporation of water. In later stages of the process, the temperature profiles become equal again.

Further computations were carried out for higher laser powers, namely: $I_0 = 5 \cdot 10^5\,\text{W/m}^2$, $I_0 = 10{\cdot}10^5\,\text{W/m}^2$, $I_0 = 15{\cdot}10^5\,\text{W/m}^2$, $I_0 = 20{\cdot}10^5\,\text{W/m}^2$ and $I_0 = 25{\cdot}10^5\,\text{W/m}^2$ with the same exposure time 120 s. In Fig. 6, the temperature history at the point $(0.0002\,\text{m}, 0.002\,\text{m})$ for all variants of computations is shown. For high laser powers, the temperature at this point exceeds

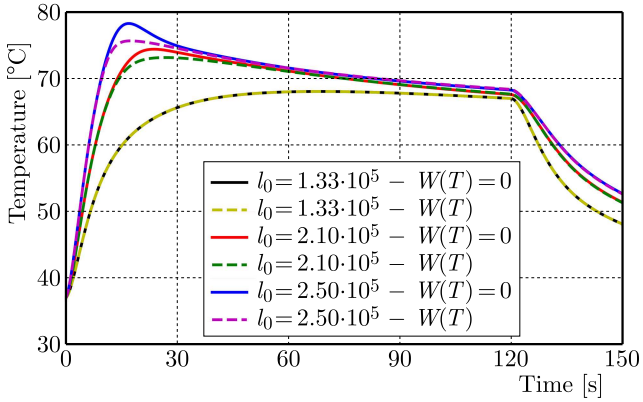


Fig. 5. Temperature history for different laser powers with a zero and non-zero $W(T)$ function – point $(0.0002\,\mathrm{m}, 0.0002\,\mathrm{m})$, exposure time 120 s

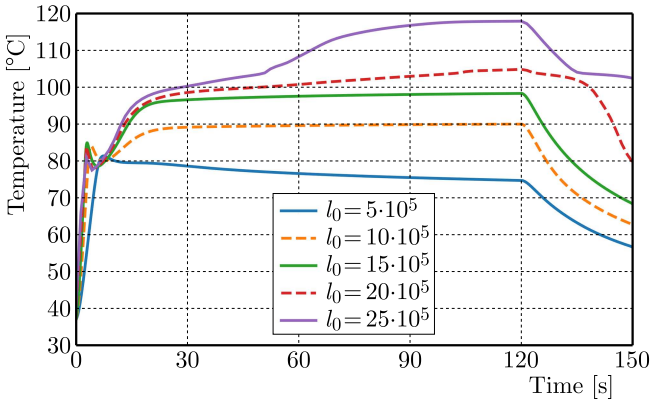


Fig. 6. Temperature history for different laser powers – exposure time 120 s – point $(0.0002\,\mathrm{m}, 0.0002\,\mathrm{m})$

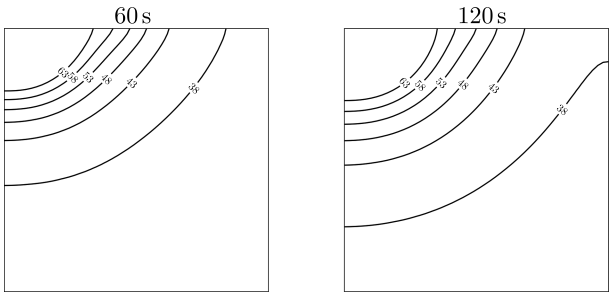


Fig. 7. Temperature distribution after 60 s and 120 s, $I_0 = 25 \cdot 10^5\,\mathrm{W/m^2}$, exposure time 120 s

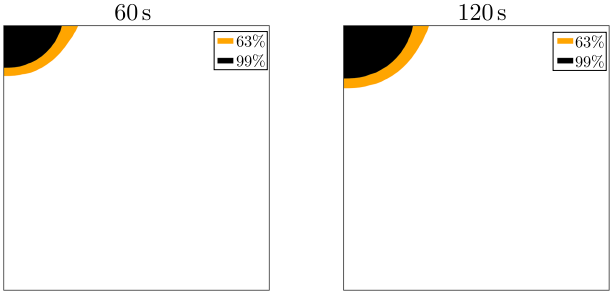


Fig. 8. Arrhenius integral distribution after 60 s and 120 s, $I_0 = 25 \cdot 10^5\,\mathrm{W/m^2}$, exposure time 120 s

100°C, and the water contained in the tissue undergoes an intensive evaporation process. In Figs. 7 and 8, the temperature and Arrhenius integral distributions in the domain for the time 60 s and 120 s are shown. It is clearly visible that for the laser power $I_0 = 25 \cdot 10^5\,\mathrm{W/m^2}$, the region of complete tissue destruction is quite large (e.g. for 60 s) and increases as the heating process continues (e.g. for 120 s).

5. Conclusions

The presented method of modeling of interactions of laser with biological tissue is based on the dual-phase lag equation coupled with the optical diffusion equation. The equations considered take into account thermophysical parameters of tissue that change with temperature and the optical parameters of biological tissue that change with the Arrhenius integral. It has been shown that in the modeling of high-temperature hyperthermia, it is important to use variable optical parameters of the tissue. The use of constant optical parameters leads to excessive temperatures, which may result in an inaccurate assessment of the degree of damage to biological tissues, which may consequently affect the correct planning of artificial hyperthermia treatments. An important element of the computations is the analysis of water percentage in the tissue. Taking into account the function $W(T)$ allows for modeling of the evaporation process. The evaporation of water and other fluids from the tissue significantly affects the obtained temperature values and, consequently, the size of the estimated domain of damage.

Acknowledgment

The research was funded from financial resources from the statutory subsidy of the Faculty of Mechanical Engineering, Silesian University of Technology.

References

1. Abraham J., Sparrow E., 2007, A thermal-ablation bioheat model including liquid-to-vapor phase change, pressure and necrosis-dependent perfusion, and moisture-dependent properties, *International Journal of Heat and Mass Transfer*, **50**, 13-14, 2537-2544
2. Ashley J., Welch M.A.J., Gemert M.J.C. [Edit.], 1995, *Optical-Thermal Response of Laser-Irradiated Tissue*, Plenum Press, New York
3. Barnoon P., Bakhshandehfard F., 2021, Thermal management in a biological tissue in order to destroy tissue under local heating process, *Case Studies in Thermal Engineering*, **26**, ID 101105
4. Dombrovsky L.A., Baillis D., 2010, *Thermal Radiation in Disperse Systems: An Engineering Approach*, Begell House, New York
5. Dombrovsky L.A., Timchenko V., Jackson M., 2012, Indirect heating strategy of laser induced hyperthermia: an advanced thermal model, *International Journal of Heat and Mass Transfer*, **55**, 17-18, 4688-4700
6. Ellebrecht D.B., Theisen-Kunde D., Kuempers Ch., Keck T., Kleemann M., Wolken H., 2018, Analysis of laparoscopic laser liver resection in standardized porcine model, *Surgical Endoscopy*, **32**, 4966-4972
7. Fasano A.,Hömberg D, Naumov D., 2010, On a mathematical model for laser-induced thermotherapy, *Applied Mathematical Modelling*, **34**, 12, 3831-3840
8. Foster J., Hodder S.G., Lloyd A.B., Havenith G., 2020, Individual responses to heat stress: implications for hyperthermia and physical work capacity, *Frontiers in Physiology*, **11**, 28 pp
9. Giglio M.C., Logghe B., Garofalo E., Tomassini F., Vanlander A., *et al.*, 2020, Laparoscopic versus open thermal ablation of colorectal liver metastases: a propensity score-based analysis of local control of the ablated tumors, *Annals of Surgical Oncology*, **27**, 2370-2380

10. Jacques S.L., Pogue B.W., 2008, Tutorial on diffuse light transport, *Journal of Biomedical Optics*, **13**, 4, 1-19

11. Jasiński M., Majchrzak E., Turchan Ł., 2016, Numerical analysis of the interactions between laser and soft tissues using generalized dual-phase lag model, *Applied Mathematical Modeling*, **40**, 2, 750-762

12. Jaunich M., Raje S., Kim K., Mitra K., Guo Z., 2008, Bio-heat transfer analysis during short--pulse laser irradiation of tissues, *International Journal of Heat and Mass Transfer*, **51**, 5511-5521

13. Kim B.M., Jacques S.L., Rastegar S., Thomsen S., Motamedi M., 1996, Nonlinear finite--element analysis of the role of dynamic changes in blood perfusion and optical properties in laser coagulation of tissue, *IEEE Journal of Selected Topics in Quantum Electronics*, **2**, 4, 922-933

14. Kim K., Guo Z., 2007, Multi-time-scale heat transfer modeling of turbid tissues exposed to short-pulsed irradiations, *Computer Methods and Programs in Biomedicine*, **86**, 112-123

15. Lopresto V., Argentieri A., Pinto R., Cavagnaro M., 2019, Temperature dependence of thermal properties of ex vivo liver tissue up to ablative temperatures, *Physics in Medicine and Biology*, **64**, 10, 13 pp

16. Majchrzak E., Stryczyński M., 2022, Numerical analysis of biological tissue heating using the dual-phase lag equation with temperature-dependent parameters, *Journal of Applied Mathematics and Computational Mechanics*, **21**, 3, 85-98

17. Majchrzak E., Turchan Ł., Jasiński M., 2019, Identification of laser intensity assuring the destruction of target region of biological tissue using the gradient method and generalized dual-phase lag equation, *Iranian Journal of Science and Technology – Transactions of Mechanical Engineering*, **43**, 3, 539-548

18. Mochnacki B., Majchrzak E., 2007, Identification of macro and micro parameters in solidification model, *Bulletin of the Polish Academy of Sciences, Technical Sciences*, **55**, 1, 107-113

19. Niemz M.H., 2007, *Laser-Tissue Interaction: Fundamentals and Applications*, Springer-Verlag, Berlin, Heidelberg, New York

20. Szasz A., Szasz N., Szasz O., 2011, *Oncothermia: Principles and Practices*, Springer

21. Tzou D.Y., 1995, A unified field approach for heat conduction from macro- to micro-scales, *Journal of Heat Transfer*, **117**, 1, 8-16

22. Yang D., Converse M.C., Mahvi D.M., Webster J.G., 2007, Expanding the bioheat equation to include tissue internal water evaporation during heating, *IEEE Transactions on Biomedical Engineering*, **54**, 8, 1382-1388

23. Zhou J., Zhang Y., Chen J. K., 2009, An axisymmetric dual-phase lag bio-heat model for laser heating of living tissues, *International Journal of Thermal Sciences*, **48**, 8, 1477-1485

Manuscript received November 8, 2023; accepted for print January 15, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 403-413, Warsaw 2024
https://doi.org/10.15632/jtam-pl/185352

HEAT TRANSFER IN A THIN METAL FILM SUBJECTED TO THE ULTRA-SHORT LASER PULSE MODELED BY A NONLINEAR TWO-TEMPERATURE MODEL[1]

Jolanta Dziatkiewicz, Ewa Majchrzak

Silesian University of Technology, Department of Computational Mechanics and Engineering, Gliwice, Poland
corresponding author J. Dziatkiewicz, e-mail: jolanta.dziatkiewicz@polsl.pl

The heating of a thin metal film subjected to the ultra-short laser pulse is presented. Mathematical description of this process is based on the system of equations describing the electron and lattice temperatures and dependences between intensity of heat fluxes and temperature gradients supplemented by appropriate boundary and initial conditions. In this approach, a system of four equations needs to be solved. In this paper, another method of solution of the above formulated problem is proposed. Using appropriate mathematical manipulations, instead of four equations, two equations describing the lattice and electron temperature distributions are obtained. This system of two equations is solved using an implicit scheme of the finite difference method. The results obtained using both approaches were compared. They were almost identical, which confirms the correctness of the proposed method.

Keywords: microscale heat transfer, two-temperature model, laser, finite difference method

1. Introduction

Heat transfer in the thin metal film domain subjected to the ultra-short laser pulse can be described by different mathematical models. One of them is a two-temperature model (TTM) firstly formulated by Anisimov and co-workers (Anisimov *et al.*, 1974). In this model, two different temperatures, the electron temperature and the lattice temperature appear, and they are described by two coupled Fourier equations. The model based on the classical Fourier law is called the parabolic TTM and it has some limitations. It means, it is valid only when the characteristic space and time scales of the temperature field are much greater compared to the electrons mean free path and the relaxation time (Alexopoulou and Markopoulos, 2023). In turn, the model to be used when the characteristic space and time scales of the temperature field are comparable with the electrons mean free path and relaxation time is called the hyperbolic two--temperature model (Qiu and Tien, 1993; Chen *et al.*, 2004; Smith and Norris, 2003). This model is based on the generalized Fourier law, in which the relaxation time appears, and consists of four equations while two of them describe distributions of lattice and electron temperatures, and two of them describe relationships between intensity of heat fluxes and lattice and electron temperature gradients.

Currently, these models are used for the modeling of thermal processes occurring in the laser treated materials, see e.g. (Chen and Beraun, 2001; Majchrzak and Dziatkiewicz, 2015; Sobolev, 2016). Here one can mention the analytical or semi-analytical methods, e.g. (Oane *et al.* 2019), finite difference method, e.g. (Niu and Dai, 2009; Huang *et al.*, 2011; Dziatkiewicz *et al.*, 2014), finite volume method, e.g. (Qiu and Tien, 1993) and finite element method, e.g. (Saghebfar *et al.*, 2017).

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

A broad literature review on two-temperature models can be found in the paper by Alexopoulou and Markopoulos (2023).

2. Statement of the problem

A thin metal film subjected to the ultra-short laser pulse is considered. Usually, the laser spot size is much larger than film thickness and then it is possible to treat the interactions as a one-dimensional (1D) heat transfer process, wherein the front surface $x = 0$ is irradiated by the laser pulse, and this simplification is used here.

The two-temperature model describes temporal and spatial evolution of the lattice and electron temperatures in the irradiated metal by two coupled nonlinear differential equations (Tzou, 1997; Zhang, 2007)

$$\begin{aligned} C_e(T_e)\frac{\partial T_e(x,t)}{\partial t} &= -\frac{\partial q_e(x,t)}{\partial x} - G(T_e,T_l)[T_e(x,t) - T_l(x,t)] + Q(x,t) \\ C_l(T_l)\frac{\partial T_l(x,t)}{\partial t} &= -\frac{\partial q_l(x,t)}{\partial x} + G(T_e,T_l)[T_e(x,t) - T_l(x,t)] \end{aligned} \tag{2.1}$$

where $T_e(x,t)$, $T_l(x,t)$, $q_e(x,t)$, $q_l(x,t)$ are temperatures and heat fluxes of the electrons and lattice, respectively, $C_e(T_e)$, $C_l(T_l)$ are volumetric specific heats, $G(T_e,T_l)$ is the electron-phonon coupling factor which characterizes the energy exchange between electrons and phonons, $Q(x,t)$ is the source function associated with laser irradiation, x is the spatial coordinate and t denotes time.

The following relationships between the intensity of heat fluxes and temperature gradients proposed by Qiu and Tien (1993) are used

$$\begin{aligned} q_e(x,t+\tau_e) &= -\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x} \\ q_l(x,t+\tau_l) &= -\lambda_l(T_l)\frac{\partial T_l(x,t)}{\partial x} \end{aligned} \tag{2.2}$$

where τ_e is the relaxation time of free electrons in metals (the mean time for electrons to change their states), τ_l is the relaxation time in phonon collisions, $\lambda_e(T_e,T_l)$, $\lambda_l(T_l)$ are the thermal conductivities of electrons and lattice, respectively.

Expanding the left-hand sides of equations (2.2) into the Taylor series with an accuracy of two terms, one obtains

$$\begin{aligned} q_e(x,t) + \tau_e\frac{\partial q_e(x,t)}{\partial t} &= -\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x} \\ q_l(x,t) + \tau_l\frac{\partial q_l(x,t)}{\partial t} &= -\lambda_l(T_l)\frac{\partial T_l(x,t)}{\partial x} \end{aligned} \tag{2.3}$$

The source function $Q(x,t)$ is associated with laser irradiation (Chen and Beraun, 2001; Majchrzak and Dziatkiewicz, 2019)

$$Q(x,t) = \sqrt{\frac{\beta}{\pi}}\frac{1-R}{t_p\delta}I\exp\Big[-\frac{x}{\delta} - \beta\frac{(t-2t_p)^2}{t_p^2}\Big] \tag{2.4}$$

where I is the laser intensity, t_p is the characteristic time of the laser pulse, δ is the optical penetration depth, R is the reflectivity of the irradiated surface and $\beta = 4\ln 2$.

For $x = 0$ and $x = L$ the non-flux conditions are assumed, and the initial condition $T_e(x,0) = T_l(x,0) = T_p$, where T_p is the initial temperature of electrons and lattice, is also known.

In literature, the algorithms that involve simultaneous solution of equations (2.1) and (2.3) using a staggered grid are presented, e.g. (Huang *et al.*, 2009; Majchrzak *et al.*, 2017; Wang *et al.*, 2006, 2008). It means that for even nodes the temperatures are calculated, and for odd nodes the intensity of the heat fluxes are determined (1D problem). In this paper, another method of solution of the above formulated problem is proposed. Using appropriate mathematical manipulations, instead of four equations, two equations describing the lattice and electron temperature distributions are obtained. This system of two equations is solved using an implicit scheme of the finite difference method.

3. Mathematical model

In this Section, the mathematical manipulations leading to two equations describing the heat transfer in a thin metal layer subjected to the ultra-short laser pulse are presented.

From equation $(2.3)_1$ it follows that

$$-q_e(x,t) = \tau_e \frac{\partial q_e(x,t)}{\partial t} + \lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x} \tag{3.1}$$

therefore

$$-\frac{\partial q_e(x,t)}{\partial x} = \tau_e \frac{\partial^2 q_e(x,t)}{\partial t \partial x} + \frac{\partial}{\partial x}\Big[\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x}\Big] \tag{3.2}$$

Formula (3.2) is introduced into equation $(2.1)_1$, and then

$$\begin{aligned} C_e(T_e)\frac{\partial T_e(x,t)}{\partial t} &= \tau_e \frac{\partial^2 q_e(x,t)}{\partial t \partial x} + \frac{\partial}{\partial x}\Big[\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x}\Big] \\ &- G(T_e,T_l)[T_e(x,t) - T_l(x,t)] + Q(x,t) \end{aligned} \tag{3.3}$$

It follows from equation $(2.1)_1$ that

$$\frac{\partial q_e(x,t)}{\partial x} = -C_e(T_e)\frac{\partial T_e(x,t)}{\partial t} - G(T_e,T_l)[T_e(x,t) - T_l(x,t)] + Q(x,t) \tag{3.4}$$

Introducing (3.4) into equation (3.3), one has

$$\begin{aligned} C_e(T_e)\frac{\partial T_e(x,t)}{\partial t} &= \tau_e \frac{\partial}{\partial t}\Big[-C_e(T_e)\frac{\partial T_e(x,t)}{\partial t} - G(T_e,T_l)[T_e(x,t) - T_l(x,t)] + Q(x,t)\Big] \\ &+ \frac{\partial}{\partial x}\Big[\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x}\Big] - G(T_e,T_l)[T_e(x,t) - T_l(x,t)] + Q(x,t) \end{aligned} \tag{3.5}$$

or

$$\begin{aligned} C_e(T_e)\frac{\partial T_e(x,t)}{\partial t} + \tau_e \frac{\partial}{\partial t}\Big[C_e(T_e)\frac{\partial T_e(x,t)}{\partial t}\Big] &= \frac{\partial}{\partial x}\Big[\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x}\Big] \\ - G(T_e,T_l)[T_e(x,t) - T_l(x,t)] - \tau_e \frac{\partial}{\partial t}\Big\{G(T_e,T_l)[T_e(x,t) - T_l(x,t)]\Big\} \\ + Q(x,t) + \tau_e \frac{\partial Q(x,t)}{\partial t} \end{aligned} \tag{3.6}$$

In a similar way, the equation describing the lattice temperature can be derived

$$\begin{aligned} C_l(T_l)\frac{\partial T_l(x,t)}{\partial t} + \tau_l \frac{\partial}{\partial t}\Big[C_l(T_l)\frac{\partial T_l(x,t)}{\partial t}\Big] &= \frac{\partial}{\partial x}\Big[\lambda_l(T_l)\frac{\partial T_l(x,t)}{\partial x}\Big] \\ + G(T_e,T_l)[T_e(x,t) - T_l(x,t)] + \tau_l \frac{\partial}{\partial t}\{G(T_e,T_l)[T_e(x,t) - T_l(x,t)]\} \end{aligned} \tag{3.7}$$

Equations (3.6) and (3.7) can be written in the form

$$
\begin{aligned}
&C_e(T_e)\Big[\frac{\partial T_e(x,t)}{\partial t}+\tau_e\frac{\partial^2 T_e(x,t)}{\partial t^2}\Big]+\tau_e\frac{\partial C_e(T_e)}{\partial t}\frac{\partial T_e(x,t)}{\partial t}=\frac{\partial}{\partial x}\Big[\lambda_e(T_e,T_l)\frac{\partial T_e(x,t)}{\partial x}\Big]\\
&\quad-G(T_e,T_l)[T_e(x,t)-T_l(x,t)]-\tau_e\frac{\partial G(T_e,T_l)}{\partial t}[T_e(x,t)-T_l(x,t)]\\
&\quad-\tau_e G(T_e,T_l)\Big[\frac{\partial T_e(x,t)}{\partial t}-\frac{\partial T_l(x,t)}{\partial t}\Big]+Q(x,t)+\tau_e\frac{\partial Q(x,t)}{\partial t}\\
&C_l(T_l)\Big[\frac{\partial T_l(x,t)}{\partial t}+\tau_l\frac{\partial^2 T_l(x,t)}{\partial t^2}\Big]+\tau_l\frac{\partial C_l(T_l)}{\partial t}\frac{\partial T_l(x,t)}{\partial t}=\frac{\partial}{\partial x}\Big[\lambda_l(T_l)\frac{\partial T_l(x,t)}{\partial x}\Big]\\
&\quad+G(T_e,T_l)[T_e(x,t)-T_l(x,t)]+\tau_l\frac{\partial G(T_e,T_l)}{\partial t}[T_e(x,t)-T_l(x,t)]\\
&\quad+\tau_l G(T_e,T_l)\Big[\frac{\partial T_e(x,t)}{\partial t}-\frac{\partial T_l(x,t)}{\partial t}\Big]
\end{aligned}
\tag{3.8}
$$

By performing derivative operations, one has (arguments omitted for simplicity)

$$
\begin{aligned}
&C_e(T_e)\Big(\frac{\partial T_e}{\partial t}+\tau_e\frac{\partial^2 T_e}{\partial t^2}\Big)+\tau_e\frac{dC_e(T_e)}{dT_e}\Big(\frac{\partial T_e}{\partial t}\Big)^2=\lambda_e(T_e,T_l)\frac{\partial^2 T_e}{\partial x^2}\\
&\quad+\Big[\frac{\partial\lambda_e(T_e,T_l)}{\partial T_e}\frac{\partial T_e}{\partial x}+\frac{\partial\lambda_e(T_e,T_l)}{\partial T_l}\frac{\partial T_l}{\partial x}\Big]\frac{\partial T_e}{\partial x}-G(T_e,T_l)(T_e-T_l)\\
&\quad-\tau_e\Big[\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e}{\partial t}+\frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l}{\partial t}\Big](T_e-T_l)\\
&\quad-\tau_e G(T_e,T_l)\Big(\frac{\partial T_e}{\partial t}-\frac{\partial T_l}{\partial t}\Big)+Q+\tau_e\frac{\partial Q}{\partial t}\\
&C_l(T_l)\Big(\frac{\partial T_l}{\partial t}+\tau_l\frac{\partial^2 T_l}{\partial t^2}\Big)+\tau_l\frac{dC_l(T_l)}{dT_l}\Big(\frac{\partial T_l}{\partial t}\Big)^2=\lambda_l(T_l)\frac{\partial^2 T_l}{\partial x^2}+\frac{d\lambda_l(T_l)}{dT_l}\Big(\frac{\partial T_l}{\partial x}\Big)^2\\
&\quad+G(T_e,T_l)(T_e-T_l)+\tau_l\Big[\frac{\partial G(T_e,T_l)}{\partial T_e}\frac{\partial T_e}{\partial t}+\frac{\partial G(T_e,T_l)}{\partial T_l}\frac{\partial T_l}{\partial t}\Big](T_e-T_l)\\
&\quad+\tau_l G(T_e,T_l)\Big(\frac{\partial T_e}{\partial t}-\frac{\partial T_l}{\partial t}\Big)
\end{aligned}
\tag{3.9}
$$

Summing up, in the proposed approach instead of solving four equations (2.1) and (2.3) it is enough to solve two equations (3.9) supplemented by appropriate boundary and initial conditions.

4. Method of solution

The problem formulated is solved using an implicit scheme of the finite difference method (Majchrzak and Dziatkiewicz, 2015; Niu and Dai, 2009; Wang *et al.*, 2008). Let us denote $T_i^f = T(ih, f\Delta t)$, where h is the mesh step, Δt is the time step, $i = 0, 1, 2, \ldots, n$. Using the appropriate difference quotients, the following approximation of equation $(3.9)_1$ is proposed

$$
\begin{aligned}
&C_{ei}^{f-1}\Big(\frac{T_{ei}^f-T_{ei}^{f-1}}{\Delta t}+\tau_e\frac{T_{ei}^f-2T_{ei}^{f-1}+T_{ei}^{f-2}}{(\Delta t)^2}\Big)+\tau_e\Big[\frac{dC_e(T_e)}{dT_e}\Big]_i^{f-1}\Big(\frac{T_{ei}^{f-1}-T_{ei}^{f-2}}{\Delta t}\Big)^2\\
&\quad=\lambda_{ei}^{f-1}\frac{T_{ei-1}^f-2T_{ei}^f+T_{ei+1}^f}{h^2}+D_{ei}^{f-1}-G_i^{f-1}(T_{ei}^{f-1}-T_{li}^{f-1})-E_{ei}^{f-1}\\
&\quad-\tau_e G_i^{f-1}\Big(\frac{T_{ei}^{f-1}-T_{ei}^{f-2}}{\Delta t}-\frac{T_{li}^{f-1}-T_{li}^{f-2}}{\Delta t}\Big)+Q_i^f+\tau_e\Big(\frac{\partial Q}{\partial t}\Big)_i^f
\end{aligned}
\tag{4.1}
$$

where

$$D_{ei}^{f-1} = \left\{ \Big[\frac{\partial \lambda_e(T_e,T_l)}{\partial T_e}\Big]_i^{f-1} \frac{T_{ei+1}^{f-1} - T_{ei-1}^{f-1}}{2h} + \Big[\frac{\partial \lambda_e(T_e,T_l)}{\partial T_l}\Big]_i^{f-1} \frac{T_{li+1}^{f-1} - T_{li-1}^{f-1}}{2h} \right\} \frac{T_{ei+1}^{f-1} - T_{ei-1}^{f-1}}{2h}$$
$$E_{ei}^{f-1} = \tau_e \left\{ \Big[\frac{\partial G(T_e,T_l)}{\partial T_e}\Big]_i^{f-1} \frac{T_{ei}^{f-1} - T_{ei}^{f-2}}{\Delta t} + \Big[\frac{\partial G(T_e,T_l)}{\partial T_l}\Big]_i^{f-1} \frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t} \right\} (T_{ei}^{f-1} - T_{li}^{f-1}) \tag{4.2}$$

Equation (4.1) can be written in the form

$$T_{ei}^f = \frac{\lambda_{ei}^{f-1}}{A_{ei}^{f-1} h^2}(T_{ei-1}^f + T_{ei+1}^f) + \frac{F_{ei}^{f-1}}{A_{ei}^{f-1}} + \frac{1}{A_{ei}^{f-1}}\Big[Q_i^f + \tau_e \Big(\frac{\partial Q}{\partial t}\Big)_i^f\Big] \tag{4.3}$$

where

$$\begin{aligned} A_{ei}^{f-1} &= C_{ei}^{f-1}\frac{\Delta t + \tau_e}{(\Delta t)^2} + \frac{2\lambda_{ei}^{f-1}}{h^2} \\ F_{ei}^{f-1} &= C_{ei}^{f-1}\frac{\Delta t + 2\tau_e}{(\Delta t)^2} T_{ei}^{f-1} - C_{ei}^{f-1}\frac{\tau_e}{(\Delta t)^2} T_{ei}^{f-2} - \tau_e \Big[\frac{dC_e(T_e)}{dT_e}\Big]_i^{f-1} \Big(\frac{T_{ei}^{f-1} - T_{ei}^{f-2}}{\Delta t}\Big)^2 \\ &\quad + D_{ei}^{f-1} - G_i^{f-1}(T_{ei}^{f-1} - T_{li}^{f-1}) - E_{ei}^{f-1} - \tau_e G_i^{f-1}\Big(\frac{T_{ei}^{f-1} - T_{ei}^{f-2}}{\Delta t} - \frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t}\Big) \end{aligned} \tag{4.4}$$

In a similar way, equation $(3.9)_2$ is approximated, namely

$$\begin{aligned} &C_{li}^{f-1}\Big(\frac{T_{li}^f - T_{li}^{f-1}}{\Delta t} + \tau_l \frac{T_{li}^f - 2T_{li}^{f-1} + T_{li}^{f-2}}{(\Delta t)^2}\Big) + \tau_l \Big[\frac{dC_l(T_l)}{dT_l}\Big]_i^{f-1} \Big(\frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t}\Big)^2 \\ &\quad = \lambda_{li}^{f-1}\frac{T_{li-1}^f - 2T_{li}^f + T_{li+1}^f}{h^2} + \Big[\frac{d\lambda_l(T_l)}{dT_l}\Big]_i^{f-1}\Big(\frac{T_{li+1}^{f-1} - T_{li-1}^{f-1}}{2h}\Big)^2 \\ &\quad + G_i^{f-1}(T_{ei}^{f-1} - T_{li}^{f-1}) + E_{li}^{f-1} + \tau_l G_i^{f-1}\Big(\frac{T_{ei}^{f-1} - T_{ei}^{f-2}}{\Delta t} - \frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t}\Big) \end{aligned} \tag{4.5}$$

where

$$E_{li}^{f-1} = \tau_l \left\{ \Big[\frac{\partial G(T_e,T_l)}{\partial T_e}\Big]_i^{f-1} \frac{T_{ei}^{f-1} - T_{ei}^{f-2}}{\Delta t} + \Big[\frac{\partial G(T_e,T_l)}{\partial T_l}\Big]_i^{f-1} \frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t} \right\} (T_{ei}^{f-1} - T_{li}^{f-1}) \tag{4.6}$$

Equation (4.5) can be written in the form

$$T_{li}^f = \frac{\lambda_{li}^{f-1}}{A_{li}^{f-1} h^2}(T_{li-1}^f + T_{li+1}^f) + \frac{F_{li}^{f-1}}{A_{li}^{f-1}} \tag{4.7}$$

where

$$\begin{aligned} A_{li}^{f-1} &= C_{li}^{f-1}\frac{\Delta t + \tau_l}{(\Delta t)^2} + \frac{2\lambda_{li}^{f-1}}{h^2} \\ F_{li}^{f-1} &= C_{li}^{f-1}\frac{\Delta t + 2\tau_l}{(\Delta t)^2} T_{li}^{f-1} - C_{li}^{f-1}\frac{\tau_l}{(\Delta t)^2} T_{li}^{f-2} - \tau_l \Big[\frac{dC_l(T_l)}{dT_l}\Big]_i^{f-1} \Big(\frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t}\Big)^2 \\ &\quad + \Big[\frac{d\lambda_l(T_l)}{dT_l}\Big]_i^{f-1}\Big(\frac{T_{li+1}^{f-1} - T_{li-1}^{f-1}}{2h}\Big)^2 + G_i^{f-1}(T_{ei}^{f-1} - T_{li}^{f-1}) + E_{li}^{f-1} \\ &\quad + \tau_l G_i^{f-1}\Big(\frac{T_{ei}^{f-1} - T_{ei}^{f-2}}{\Delta t} - \frac{T_{li}^{f-1} - T_{li}^{f-2}}{\Delta t}\Big) \end{aligned} \tag{4.8}$$

The non-flux boundary conditions are also approximated

$$\begin{aligned} &x=0: \quad \frac{T_{e1}^f - T_{e0}^f}{h} = 0 \qquad x=L: \quad \frac{T_{en}^f - T_{en-1}^f}{h} = 0 \\ &x=0: \quad \frac{T_{l1}^f - T_{l0}^f}{h} = 0 \qquad x=L: \quad \frac{T_{ln}^f - T_{ln-1}^f}{h} = 0 \end{aligned} \tag{4.9}$$

that is

$$T_{e0}^f = T_{e1}^f \qquad T_{en}^f = T_{en-1}^f \qquad T_{l0}^f = T_{l1}^f \qquad T_{ln}^f = T_{ln-1}^f \tag{4.10}$$

From the initial condition, it follows that

$$T_{ei}^0 = T_{ei}^1 = T_p \qquad T_{li}^0 = T_{li}^1 = T_p \qquad i = 0, 1, \ldots, n \tag{4.11}$$

For each transition $t^{f-1} \to t^f$, the system of equations (4.3), (4.7), (4.10) is solved using e.g. the Gauss-Seidel iterative method.

5. Results of computations

A gold film of thickness $L = 100\,\text{nm}$ ($1\,\text{nm} = 10^{-9}\,\text{m}$) is considered. The initial temperature is equal to $T_p = 300\,\text{K}$.

For a high laser intensity, the following formula describing temperature-dependent volumetric specific heat of electrons is proposed (Huang *et al.*, 2009, 2011; Majchrzak and Dziatkiewicz, 2019)

$$C_e(T_e) = \begin{cases} AT_e & \text{for} \quad T_e < \dfrac{T_F}{\pi^2} \\ A\dfrac{T_F}{\pi^2} + \dfrac{Nk_B - AT_F/\pi^2}{2T_F/\pi^2}\Big(T_e - \dfrac{T_F}{\pi^2}\Big) & \text{for} \quad \dfrac{T_F}{\pi^2} \leqslant T_e < 3\dfrac{T_F}{\pi^2} \\ Nk_B + \dfrac{Nk_B/2}{T_F - 3T_F/\pi^2}\Big(T_e - 3\dfrac{T_F}{\pi^2}\Big) & \text{for} \quad 3\dfrac{T_F}{\pi^2} \leqslant T_e < T_F \\ 3N\dfrac{k_B}{2} & \text{for} \quad T_e \geqslant T_F \end{cases} \tag{5.1}$$

where $N = 5.9 \cdot 10^{28}\,\text{m}^{-1}$ is the electron concentration, $T_F = 64\,200\,\text{K}$ is the Fermi temperature, k_B is the Boltzmann constant and A is given by formula $A = \pi^2 N k_B/(2T_F) = 62.7\,\text{J/(m}^3\text{K)}$.

The electrons thermal conductivity is described by the formula (Huang, 2011; Majchrzak and Dziatkiewicz, 2015)

$$\lambda_e(T_e, T_l) = \chi \frac{[(T_e/T_F)^2 + 0.16]^{5/4}[(T_e/T_F)^2 + 0.44](T_e/T_F)}{[(T_e/T_F)^2 + 0.092]^{1/2}[(T_e/T_F)^2 + \eta(T_l/T_F)]} \tag{5.2}$$

and the coupling factor

$$G(T_e, T_l) = G_{rt}\Big[\frac{A_e}{B_l}(T_e + T_l) + 1\Big] \tag{5.3}$$

where $\chi = 353\,\text{W/(mK)}$, $\eta = 0.16$, $A_e = 1.2 \cdot 10^7\,1/(\text{K}^2\text{s})$, $B_l = 1.23 \cdot 10^{11}\,1/(\text{Ks})$ and $G_{rt} = 2.2 \cdot 10^{16}\,\text{W/(m}^3\text{K)}$ (Majchrzak and Dziatkiewicz, 2015).

Temperature dependent thermal conductivity and volumetric specific heat of gold are taken from (Huang *et al.*, 2009, 2011)

$$\lambda_l(T_l)\ \Big[\frac{\text{W}}{\text{m K}}\Big] = \begin{cases} 320.973 - 0.0111T_l - 2.747 \cdot 10^{-5}T_l^2 - 4.048 \cdot 10^{-9}T_l^3 & T_l \leqslant 1336\,\text{K} \\ 37.72 + 0.0711T_l - 1.721 \cdot 10^{-5}T_l^2 + 1.064 \cdot 10^{-9}T_l^3 & T_l > 1336\,\text{K} \end{cases} \tag{5.4}$$

and

$$C_l(T_l)\left[\frac{\mathrm{J}}{\mathrm{m^3K}}\right]=\begin{cases}(105.1+0.2941T_l-8.731\cdot10^{-4}T_l^2+1.787\cdot10^{-6}T_l^3 & \\ \quad -7.051\cdot10^{-10}T_l^4+1.538\cdot10^{-13}T_l^5)19300 & T_l\leqslant 1336\,\mathrm{K}\\ 163.205\cdot 17280 & T_l>1336\,\mathrm{K}\end{cases}\tag{5.5}$$

The other parameters are as follows: electrons relaxation time $\tau_e = 0.04\,\mathrm{ps}$, phonons relaxation time $\tau_l = 0.8\,\mathrm{ps}$ (Chen and Beraun, 2001), reflectivity $R = 0.93$, optical penetration depth $\delta = 15.3\,\mathrm{nm}$.

The problem is solved using the finite difference method on the assumption that $\Delta t = 0.002\,\mathrm{ps}$ and $h = 1\,\mathrm{nm}$.

First, calculations were performed for the laser intensity $I = 4182\,\mathrm{J/m^2}$ and the characteristic time of laser pulse $t_p = 0.1\,\mathrm{ps}$. In Figs. 1 and 2, the electrons and lattice temperature histories on the irradiated surface are presented. These temperatures were compared with the results obtained using a repeatedly verified algorithm based on the simultaneous solution of four equations (2.1) and (2.3) using the staggered grid and thoroughly described, among others in (Majchrzak and Dziatkiewicz, 2015, 2019). As can be seen, the results are almost identical, which confirms the correctness of the algorithm and the authors' computer program presented in this paper.

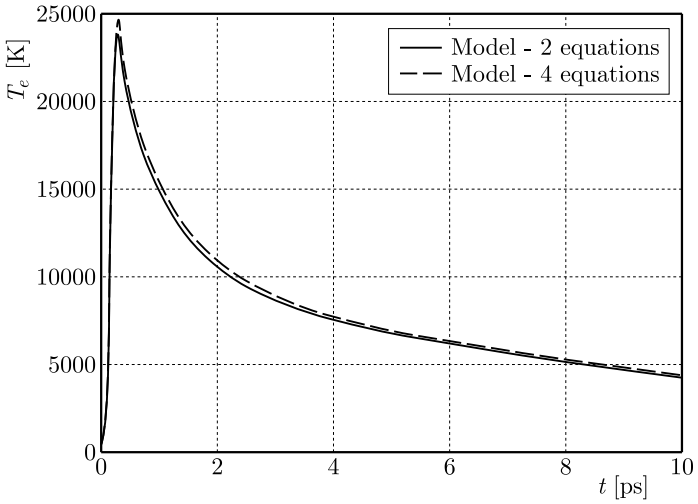


Fig. 1. Comparison of calculated electron temperature

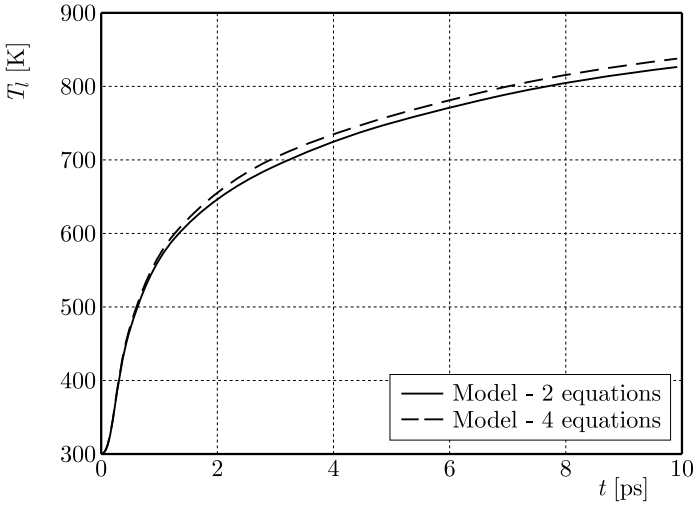


Fig. 2. Comparison of calculated lattice temperature

In Figs. 3 and 4, the temperature distributions of electrons and lattice for selected moments of time are presented. The solid line shows the calculation results for the model with two equations and the dashed line for the model with four equations. It can be seen that the obtained results are very cosistent.

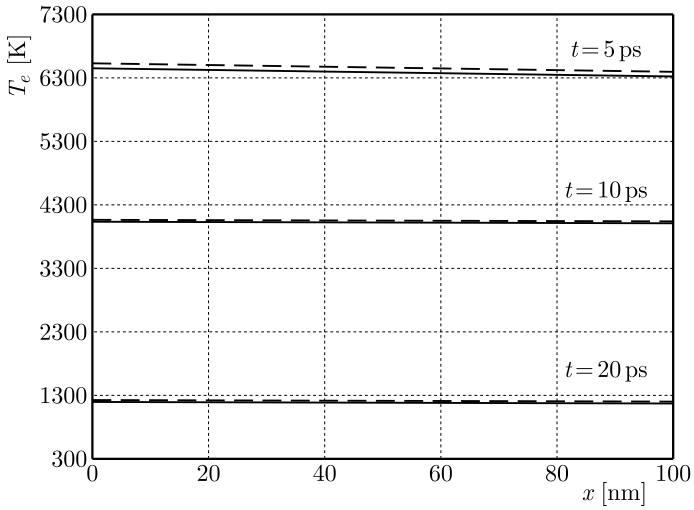


Fig. 3. Comparison of calculated electron temperature for different times

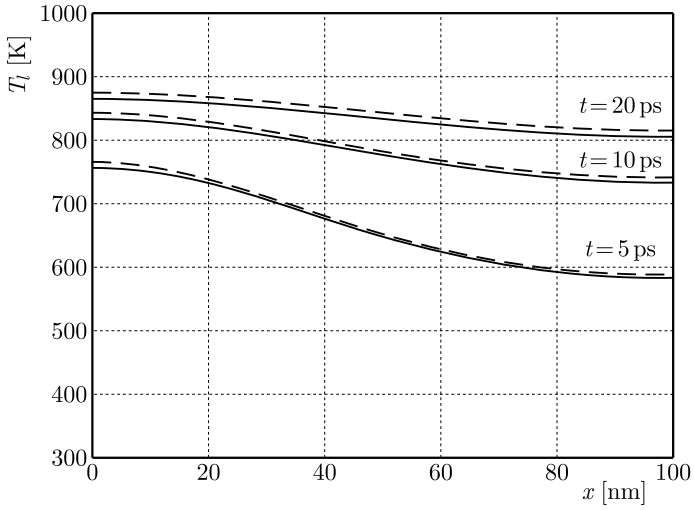


Fig. 4. Comparison of calculated lattice temperature for different times

Then, the influence of the parameters C_l and λ_l on the obtained results was checked. Calculations were prepared for constant values of C_l and λ_l equal to $2.5 \cdot 10^6$ J/(m^{3}K) and 315 W/(mK), respectively, and for values obtained from formulas (5.4) and (5.5).

Figures 5 and 6 show the distribution of temperature of electrons and lattice over time. The solid line shows the results obtained for variable parametrs and the dashed line for constant values.

Calculations were performed for low and high laser intensities. It can be noticed that as the laser intensity increases, the use of constant parameters C_l and λ_l is inappropriate and the obtained results differ significantly. For low intensities, the results are almost identical.

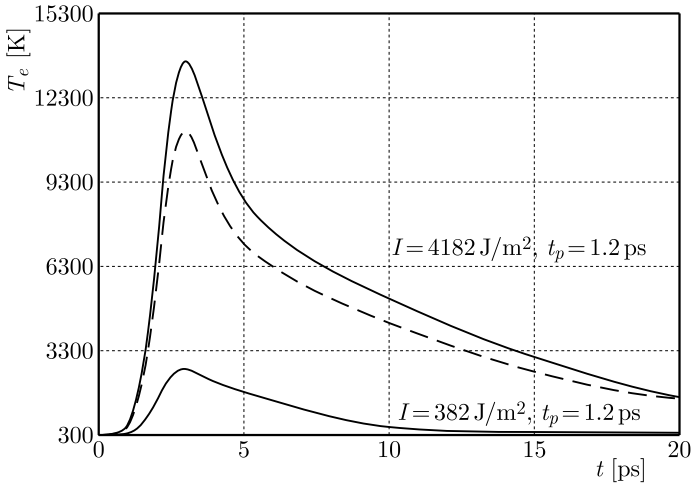


Fig. 5. Electron temperature distribution

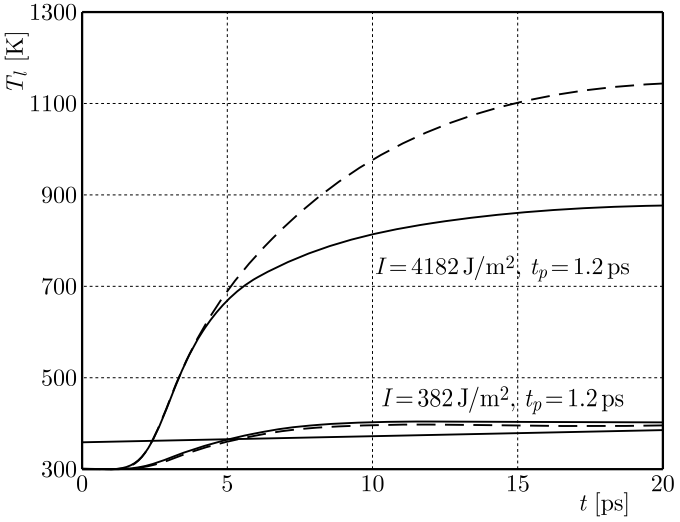


Fig. 6. Lattice temperature distribution

6. Conclusions

In this paper, the laser heating of a thin metal film made of gold is analyzed. A two-temperature model containing four equations (2.1) and (2.3) and the proposed approach with two equations (3.9) were considered. These problems were solved using the finite difference method. The problem with two equations was solved by an implicit the scheme of finite difference method, while the problem with four equations was solved by an explicit scheme of the finite difference method with the staggered grid. The results were compared, and it was shown that they are almost identical, which confirms the correctness of the proposed approach based on the two equations.

In the future, the presented approach can be extended to an axisymmetric (spatial) task, which better reflects the course of the analyzed phenomenon.

The developed algorithm and computer program should be supplemented with procedures that take into account melting, evaporation and ablation processes (Alexopoulou and Markopoulos, 2023). This will allow one to analyse thermal processes occurring in thin metal layers under the influence of higher laser powers.

Acknowledgment

The research was funded from financial resources from the statutory subsidy of the Faculty of Mechanical Engineering, Silesian University of Technology.

References

1. Alexopoulou, V.E., Markopoulos A.P., 2023, A critical assessment regarding two--temperature models: an investigation of the different forms of two-temperature models, the various ultrashort pulsed laser models and computational methods, *Archives of Computational Methods in Engineering*, **31**, 93-123
2. Anisimov S.I., Kapeliovich B.L., Perel'man T.L., 1974, Electron emission from metal surfaces exposed to ultrashort laser pulses, *Zhurnal Eksperimental'noi i Teroreticheskoi Fiziki*, **66**, 776-781
3. Chen J.K., Beraun J.E., 2001, Numerical study of ultrashort laser pulse interactions with metal films, *Numerical Heat Transfer, Part A*, **40**, 1-20
4. Chen G., Borca-Tasciuc D., Yang R.G., 2004, [In:] *Encyclopedia of Nanoscience and Nanotechnology*, Hari Singh Nalwa (Ed.), American Scientific Publishers: Stevenson Ranch, **7**, 429-459
5. Dziatkiewicz J., Kus W., Majchrzak E., Burczyński T., Turchan L., 2014, Bioinspired identification of parameters in microscale heat transfer, *International Journal for Multiscale Computational Engineering*, **12**, 1, 79-89
6. Huang J., Baheti K., Chen J. K., Zhang Y., 2011, An axisymmetric model for solid-liquid--vapor phase change in thin metal films induced by an ultrashort laser pulse, *Frontiers in Heat and Mass Transfer*, **2**, 1, 1-10
7. Huang J., Zhang Y., Chen J.K., 2009, Ultrafast solid-liquid-vapor phase change in a thin gold film irradiated by multiple femtosecond laser pulses, *International Journal of Heat and Mass Transfer*, **52**, 3091-3100
8. Lin Z., Zhigilei L.V., Celli V., 2008, Electron-phonon coupling and electron heat capacity of metals under conditions of strong electron-phonon nonequilibrium, *Physical Review B*, **77**, 075133--1-0.75133-17
9. Majchrzak E., Dziatkiewicz J., 2015, Analysis of ultrashort laser pulse interactions with metal films using a two-temperature model, *Journal of Applied Mathematics and Computational Mechanics*, **14**, 2, 31-39
10. Majchrzak E., Dziatkiewicz J., 2019, Second-order two-temperature model of heat transfer processes in a thin metal film subjected to an ultrashort laser pulse, *Archives of Mechanics*, **71**, 4-5, 377-391
11. Majchrzak E., Dziatkiewicz J., Turchan L., 2017, Analysis of thermal processes occuring in the microdomain subjected to the ultrashort laser pulse using the axisymmetric two-temperature model, *International Journal for Multiscale Computational Engineering*, **15**, 5, 395-411
12. Niu T., Dai W.A., 2009, A hyperbolic two-step model based finite difference scheme for studying thermal deformation in a double-layered thin film exposed to ultrashort-pulsed lasers, *International Journal of Thermal Sciences*, **48**, 34-49
13. Oane M., Mihailescu I.N., Sava B., 2019, The linearized Fourier thermal model applied to Au nanoparticles 1D and 2D lattices under intense nanoseconds laser irradiation pulses, *Journal of Material Sciences and Engineering*, **8**, 1, 1-6
14. Qiu T.Q., Tien C.L., 1993, Heat transfer mechanisms during short-pulse laser heating of metals, *Journal of Heat Transfer*, **115**, 835-841
15. Saghebfar M., Tehrani M.K., Darbani S.M.R., Majd A.E., 2017, Femtosecond pulse laser irradiation of gold/chromium double-layer metal film: The role of interface boundary resistance in two-temperature model simulations, *Thin Solid Films*, **636**, 464-473
16. Smith A.N., Norris P.M., 2003, [In:] *Heat Transfer Handbook*, Adrian Bejan (Ed.), John Wiley & Sons, Hoboken, 1309-1409
17. Sobolev S.L., 2016, Nonlocal two-temperature model: Application to heat transport in metals irradiated by ultrashort laser pulses, *International Journal of Heat and Mass Transfer*, **94**, 138-144

18. Tzou D.Y., 1997, *Macro- to Microscale Heat Transfer. The Lagging Behavior*, Taylor and Francis

19. Wang H., Dai W., Hewavitharana L.G., 2008, A finite difference method for studying thermal deformation in a double-layered thin film with imperfect interfacial contact exposed to ultrashort pulsed lasers, *International Journal of Thermal Sciences*, **47**, 7-24

20. Wang H., Dai W., Melnik R.A., 2006, Finite difference method for studying thermal deformation in a double-layered thin film exposed to ultrashort pulsed lasers, *International Journal of Thermal Sciences*, **45**, 1179-1196

21. Zhang Z.M., 2007, *Nano/microscale Heat Transfer*, McGraw-Hill, New York

Manuscript received November 3, 2023; accepted for print December 18, 2023

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 415-419, Warsaw 2024
https://doi.org/10.15632/jtam-pl/186510

IDENTIFICATION OF NICKEL-TITANIUM ALLOY MATERIAL MODEL PARAMETERS BASED ON EXPERIMENTAL RESEARCH[1]

Jonasz Hartwich, Sebastian Sławski, Sławomir Duda
Silesian University of Technology, Department of Theoretical and Applied Mechanics, Gliwice, Poland
e-mail: jonasz.hartwich@polsl.pl; sebastian.slawski@polsl.pl; slawomir.duda@polsl.pl

The paper presents an identification process of model parameters of a thin nickel-titanium alloy wire based on experimental research. The wire made of NiTi alloy was subjected to a tensile test to obtain the stress-strain characteristic. Parameters of the non-linear material model were identified based on the obtained experimental results. The material model used in the conducted research may be used for simulation of the shape memory effect and pseudoelasticity of the shape memory alloy. The generated results of numerical simulations have a good approximation with the conducted experimental tests.

Keywords: NiTi, shape memory alloy, numerical research, NiTi stress-strain curves

1. Introduction

The sector of smart materials has been for many years characterized by a dynamic development. Smart materials can be used both in sensor and actuator applications because of their unique properties. Another advantage of smart materials is the possibility to use them in self sensing applications. Smart materials are defined as substances which have one or more properties that can be significantly modified in a controlled manner by external stimuli; such as stress, temperature, electric or magnetic field, radiation, pH, moisture, or chemical compounds (Chopra, 1996).

An important group among smart materials are Shape-Memory Alloys (SMA). SMA, in response to a change in environmental conditions, change their internal structure (phase), which leads to a change in the properties of the alloy (Pieczyska *et al.*, 2006). The factor inducing the phase transformation in SMA depends on the type of the alloy.

Among the SMAs, the nickel-titanium alloys have raised a great scientific interest, and have the largest number of industrial applications. The shape memory effect (SME) in NiTi alloys is due to the phase transformation (from martensite to austenite and reverse) related to a temperature change of the alloy (Abel *et al.*, 2004). There are three main crystal structures in NiTi alloys: twinned martensite, detwinned martensite and austenite. In some commercially produced alloys, a rhombohedral R phase may also be observed (Kciuk *et al.*, 2019; Tobushi *et al.*, 2009). The crystal structures existing in NiTi alloys are different for their mechanical and physical properties, including their electrical resistance (Sławski *et al.*, 2021). This makes it possible to identify phase transformations occurring in NiTi alloys due to the fact that electrical resistance is easily measurable (Antonucci *et al.*, 2007). In addition, the correlation between resistance and internal transformations of NiTi alloys makes it possible to rationalize self-sensing control using resistance as a feedback signal (Sławski *et al.*, 2022).

The main purpose of this paper is identification of the NiTi material model parameters. This paper presents a process of identification of thin NiTi wires material model parameters, which is

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

provided with the use of a dedicated test stand. The identified properties are used in a numerical research based on the finite element method (FEM). Results obtained from complex numerical models depend on parameters declared for each component of the model. So, it is important to perform a test concerning material parameters identification to obtain a more accurate numerical model. Validated NiTi material model parameters established as results of the presented paper will be used in further numerical research concerning analysis of polymer based composites containing NiTi wires. Section 2 discusses the material properties and, in addition, the method of performing the experimental research. In Section 3, the experimental results are presented, the identification process is discussed, and the obtained simulation response is demonstrated. The final Section summarizes the key aspects of the performed research.

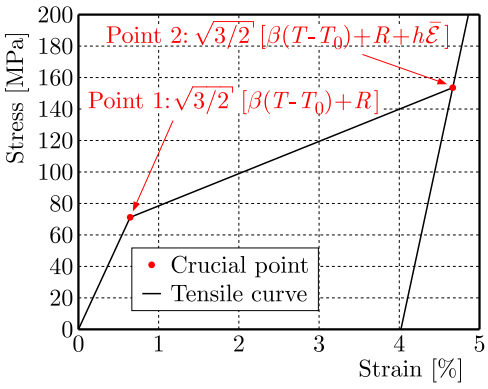
2. Materials and method

This paper discusses the identification process of NiTi alloy material properties based on conducted experimental research. The tests were carried out using a material supplied by Dynalloy (Dynalloy, Irvine, CA, USA), whose trade name is Flexinol. To perform the research, a low--temperature alloy (designation LT) characterized by full transformation at temperatures above 70°C was used (Dynalloy Inc., 2023). The crucial parameter of the investigated material is Young's modulus which depends on the crystal structure of the alloy. According to the manufacturer's information, it is 28 GPa for martensite and 75 GPa for austenite (Dynalloy Inc., 2023).

In order to identify the material model properties, the tensile curve of the NiTi alloy was determined. In most cases, investigations of the mechanical properties of NiTi alloys were carried out on normalized samples using universal testing systems (Hartl and Lagoudas, 2008; Pieczyska *et al.*, 2005). In the present paper, tensile tests were performed on thin NiTi wires using a dedicated automated test stand developed for this research. The used test stand was composed of: STAV 500/280 stand (AXIS Sp. z o.o., Gdańsk, Poland) for mounting the sample (using the designed handle) and measure the displacement, FB50 force gauge (AXIS Sp. z o.o., Gdańsk, Poland). The process of recording measurement data was performed by the cDAQ-9174 data acquisition system (NI, Austin, TX, USA) (Hartwich *et al.*, 2023). The test samples were thin nickel-titanium alloy wires with a diameter of 150 μm and an equal length of 150 mm.

The applied material model allows for simulation of the shape memory effect and generation of stress-strain characteristics of NiTi alloy. The selected material model was developed as a part of the work of Auricchio (2001) and was implemented in the Ansys environment (Ansys, 2023). A schematic curve which represents behavior of the considered material mode, along with parameters that allow defining its shape, are presented in Fig. 1.

(a)



(b)

Property	Symbol	Unit
Hardening parameter	h	GPa
Elastic limit	R	MPa
Maximum transformation strain	$\bar{\mathcal{E}}$	$\frac{1}{\text{mm}}$
Temperature scaling parameter	β	$\frac{\text{MPa}}{\text{°C}}$
Reference temperature	T_0	°C
Element temperature	T	°C

Fig. 1. Schematic tensile curve with crucial points marked

Identification of the material model properties is possible by mapping the experimentally determined stress-strain characteristics. Five tensile tests were carried out as a part of the research, and then the extreme results were dismissed. The experimental research consisted of stretching the mounted sample at a constant speed (50 mm/min) until a set strain of 4.9% was achieved, then the stand was returned to the initial position at the same speed. Experimental research was carried out at room temperature, in addition, measurements were made under stagnant air conditions in an unventilated room.

The numerical research has been conducted with the use of the FEM. The discretized numerical model was assembled from 21 beam elements with length of about 4.76 mm and diameter of 150 μm in the cross-section. All degrees of freedom were blocked for the node which was located at one end of the model. Stresses were induced by displacing the node located at the second end of the model in the longitudinal axis. For the model used to determine temperature versus strain characteristics similarly, all degrees of freedom were blocked for the end node. The node located at the second end was loaded with a constant force inducing stress equal to 172 MPa. At the same time the model was subjected to a variable temperature rising and falling within the range from 22°C to 100°C.

3. Results and discussion

The tensile curves of a thin NiTi alloy wire were determined for five samples of the wire with diameter of 150 μm. The obtained waveform is consistent with predictions based on theoretical knowledge (Mohd Jani *et al.*, 2014). It is possible to distinguish successive areas of stress-strain characteristics correlated with changes in the crystal structure of the NiTi alloy. The research results divided into different areas, together with the linear regression determined for them, is presented in Fig. 2.

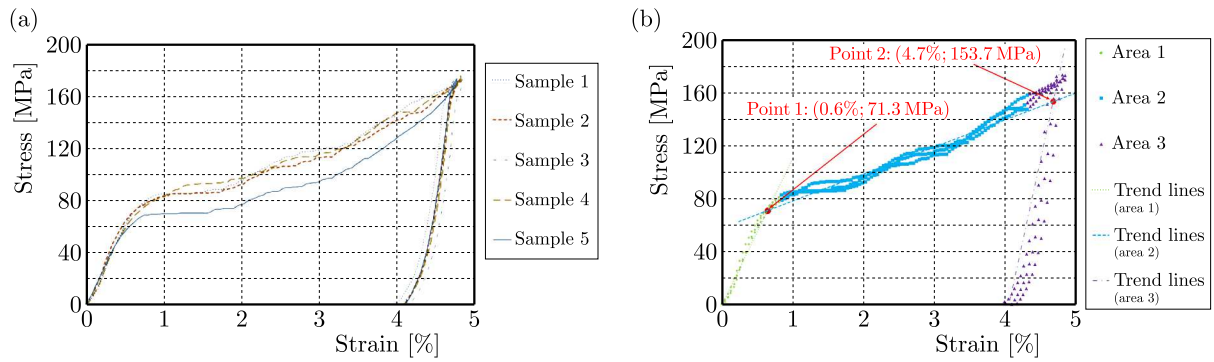


Fig. 2. (a) Experimental results, (b) the tensile result of the NiTi wire along with the trend lines determined for each loading stage

In the first stage of the loading, elastic stretching of martensite takes place, then followed by a phase transformation of the alloy – the crystalline structure changes from martensite to detwinned martensite at the strain range of approximately 0.6% to 4.6%. The last loading stage is associated with further deformation of the already completely detwinned martensite until the target stress is achieved. During the unloading, the stress gradually decreases to the initial position, however significant residual strain of approximately 4% is recorded. The crossing points of the trend lines are equivalent to the points separating the different areas of the characteristic. Point 1 (strain: 0.6%, stresses 71.3 MPa) defines the start of material transformation between martensite to detwinned martensite while point 2 (strain: 4.7%, stresses 153.7 MPa) determines the end of transformation. The stress-strain characteristics generated by the calibrated numerical

model compared to the experimental results and temperature-strain characteristics are presented in Figs. 3a and 3b, respectively.

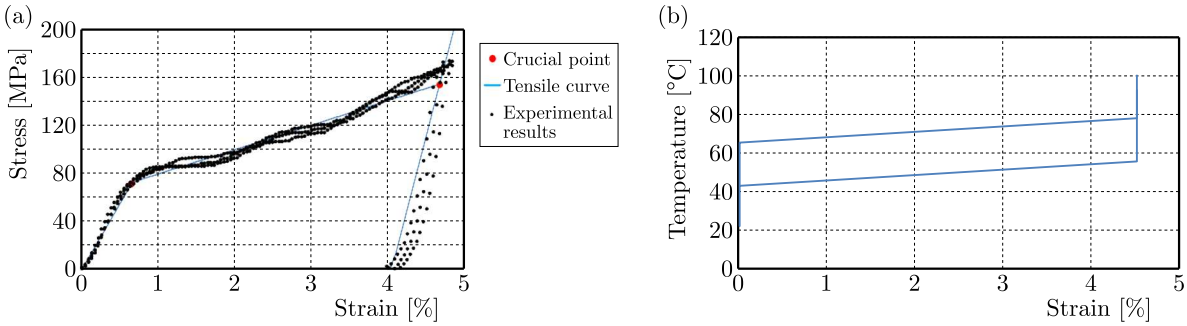


Fig. 3. (a) Stress-strain curves obtained during the tensile loading of five NiTi wire samples and the curve generated by the numerical model, (b) temperature to strain characteristics generated by the numerical model with 172 MPa maximal load

The tensile curve obtained from the model is a close reproduction of the experimental results, which is clearly visible in Fig. 3a. In addition, the model has been calibrated in such a way that the full transformation takes place in the temperature range from 40°C to 78°C along with the maximum strain of 4.5% according to the manufacturer's information (Dynalloy Inc., 2023). This can be seen in the temperature-to-strain characteristics (Fig. 3b).

4. Conclusions

On the basis of the conducted experimental investigation, it can be concluded that:

- the obtained results for successive tensile tests are similar to each other,
- the obtained results are consistent with the predicted ones based on theoretical knowledge of NiTi alloys. Therefore, different stages of stress-strain curves can be easily interpreted.

In addition, it can be claimed that the material model developed on the basis of experimental results:

- reproduces the experimental results with high accuracy,
- is consistent with the catalog data provided by the material manufacturer,
- has the potential to support further, more complex numerical research.

References

1. Abel E., Luo H., Pridham M., Slade A., 2004, Issues concerning the measurement of transformation temperatures of NiTi alloys, *Smart Materials and Structures*, **13**, 1110-1117
2. Ansys, 2023, *Shape Memory Alloy (SMA)*, https://ansyshelp.ansys.com/account/secured?returnurl=/Views/Secured/corp/v232/en/ans_mat/smas.html [accessed on 7 June 2023]
3. Antonucci V., Faiella G., Giordano M., Mennella F., Nicolais L., 2007, Electrical resistivity study and characterization during NiTi phase transformations, *Thermochimica Acta*, **462**, 64-69
4. Auricchio F., 2001, A robust integration-algorithm for a finite-strain shape-memory-alloy superelastic model, *International Journal of Plasticity*, **17**, 971-990
5. Chopra I., 1996, Review of current status of smart structures and integrated systems, *Proceedings of SPIE – The International Society for Optical Engineering*, **2717**, 20-62

6. Dynalloy Inc., 2023, *Flexinol® Nickel-Titanium Alloy Physical Properties*, https://www.dynalloy.com/pdfs/TCF1140.pdf [accessed on 19 May 2023]
7. Hartl D.J., Lagoudas D.C., 2008, Thermomechanical characterization of shape memory alloy materials, *Shape Memory Alloys*, 53-119
8. Hartwich J., Sławski S., Kciuk M., Duda S., 2023, Determination of thin NiTi wires' mechanical properties during phase transformations, *Sensors*, **23**, 1153
9. Kciuk M., Kuchcik W., Pilch Z., Klein W., 2019, A novel SMA drive based on the Graham Clock escapement and resistance feedback, *Sensors and Actuators A: Physical*, **285**, 406-413
10. Mohd Jani J., Leary M., Subic A., Gibson M.A., 2014, A review of shape memory alloy research, applications and opportunities, *Materials and Design*, **56**, 1078-1113
11. Pieczyska E., Gadaj S., Nowacki W.K., Hoshio K., Makino Y., Tobushi H., 2005, Characteristics of energy storage and dissipation in TiNi shape memory alloy, *Science and Technology of Advanced Materials*, **6**, 889-894
12. Pieczyska E.A., Tobushi H., Gadaj S.P., Nowacki W.K., 2006, Superelastic deformation behaviors based on phase transformation bands in TiNi shape memory alloy, *Materials Transactions*, **47**, 670-676
13. Sławski S., Kciuk M., Klein W., 2021, Assessment of SMA electrical resistance change during cyclic stretching with small elongation, *Sensors*, **21**, 6804
14. Sławski S., Kciuk M., Klein W., 2022, Change in electrical resistance of SMA (NiTi) wires during cyclic stretching, *Sensors*, **22**, 3584
15. Tobushi H., Pieczyska E., Ejiri Y., Sakuragi T., 2009, Thermomechanical properties of shape-memory alloy and polymer and their composites, *Mechanics of Advanced Materials and Structures*. **16**, 236-247

Manuscript received November 8, 2023; accepted for print February 29, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 421-434, Warsaw 2024
https://doi.org/10.15632/jtam-pl/186899

TRANSIENT VIBRATION OF A FRACTIONAL VISCOELASTIC CANTILEVER BEAM WITH AN ECCENTRIC MASS ELEMENT AT THE END[1]

Jan Freundlich
Warsaw Univerity of Thechnology, Warsaw, Poland
e-mail: jan.freundlich@pw.edu.pl

The work focuses on the transient forced vibration of a cantilever beam with a rigid eccentric mass element attached at the free end. The Euler-Bernoulli beam theory and the viscoelastic fractional Kelvin-Voigt material model are adopted. The equation of motion of the beam is derived using Hamilton's principle. The first eigenfunction of linear vibrations is used as an approximate solution for the nonlinear vibrations. The equations of motion of the system are solved numerically. The impact of the order of the fractional derivative on the beam transient linear and nonlinear vibrations is studied.

Keywords: fractional viscoelasticity, beam vibration, transient dynamic analysis, nonlinear vibrations

1. Introduction

Cantilever beams with a tip mass element are commonly used to model various engineering structures, such as tall buildings, offshore structures, moving cranes, masts, accelerometers, military airplane wings, accelerometers, Stockbridge dampers, energy harvesters, turbine blades (Rama Bhat and Wagner, 1976; Erturk and Inman, 2011; Gürgöze and Zeren, 2011; Markiewicz, 1995; Seidel and Csepregi, 1984). This issue has been studied extensively by many researchers for different variants of cantilever beams (Gürgöze and Zeren, 2011; Suzuki *et al.*, 2021; Yang, 2017). However, many studies have omitted some important issues, such as the effects of material damping and eccentricity on system dynamics. It is rather obvious that in some vibration studies of such beam systems it is necessary to take eccentricity into account, namely that the center of mass of the element does not coincide with its point of attachment to the end of the beam. This eccentricity can affect dynamic properties of the analyzed system (Gürgöze and Zeren, 2011; Suzuki *et al.*, 2021; Yang, 2017). Similarly, viscoelastic properties of the material may significantly affect the dynamic behavior of the system, thus a proper viscoelastic material model should be used in dynamic analysis. Some experiments revealed that numerous engineering materials show a weak frequency dependence of their damping properties within a wide frequency range (Torvik and Bagley, 1984; Caputo, 1967). The description of this feature is complicated, and it is usually performed with the help of integer order derivatives (Caputo, 1967).

In recent decades, fractional calculus has been increasingly used in many scientific researches. Fractional derivatives are widely utilized in mechanics of materials, control systems, mechatronics, thermoelasticity, signal and image processing, engineering biology and many other (Freundlich, 2016; Podlubny, 1999; Shen *et al.*, 2022; Sumelka, 2016; Sumelka *et al.*, 2020; Tayel and Hassan, 2019). Since, fractional derivatives are not local, they are used for modeling of non-local phenomena i.e. depending on the history process, therefore fractional derivatives are widely used in a description of viscoelastic material behavior (Torvik and Bagley, 1984; Rossikhin and

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

Shitikova, 2009). Fractional derivatives allows more accurate modeling of viscoelastic material behavior in a wide range of frequencies (Torvik and Bagley, 1984; Caputo, 1967).

This paper presents a study of transient vibrations of a fractional cantilever beam with a rigid mass element attached at its free end, whose center of mass does not coincide with the point of its attachment. The study is a continuation and extension of the author's earlier works (Freundlich, 2019, 2021). The mentioned works have been focused on vibration of a fractional viscoelastic cantilever beam with an end mass element, whose center of gravity coincides with the point of its attachment. In the first mentioned work, a fractional viscoelastic Kelvin-Voigt material model was used (Freundlich, 2019), whereas the second work adopted a fractional viscoelastic Zener material model (Freundlich, 2021). Therefore, this study is dedicated to transient dynamic analysis of a cantilever beam having the end mass element, whose center of gravity is not coincident with the point of its attachment.

2. Problem formulation

In this work, dynamic analysis of a homogeneous cantilever beam of length l having an eccentric heavy element of mass m_p and moment of inertia I_B, which is attached at the beam free end is presented. The mass center of the mass element does not coincide with the free end of the beam, and there is an eccentricity of distance e (Fig. 1). The analyzed beam has uniform cross-

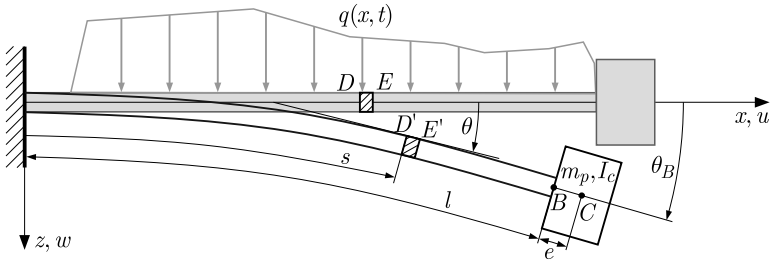


Fig. 1. Schematic of the analyzed beam

-section A and mass density ρ. The case of thin inextensible beam subjected to large deformation is studied. The Euler-Bernoulli theory is assumed, namely, that the rotary inertia and shear deformation are neglected. Moreover, the beam motion is assumed only in the xz-plane and that the gravitational force is perpendicular to this plane, thus the gravitational force has no effect on the beam motion. The viscoelastic beam material properties are assumed to be described using a fractional Kelvin-Voigt model, which is defined below (Torvik and Bagley, 1984; Mainardi and Spada, 2011)

$$\sigma(t) = E\big(\varepsilon(t) + \mu_\gamma D^{(\gamma)}(\varepsilon(t))\big) \tag{2.1}$$

where $\sigma(t)$ and $\varepsilon(t)$ are the stress and strain functions of time, μ_γ is a time constant (Mainardi and Spada, 2011), E is the relaxed modulus, t is time and $D^{(\gamma)}(\cdot)$ is the Caputo fractional derivative of the order γ, formulated as Eq. (2.2) (Caputo, 1967; Mainardi and Spada, 2011; Podlubny, 1999). For integer order derivative i.e. $\gamma = 1.0$, time constant μ_γ reduces to retardation time of the classical Kelvin-Voigt material. The unit of μ_γ is s^γ

$$D^{(\gamma)}(f(t)) \equiv \frac{d^\gamma}{dt^\gamma}(f(t)) \equiv \frac{1}{\Gamma(\mathfrak{M}-\gamma)} \int_0^t \frac{D^{(\mathfrak{M})}(f(\tau))}{(t-\tau)^{\gamma+1-\mathfrak{M}}}\, d\tau \tag{2.2}$$

where $\Gamma(\mathfrak{M}-\gamma)$ is the Euler gamma function (Podlubny, 1999), $D^{(\mathfrak{M})}(f(\cdot)) = (\partial^{\mathfrak{M}}/\partial t^{\mathfrak{M}})(\cdot)$ is the $\mathfrak{M}$-th derivative of a function $f(\cdot)$ with respect to time, $\mathfrak{M}$ is a positive integer number satisfying the inequality $\mathfrak{M}-1 < \gamma < \mathfrak{M}$, and $t > 0$.

In the case of dissipative forces, the value of γ is assumed to be in the range $0 < \gamma \leqslant 2$ (Malendowski and Sumelka, 2023), however for many real viscoelastic materials, the order of the fractional derivative is often assumed to be in the range $0 < \gamma \leqslant 1$ (Torvik and Bagley, 1984; Caputo, 1967) and then $\mathfrak{M} = 1$. Eq. (2.2), where $\gamma = 1.0$ is in the case of an integer order derivative (Mainardi and Spada, 2011; Podlubny, 1999).

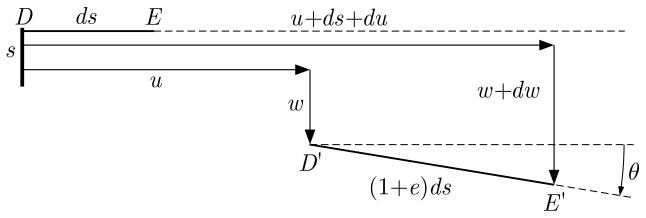


Fig. 2. Schematic of a beam displacements

The kinematics of the considered beam is presented in Fig. 2. From kinematic analysis it follows that

$$rd\theta = ds \;\rightarrow\; \frac{1}{r} = \kappa = \frac{\partial\theta}{\partial s} = \theta' \qquad \sin\theta = \frac{\partial w}{\partial s} = w' \;\Rightarrow\; \theta = \arcsin w' \tag{2.3}$$

where r is radius of curvature, κ is curvature.

Since the beam is assumed to be inextensible, therefore

$$\cos\theta = \frac{u + ds + du - u}{(1+\mathfrak{e})ds} = \frac{ds + du}{(1+\mathfrak{e})ds} = \frac{1+u'}{(1+\mathfrak{e})} \qquad \text{for} \qquad \mathfrak{e} = 0 \qquad \cos\theta = 1 + u' \tag{2.4}$$

then

$$\kappa = \theta' = \frac{\partial}{\partial s}(\arcsin w') = \frac{1}{\sqrt{1-(w')^2}} w'' \approx w''\Big(1 + \frac{1}{2}(w')^2\Big) \tag{2.5}$$

From the beam theory it follows that the strain is

$$\varepsilon(t) = -z\kappa = -zw''\Big(1 + \frac{1}{2}(w')^2\Big) \tag{2.6}$$

The extended Hamilton principle is utilized to obtain the equation of motion (Meirovitch, 1967)

$$\int_{t_1}^{t_2} (\delta T - \delta \Pi + \delta L) = 0 \tag{2.7}$$

where: δT and $\delta \Pi$ are variations of the system total kinetic and potential energy, respectively, δL is the total virtual work done by non-conservative forces.

The total kinetic energy of the system is the sum of kinetic energy of the beam and the end mass element. From literature it follows that the impact of beam longitudinal velocity u on the total system kinetic energy can be omitted, thus the total system kinetic energy is expressed as

$$T = \frac{1}{2}\int_0^l m\dot{w}^2 \, dx + \frac{1}{2} m_p \dot{w}_C^2 + \frac{1}{2} I_C (\dot{\theta}_C)^2 \tag{2.8}$$

where over-dots $\dot{(\cdot)}$ and $\ddot{(\cdot)}$ mean the first and second derivatives with respect to time, m is mass density per unit length, $\dot{w}$ is velocity of the neutral beam axis point in z direction, $\dot{w}_C$ is velocity of the center of mass C of the end mass element, $\dot{\theta}_C$ is angular velocity of the end mass element

(Fig. 1). Substituting the expression for time derivative of angle θ Eq. (2.3) into Eq. (2.8), the kinetic energy reads

$$T = \frac{1}{2}\int_0^l m\dot{w}^2\,dx + \frac{1}{2}m_p\Bigg[\bigg(-\frac{1}{\sqrt{1-(w'_B)^2}}\dot{w}'_B e\sin\theta_B\bigg)^2 + \bigg(\dot{w}_B + \frac{1}{\sqrt{1-(w'_B)^2}}\dot{w}'_B e\cos\theta_B\bigg)^2\Bigg] + \frac{1}{2}I_C\bigg(\frac{1}{\sqrt{1-(w'_B)^2}}\dot{w}'_B\bigg)^2 \tag{2.9}$$

Using an approximate relationship

$$\frac{1}{\sqrt{1-(w')^2}} \approx 1 + \frac{1}{2}(w')^2$$

the variation of the system kinetic energy is expressed as

$$\delta T = \delta\bigg(\frac{1}{2}\int_0^l m\dot{w}^2\,ds + \frac{1}{2}(m_p e^2 + I_C)(\dot{w}'_B)^2(1+(w'_B)^2) + \frac{1}{2}m_p(\dot{w}_B^2 + 2\dot{w}_B\dot{w}'_B e)\bigg) \tag{2.10}$$

From the assumed beam model it follows that the total potential energy is the strain energy of the beam. Utilizing relations (2.5) and (2.6), the variation of the beam strain energy can be expressed as

$$\begin{aligned} &\Pi_b = \frac{1}{2}\int_A\int_0^l E\varepsilon^2\,dA\,dx = \frac{1}{2}\int_A\int_0^l E(z)^2\kappa^2\,dA\,dx = \frac{1}{2}\int_0^l EJ\kappa^2\,dx \\ &\Rightarrow\quad \delta\Pi_b = \delta\bigg(\frac{1}{2}\int_0^l EJ\frac{(w'')^2}{1-(w')^2}\,dx\bigg) \approx \delta\bigg(\frac{1}{2}\int_0^l EJ[(w')^2 + (w'')^2(w')^2]\,dx\bigg) \end{aligned} \tag{2.11}$$

where E is Young's modulus of the beam material, A is cross-section area of the beam, J is cross-section moment of inertia with respect to the neutral beam axis.

Virtual work of non-conservative forces is a sum of work done by internal dissipation forces and external forces acting on the beam, namely

$$\begin{aligned} \delta L_{nc} &= -\int_0^l\int_A \sigma_{dis}\delta\varepsilon\,dA + \int_0^l q\delta w\,ds \\ &= -E'_\gamma J\int_0^l \frac{d^\gamma}{dt^\gamma}\Big[w''\Big(1+\frac{1}{2}(w')^2\Big)\Big]\Big(1+\frac{1}{2}(w')^2\Big)\delta(w'')\,ds \\ &\quad - E'_\gamma J\int_0^l \frac{d^\gamma}{dt^\gamma}\Big[w''\Big(1+\frac{1}{2}(w')^2\Big)\Big]w''w'\delta(w')\,ds + \int_0^l q\delta w\,ds \end{aligned} \tag{2.12}$$

Substituting the expanded and transformed expression for variations of kinetic energy (Eq. (2.10)), strain potential energy (Eq. (2.11)), and virtual work (Eq. (2.12)) into Hamilton's extended principle Eq. (2.7), the following equation of motion and boundary conditions of the analyzed system is obtained

$$\begin{aligned} &m\ddot{w} + EJw'''' + EJ[w''''(w')^2 + 6w''w'''w' + 3(w'')^3] \\ &\quad + E'_\gamma J\Big\{\frac{d^\gamma}{dt^\gamma}\Big[w''''\Big(1+\frac{1}{2}(w')^2\Big) + 3w''w'''w' + (w'')^3\Big]\Big(1+\frac{1}{2}(w')^2\Big)\Big\} \\ &\quad + E'_\gamma J\frac{d^\gamma}{dt^\gamma}\Big[w'''\Big(1+\frac{1}{2}(w')^2\Big) + (w'')^2w'\Big]w'w'' = q \end{aligned} \tag{2.13}$$

Boundary conditions are obtained directly from Hamilton's principle, and for $s = 0$, the beam deflection and slope equals 0, thus

$$w = w' = 0 \tag{2.14}$$

Whereas, the boundary conditions for $s = l$ are as follows

$$\begin{aligned}
&- m_p\ddot{w}_B - m_p\ddot{w}'_B e + EJ[w''' + w'''(w')^2 + 2w'(w'')^2 - (w'')^2 w'] \\
&\quad + E'_\gamma J\Big\{\frac{d^\gamma}{dt^\gamma}\Big[w'''\Big(1+\frac{1}{2}(w')^2\Big) + (w'')^2 w'\Big]\Big(1+\frac{1}{2}(w')^2\Big)\Big\} \\
&\quad + E'_\gamma J\Big\{\frac{d^\gamma}{dt^\gamma}\Big[w''\Big(1+\frac{1}{2}(w')^2\Big)\Big]w'w'' - \frac{d^\gamma}{dt^\gamma}\Big[w''\Big(1+\frac{1}{2}(w')^2\Big)\Big]w''w'\Big\} = 0 \\
&- (m_p e^2 + J_C)[\ddot{w}'_B(1+(w'_B)^2) + 2(\dot{w}'_B)^2 w'_B] - m_p\ddot{w}_B e - EJ(w'' + w''(w')^2) \\
&\quad - E'_\gamma J\Big\{\frac{d^\gamma}{dt^\gamma}\Big[w''\Big(1+\frac{1}{2}(w')^2\Big)\Big]\Big(1+\frac{1}{2}(w')^2\Big)\Big\} = 0
\end{aligned} \tag{2.15}$$

By introducing dimensionless parameters

$$\begin{aligned}
&\tau = \sqrt{\frac{EJ}{\rho A l^4}}t = ct \qquad \widetilde{x} = \frac{s}{l} \qquad \widetilde{w} = \frac{w}{l} \\
&\widetilde{\mu}_\gamma = \mu_\gamma\sqrt{\Big(\frac{EJ}{\rho A l^4}\Big)^\gamma} = \mu_\gamma c^\gamma \qquad \widetilde{q} = \frac{ql^3}{EJ} \qquad \alpha = \frac{m_p}{\rho A l} \\
&\beta = \frac{I_C}{\rho A l^3} \qquad z = kl \qquad \eta = \frac{e}{l}
\end{aligned} \tag{2.16}$$

and substituting them into Eq. (2.13), the following dimensionless equation of motion is obtained

$$\begin{aligned}
&\frac{\partial^2\widetilde{w}}{\partial\tau^2} + \frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4} + \frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}\Big(\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\Big)^2 + 6\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\frac{\partial^3\widetilde{w}}{\partial\widetilde{x}^3}\frac{\partial\widetilde{w}}{\partial\widetilde{x}} + 3\Big(\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Big)^3 + \widetilde{\mu}_\gamma\Bigg\{\frac{d^\gamma}{d\tau^\gamma}\Bigg[\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\Big)^2\Bigg) \\
&\quad + 3\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\frac{\partial^3\widetilde{w}}{\partial\widetilde{x}^3}\frac{\partial\widetilde{w}}{\partial\widetilde{x}} + \Big(\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Big)^2\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Bigg]\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}'}{\partial\widetilde{x}}\Big)^2\Bigg)\Bigg\} \\
&\quad + \widetilde{\mu}_\gamma\Bigg\{\frac{d^\gamma}{d\tau^\gamma}\Bigg[\frac{\partial^3\widetilde{w}}{\partial\widetilde{x}^3}\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}'}{\partial\widetilde{x}}\Big)^2\Bigg) + \Big(\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Big)^2\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\Bigg]\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Bigg\} = \widetilde{q}
\end{aligned} \tag{2.17}$$

with dimensionless boundary conditions for $\widetilde{x} = 0$, $\widetilde{w} = \widetilde{w}' = 0$ and for $\widetilde{x} = 1$

$$\begin{aligned}
&- \alpha\Big(\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\tau^2} + \eta\frac{\partial^3\widetilde{w}(1,\tau)}{\partial\tau^2\partial\widetilde{x}}\Big) + \frac{\partial^3\widetilde{w}(1,\tau)}{\partial\widetilde{x}^3} + \frac{\partial^3\widetilde{w}(1,\tau)}{\partial\widetilde{x}^3}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2 + \frac{\partial\widetilde{w}}{\partial\widetilde{x}}\Big(\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}\Big)^2 \\
&\quad + \widetilde{\mu}_\gamma\Bigg\{\frac{d^\gamma}{dt^\gamma}\Bigg[\frac{\partial^3\widetilde{w}(1,\tau)}{\partial\widetilde{x}^3}\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg)\Bigg]\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg)\Bigg\} \\
&\quad - \widetilde{\mu}_\gamma\Bigg\{\frac{d^\gamma}{dt^\gamma}\Bigg[\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg)\Bigg]\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Bigg\} = 0 \\
&- (\alpha\eta^2 + \beta)\Bigg[\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\tau^2}\Bigg(1+\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg) + 2\Big(\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\tau\partial\widetilde{x}}\Big)^2\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Bigg] \\
&\quad - \Bigg[\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2} + \frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg] \\
&\quad - \mu_\gamma\Bigg\{\frac{d^\gamma}{dt^\gamma}\Bigg[\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg)\Bigg]\Bigg(1+\frac{1}{2}\Big(\frac{\partial\widetilde{w}(1,\tau)}{\partial\widetilde{x}}\Big)^2\Bigg)\Bigg\} = 0
\end{aligned} \tag{2.18}$$

The approximate solution is assumed in the form of the first eigenfunction of linearized nonlinear equation Eq. (2.17). This assumption can be made because the dynamic behavior of the analyzed beam is studied in the vicinity of the first resonance, and the applied load causes only first mode vibrations. Moreover, from the structure of Green's functions of characteristic equations (see Freundlich, 2019; Podlubny, 1999) and the fact that the first natural frequency is several times lower then the second natural frequency of the analyzed cantilever beam, it follows that the first eigenfunction has the greatest impact on the vibration amplitude. Therefore, in the first step, the equation of motion is simplified, namely, only expressions up to the third order are considered. Grouping linear and nonlinear terms, we obtain

$$\begin{aligned}&\frac{\partial^2\widetilde{w}}{\partial\tau^2}+\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}+\widetilde{\mu}_\gamma\frac{d^\gamma}{d\tau^\gamma}\Big(\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}\Big)+\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}\Big(\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\Big)^2+6\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\frac{\partial^3\widetilde{w}}{\partial\widetilde{x}^3}\frac{\partial\widetilde{w}}{\partial\widetilde{x}}+3\Big(\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Big)^3\\&\quad+\widetilde{\mu}_\gamma\Big\{\frac{d^\gamma}{d\tau^\gamma}\Big[\frac{1}{2}\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}\Big(\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\Big)^2+3\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\frac{\partial^3\widetilde{w}}{\partial\widetilde{x}^3}\frac{\partial\widetilde{w}}{\partial\widetilde{x}}+\Big(\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Big)^3\Big]+\frac{d^\gamma}{d\tau^\gamma}\Big(\frac{\partial^3\widetilde{w}}{\partial\widetilde{x}^3}\Big)\frac{\partial\widetilde{w}}{\partial\widetilde{x}}\frac{\partial^2\widetilde{w}}{\partial\widetilde{x}^2}\Big\}=\widetilde{q}\end{aligned}\tag{2.19}$$

Using the first mode approximation, namely $\widetilde{w}(\widetilde{x},\tau)=W_1(\widetilde{x})\xi_1(\tau)$

$$\begin{aligned}&W_1D^{(2)}(\xi_1)+W_1''''\xi_1+\widetilde{\mu}_\gamma W_1''''D^{(\gamma)}(\xi_1)+W_1''''\xi_1(W_1'\xi_1)^2+6W_1''\xi_1W_1'''\xi_1W_1'\xi_1\\&\quad+3(W_1''\xi_1)^3+\widetilde{\mu}_\gamma\Big[D^{(\gamma)}\Big(\frac{1}{2}W_1''''\xi_1(W_1'\xi_1)^2+3W_1''\xi_1W_1'''\xi_1W_1'\xi_1+(W_1''\xi_1)^3\Big)\\&\quad+D^{(\gamma)}(W_1'''\xi_1)W_1'\xi_1W_1''\xi_1\Big]=\widetilde{q}\end{aligned}\tag{2.20}$$

where W_1 is the first eigenfunction of linearized nonlinear equation (2.17).

Next, multiplying both sides by W_1, substituting the relationship $W''''(\widetilde{x})=\widehat{k}^4W(\widetilde{x})$ (Eq. (2.27)), and integrating from 0 to 1, we obtain an equation of the generalized coordinate

$$D^{(2)}(\xi_1)+\widehat{k}^4\xi_1+\widetilde{\mu}_\gamma\widehat{k}^4D^{(\gamma)}(\xi_1)+a_3\xi_1^3+b_3\widetilde{\mu}_\gamma D^{(\gamma)}(\xi_1^3)+b_2\widetilde{\mu}_\gamma D^{(\gamma)}(\xi_1)\xi_1^2=Q\tag{2.21}$$

where

$$\begin{aligned}a_3&=\frac{\int_0^1\widehat{k}^4W_1^2(W_1')^2\,d\widetilde{x}+6\int_0^1W_1''W_1'''W_1'W_1\,d\widetilde{x}+3\int_0^1(W_1'')^3W_1\,d\widetilde{x}}{\int_0^1W_1^2\,d\widetilde{x}}\\b_2&=\frac{\int_0^1W_1'''W_1'W_1''W_1\,d\widetilde{x}}{\inf_0^1W_1^2\,d\widetilde{x}}\qquad Q=\frac{\int_0^1\widetilde{q}(\widetilde{x},\tau)W_1\,d\widetilde{x}}{\int_0^1W_1^2\,d\widetilde{x}}\\b_3&=\frac{\int_0^1\frac{1}{2}\widehat{k}^4W_1^2(W_1')^2\,d\widetilde{x}+3\int_0^1W_1''W_1'''W_1'W_1\,d\widetilde{x}+\int_0^1(W_1'')^3W_1\,d\widetilde{x}}{\int_0^1W_1^2\,d\widetilde{x}}\end{aligned}\tag{2.22}$$

As was mentioned earlier, the approximate solution to Eq. (2.20) is in the form of linear modes of linearized Eq. (2.19). This linearized equation has a form

$$\frac{\partial^2\widetilde{w}}{\partial\tau^2}+\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}+\mu_\gamma\frac{d^\gamma}{d\tau^\gamma}\Big(\frac{\partial^4\widetilde{w}}{\partial\widetilde{x}^4}\Big)=\widetilde{q}\tag{2.23}$$

with linearized boundary conditions, namely, for the clamped beam end $\widetilde{x}=0$

$$\widetilde{w}(0,\tau)=\frac{\partial\widetilde{w}(0,\tau)}{\partial\widetilde{x}}=0\tag{2.24}$$

and for $\tilde{x}=1$

$$\begin{aligned}&\alpha\Big(\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\tau^2}+\eta\frac{\partial^3\widetilde{w}(1,\tau)}{\partial\tau^2\partial\widetilde{x}}\Big)-\Big(\frac{\partial^3\widetilde{w}(1,\tau)}{\partial\widetilde{x}^3}+\widetilde{\mu}_\gamma\frac{d^\gamma}{d\tau^\gamma}\frac{\partial^3\widetilde{w}(1,\tau)}{\partial\widetilde{x}^3}\Big)=0\\&\alpha\eta\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\tau^2}+(\alpha\eta^2+\beta)\frac{\partial^3\widetilde{w}(1,\tau)}{\partial\tau^2\partial\widetilde{x}}+\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}+\widetilde{\mu}_\gamma\frac{d^\gamma}{d\tau^\gamma}\frac{\partial^2\widetilde{w}(1,\tau)}{\partial\widetilde{x}^2}=0\end{aligned}\tag{2.25}$$

The solution to the problem formulated by Eqs. (2.23)-(2.25) is sought in the form of a convergent series of the dimensionless beam eigenfunctions

$$\widetilde{w}(\widetilde{x},\tau)=\sum_{n=1}^{\infty}\xi_n(\tau)W_n(\widetilde{x}) \tag{2.26}$$

where $W_n(\widetilde{x})$ is the n-th eigenfunction of the beam, $\xi_n(\tau)$ is the n-th time depending generalized coordinate (Meirovitch, 1967).

The functions $W_n(\widetilde{x})$ can be determined with the help of a well-known procedure, i.e. by solving Eq. (2.23) with its right-hand side equal to zero (homogeneous equation). Namely, utilizing separation of variables, the subsequent equation may be obtained

$$W''''(\widetilde{x})-\widehat{k}^4W(\widetilde{x})=0 \tag{2.27}$$

The solution to equation above (2.27) is sought as

$$W(\widetilde{x})=A\sin(\widehat{k}\widetilde{x})+B\cos(\widehat{k}\widetilde{x})+C\sinh(\widehat{k}\widetilde{x})+D\cosh(\widehat{k}\widetilde{x}) \tag{2.28}$$

where A, B, C, D are arbitrary unknown constants.

Using the first two boundary conditions Eq. (2.24), the following relations between the constants may be found, namely, $A=-C$ and $B=-D$. Then, using these relationships and the next two boundary conditions Eq. (2.25), the following system of equations for constants A and B is derived

$$\begin{aligned}
&A[\alpha\widehat{k}(\sin\widehat{k}-\sinh\widehat{k})+\alpha\widehat{k}^2\eta(\cos\widehat{k}-\cosh\widehat{k})-(\cos\widehat{k}+\cosh\widehat{k})]\\
&\quad+B[\alpha\widehat{k}(\cos\widehat{k}-\cosh\widehat{k})-\alpha\widehat{k}^2\eta(\sin\widehat{k}+\sinh\widehat{k})-(\sin\widehat{k}-\sinh\widehat{k})]=0\\
&A[\alpha\eta\widehat{k}^2(\sin\widehat{k}-\sinh\widehat{k})+(\alpha\eta^2+\beta)\widehat{k}^3(\cos\widehat{k}-\cosh\widehat{k})+\sin\widehat{k}+\sinh\widehat{k}]\\
&\quad+B[\alpha\eta\widehat{k}^2(\cos\widehat{k}-\cosh\widehat{k})-(\alpha\eta^2+\beta)\widehat{k}^3(\sin\widehat{k}+\sinh\widehat{k})+\cos\widehat{k}+\cosh\widehat{k}]=0
\end{aligned} \tag{2.29}$$

The system of equations presented above, Eq. (2.29), is satisfied if the determinant of the coefficients matrix of the system of equation equals zero. Then, equating the determinant of (2.29) to zero, after long and arduous mathematical transformations, the characteristic equation of the system can be obtained

$$\begin{aligned}
&-\widehat{k}^4\alpha\beta(1-\cos\widehat{k}\cosh\widehat{k})+\widehat{k}^3(\beta+\alpha\eta^2)(\cos\widehat{k}\sinh\widehat{k}+\sin\widehat{k}\cosh\widehat{k})\\
&\quad+2\alpha\eta\widehat{k}^2\sin\widehat{k}\sinh\widehat{k}-\alpha(\cos\widehat{k}\cosh\widehat{k}-\sin\widehat{k}\cosh\widehat{k})-\cos\widehat{k}\cosh\widehat{k}-1=0
\end{aligned} \tag{2.30}$$

Characteristic equation (2.30) has of a countable infinitive set of roots $\widehat{k}_n$ corresponding to the n-th natural undamped dimensionless frequency of the beam. Next, substituting the derived roots into Eqs. (2.27) and (2.29), the expression for eigenfunctions can be obtained

$$W_n(\widetilde{x})=A_n[\sin(\widehat{k}_n\widetilde{x})-\sinh(\widehat{k}_n\widetilde{x})-\lambda_n(\cos(\widehat{k}_n\widetilde{x})-\cosh(\widehat{k}_n\widetilde{x}))] \tag{2.31}$$

where

$$\lambda_n=\frac{\alpha\widehat{k}_n(\sin\widehat{k}_n-\sinh\widehat{k}_n)+\alpha\widehat{k}_n^2\eta(\cos\widehat{k}_n-\cosh\widehat{k}_n)-(\cos\widehat{k}_n+\cosh\widehat{k}_n)}{\alpha\widehat{k}_n(\cos\widehat{k}_n-\cosh\widehat{k}_n)-\alpha\widehat{k}_n^2\eta(\sin\widehat{k}_n+\sinh\widehat{k}_n)-(\sin\widehat{k}_n-\sinh\widehat{k}_n)}$$

Eigenfunctions (2.31) must satisfy the orthogonality condition. Using the well-known standard procedure (see e.g. Meirovitch, 1967), it can be shown that the orthogonality condition has a following form

$$\begin{aligned}
&\int_0^1 W_m(\widetilde{x})W_n(\widetilde{x})\,d\widetilde{x}+\alpha[W_n(1)W_m(1)+\eta W_n'(1)W_m(1)]\\
&\quad+\alpha\eta W_n(1)W_m'(1)+(\alpha\eta^2+\beta)W_n'(1)W_m'(1)=\delta_{mn}
\end{aligned} \tag{2.32}$$

Employing orthogonality condition Eq. (2.31) and expression for eigenfunction, Eq. (2.31), coefficients A_n in Eq. (2.31) can be calculated as

$$A_n = \frac{1}{\sqrt{\int_0^1 \widetilde{W}_n^2(\widetilde{x}); d\widetilde{x} + \alpha[\widetilde{W}_n^2(1) + 2\eta\widetilde{W}_n'(1)\widetilde{W}_n(1)] + (\alpha\eta^2 + \beta)\widetilde{W}_n'^2(1)}} \tag{2.33}$$

Therefore, the function W_1 in Eqs. (2.20) and (2.22) is determined, thus the approximate solution to the problem described by Eq. (2.19) may be obtained.

Fractional differential equation (2.21) can be solved numerically using a method similar to the method presented in the book by Diethelm (2010). In this method, the fractional differential equation is converted to a system of mixed ordinary and fractional differential equations, each of the order $0 < \gamma \leqslant 1$. The converted system of equations contains integer and fractional order differential equations, which can be partitioned into two separated systems of equations and solved simultaneously (Freundlich, 2021). The system of equations is solved using own author's procedure implemented in the Matlab package. The fractional order differential equations are integrated using the trapezoidal rule for the fractional Caputo derivative worked out by Diethelm *et al.* (2005), while the integer order equations are integrated using the Adams-Bashforth-Moulton predictor-corrector method (see e.g Chapra and Canale, 2010). Roots of the nonlinear characteristic equation of system (2.30) are computed using Matlab procedure "fzero". The knowledge of damped natural frequencies are useful in dynamic analysis of the system. The natural damped frequencies of linearized sytem (2.23) can be calculated solving the characteristic equation associated with linearized fractional differential equation (2.21) with the zero right hand side, namely

$$s_n^2 + \widetilde{\mu}_\gamma \widehat{k}^4 s_n^\gamma + \widehat{k}^4 s = 0 \tag{2.34}$$

Characteristic equation (2.34) has two conjugate complex roots located in the left half-plane of the complex domain (Rossikhin and Shitikova, 1997). The absolute value of the real part of the root is the damping coefficient, whereas the imaginary part of the root is the natural damped frequency (Rossikhin and Shitikova, 1997). Equation (2.34) is solved using author's own procedure based on Newton's method of solving nonlinear complex equations (Chapra and Canale, 2010).

3. Example of numerical calculations and discussion

To demonstrate the usefulness of the method presented in the previous Section, examplary calculations of transient vibrations of the analyzed beam have been performed. The relationships obtained in the preceding Section are used to study the impact of the fractional derivative order and other parameters of the system on its transient vibrations. Additionally, responses of linear and nonlinear systems are studied and compared. Since it is important to know the modal damping and damped natural frequencies in the analysis of system dynamics, the effect of the order of the fractional derivative on the damping coefficient and damped natural frequency of the system is first determined. As mentioned previously, the damping coefficient and natural damped frequency of the system are determined by real and imaginary parts of the roots of Eq. (2.34), respectively. Numerical calculations are performed for various orders of the fractional derivative and for beam parameters $\alpha = 1$, $\beta = 0.005$, $\eta = 0.05$ and for two damping coefficients, $\widetilde{\mu}_\gamma = 0.008$ and 0.016. Computed relationships between the calculated roots and the order of the fractional derivative are shown in Figs. 3 and 4.

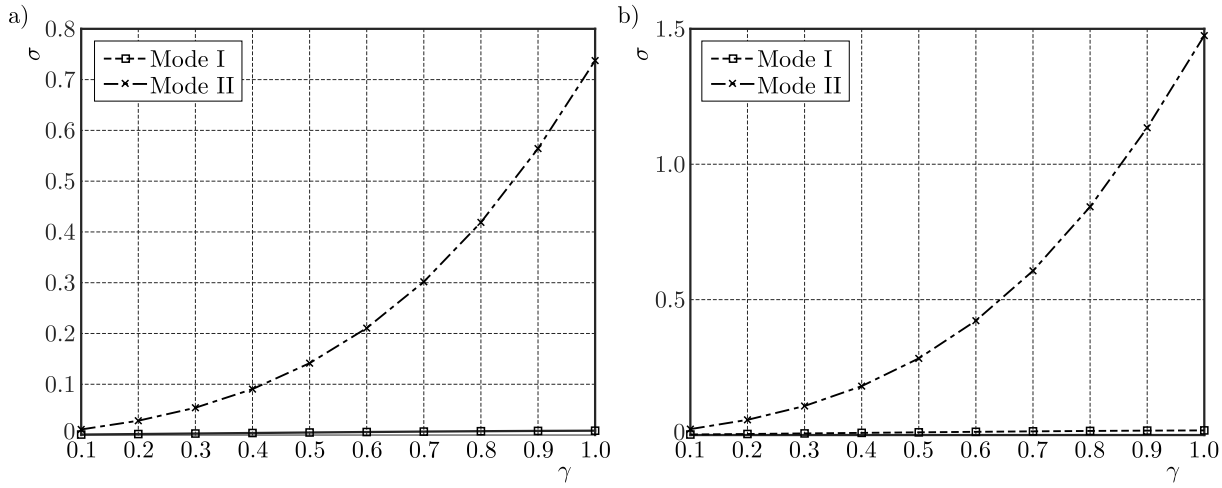


Fig. 3. The real part of the roots of characteristic equation (2.34), $\alpha = 1$, $\beta = 0.005$, $\eta = 0.05$: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

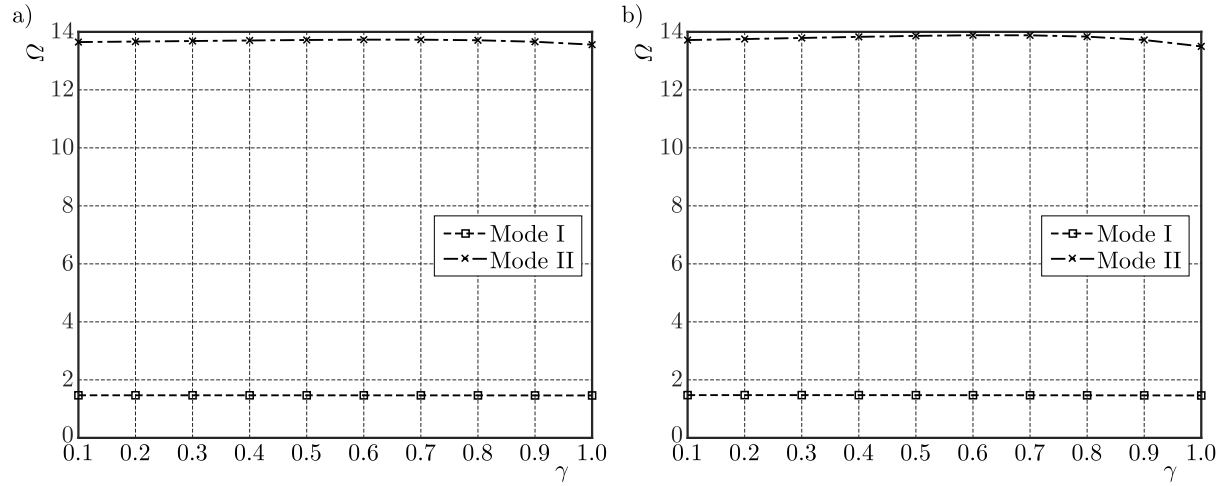


Fig. 4. The imaginary part of the roots of characteristic equation (2.34), $\alpha = 1$, $\beta = 0.005$, $\eta = 0.05$: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

It can be noticed from Fig. 3 that the damping coefficient exponentially increases with an increase of the order of the fractional derivative. The increase is significantly greater for the second mode of vibration. On the contrary, the damped natural frequency practically does not depend on the change of the order of the fractional derivative (see Fig. 4).

As noted before, in some vibration studies of the beam with attached at its end a heavy mass element, it is necessary to take into account the eccentricity. Therefore, sample calculations showing the effect of eccentricity on the damping coefficient and natural damped frequency are made. The calculations are made for various η coefficients, for $\gamma = 0.5$, $\alpha = 1$, $\beta = 0.005$, and for two damping coefficients, $\widetilde{\mu}_\gamma = 0.008$ and 0.016. An impact of the eccentricity coefficient η on the damping coefficient and damped natural frequency is shown in Figs. 5 and 6.

As can be seen from Figs. 4 and 6, an increase in the order of the fractional derivative results in a decrease of damping coefficients and natural damped frequencies. The decrease is greater for the second mode of vibrations. A relative difference between damped natural frequencies for $\eta = 0$ and $\eta = 0.2$ is about 20% for the fist mode of vibration, and about 24% for the second mode of vibration.

Next, having determined damped natural frequencies of the linearized system, the impact of the order of the fractional derivative on transient forced vibrations of the analyzed beam is

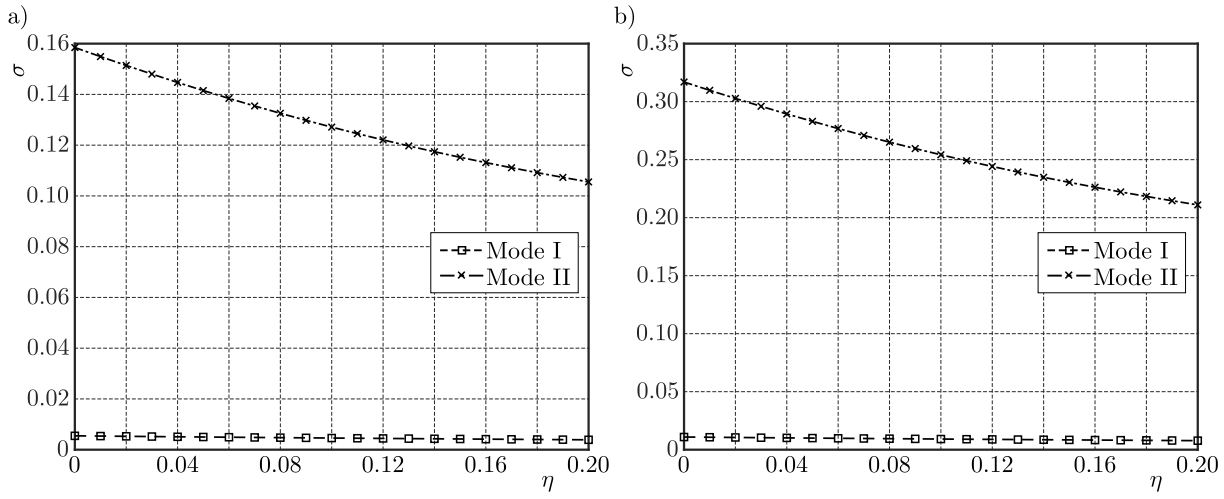


Fig. 5. The real part of the roots of characteristic equation (2.34), $\gamma = 0.5$, $\alpha = 1$, $\beta = 0.005$: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

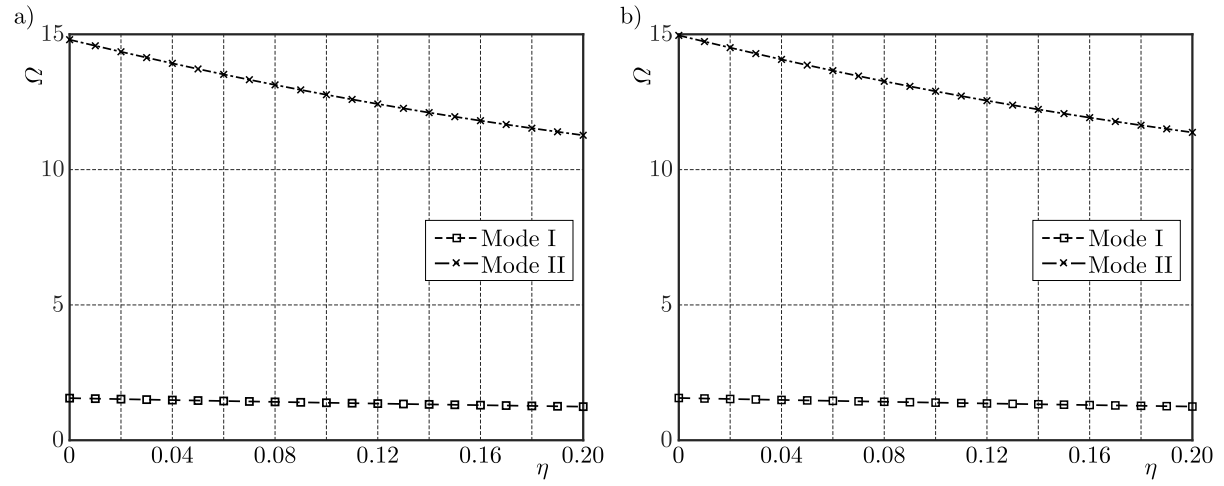


Fig. 6. The imaginary part of the roots of characteristic equation (2.34), $\gamma = 0.5$, $\alpha = 1$, $\beta = 0.005$: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

investigated. Linear and nonlinear transient vibrations are examined. In the first stage, linear and nonlinear beam responses to the harmonic excitation of amplitude F_0 are computed. The excitation frequency is assumed to be the natural damped frequency of the linearized system determined earlier (see Fig.4). These calculations are performed for the dimensionless beam parameters $\alpha = 1$, $\beta = 0.005$, $\eta = 0.05$, two damping coefficients, $\widetilde{\mu}_\gamma = 0.008$ and 0.016, and various orders of the fractional derivative $\gamma = 0.25$, 0.5, 0.75, 1.0. The calculated responses of the linearized system to the harmonic excitation are shown in Fig. 7, whereas for the nonlinear system are shown in Fig. 8. As can be seen from Fig. 7, the maximum amplitudes of the linearized responses increase monotonically until their values stabilize. Furthermore, vibration amplitudes are greater for lower values of the order of the fractional derivative γ for both coefficients $\widetilde{\mu}_\gamma$ (Fig. 7). In contrast, the maximum amplitudes of the nonlinear responses oscillate until their values stabilize (see Fig. 8). Furthermore, it can be seen from Fig. 8 that the oscilations of the maximum amplidudes of nonlinear responses are greater for lower values of the order of the fractional derivative γ. Similarly, as in the case of linear responses, the vibration amplitudes are lower as the order of the fractional derivative γ increases.

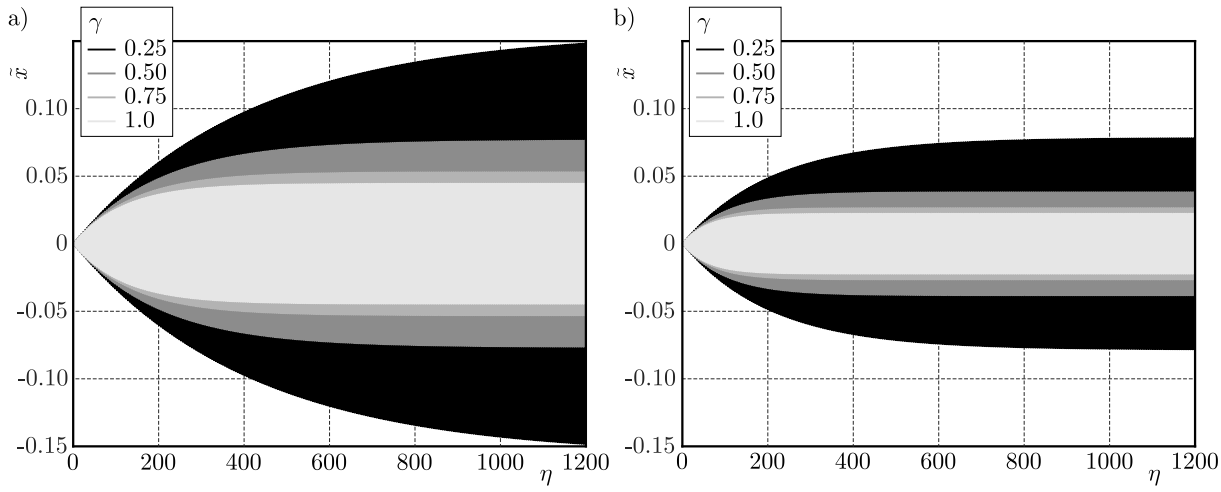


Fig. 7. Linear beam response, harmonic excitation: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

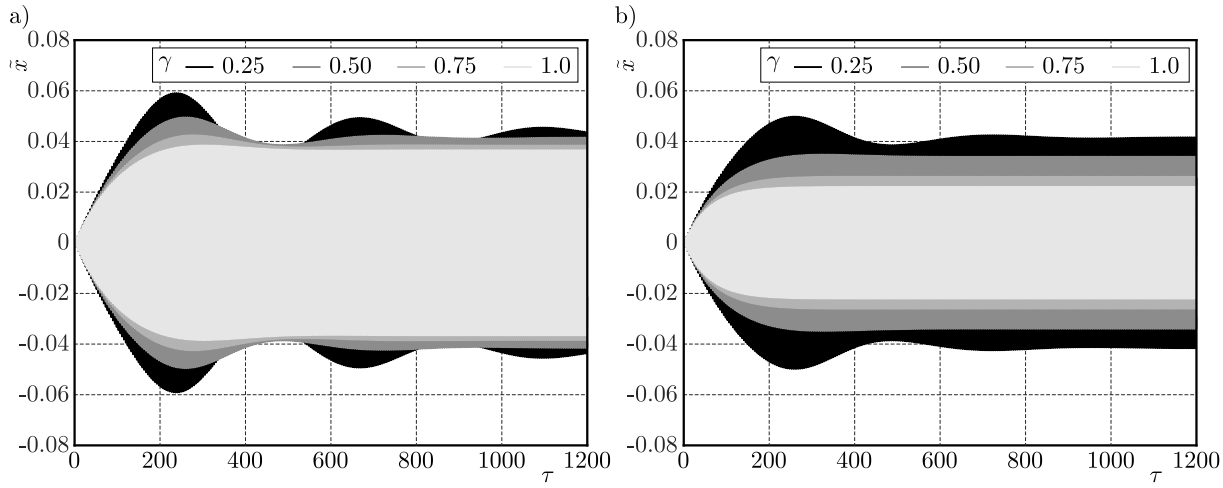


Fig. 8. Nonlinear beam response, harmonic excitation: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

In the next step, the transient responses of the beam to an excitation force of varying angular frequency are studied. The excitation force function is described by the following expression

$$F(\tau) = F_0 \sin \frac{\mathcal{E}\tau^2}{2} \tag{3.1}$$

where $\mathcal{E}$ is dimensionless angular acceleration.

The beam responses to excitation described by Eq. (3.1) are computed for the dimensionless angular acceleration $\mathcal{E} = 0.1$, dimensionless beam parameters $\gamma = 0.5$, $\alpha = 1$, $\beta = 0.005$, $\eta = 0.05$, two damping coefficients, $\widetilde{\mu}_\gamma = 0.008$ and 0.016, and the order of the fractional derivative $\gamma = 0.25$, 0.5, 0.75, 1.0. The calculated responses are shown in Fig. 9. The obtained responses of the beam show that the maximum amplitudes of vibrations, after reaching the maximum value, decrease monotonically. The decrease is faster for higher orders of the fractional derivative γ and greater coefficient $\widetilde{\mu}_\gamma$. As can be seen from Fig. 9, an increase of the order of the fractional derivative decreases the response amplitudes.

Analyzing the results shown in Figs. 7-9, it can be concluded that the order of the fractional derivative has a similar effect on vibration amplitudes as the damping coefficient or the time constant μ_γ, i.e., increasing the order of the fractional derivative γ causes a decrease in the vibration amplitudes.

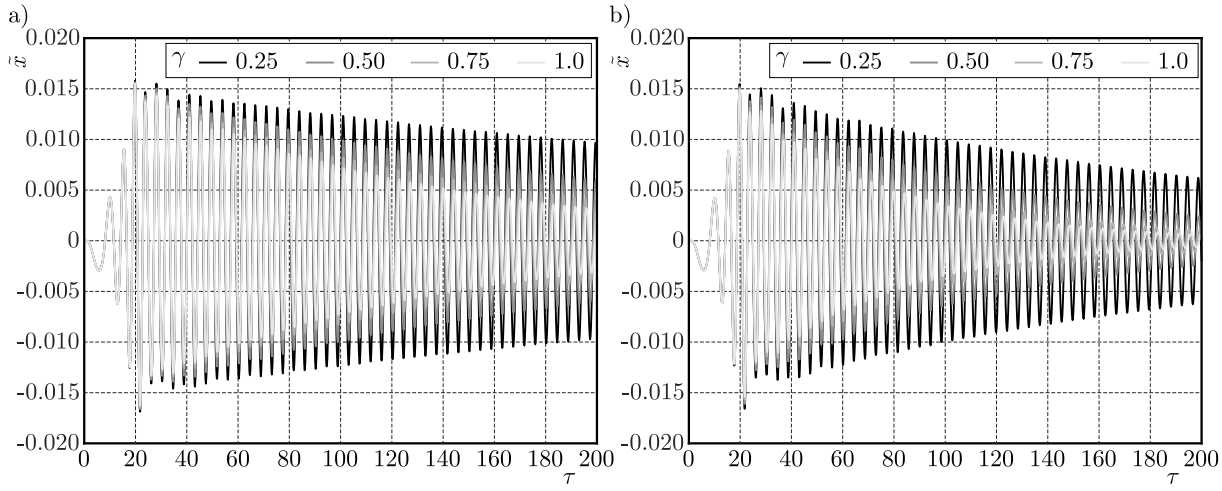


Fig. 9. Nonlinear beam response, transient excitation, $\mathcal{E} = 0.1$; (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

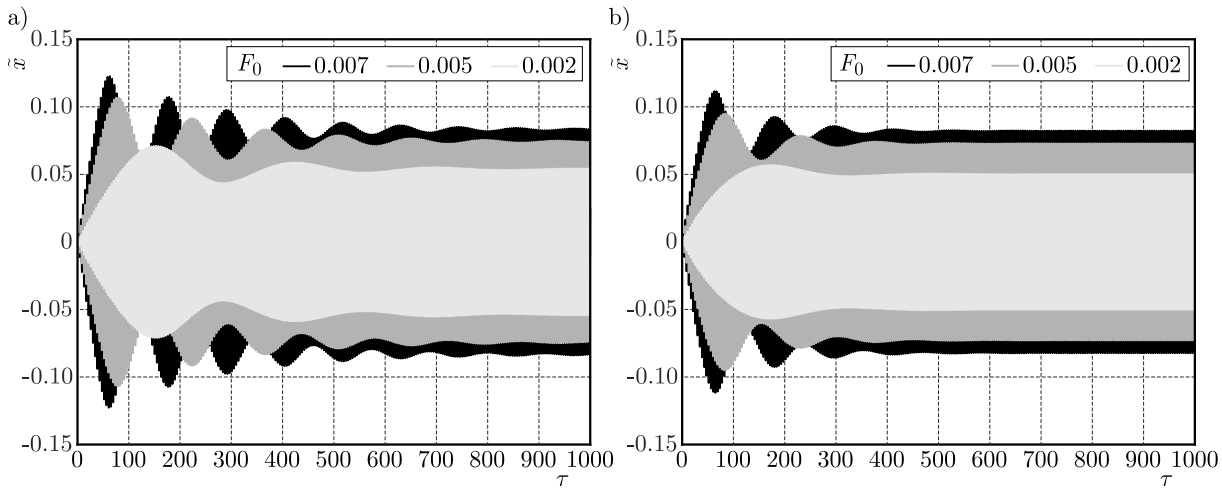


Fig. 10. Nonlinear beam response, harmonic excitation, $\gamma = 0.5$: (a) $\widetilde{\mu}_\gamma = 0.008$, (b) $\widetilde{\mu}_\gamma = 0.016$

Finally, the effect of amplitude F_0 of the sinusoidal forcing force on the transient responses of the beam is studied. The study is carried out for the order of the fractional derivative $\gamma = 0.5$, $\widetilde{\mu}_\gamma = 0.008$, 0.016, and amplitudes of the exciting force $F_0 = 0.002$, 0.005, 0.007. The computed nonlinear responses are presented in Fig. 10.

From Fig. 10 we can see that the oscillation of the maximum amplitude of the response is higher for higher amplitudes of the exciting force. Additionally, the responses reach the steady--state amplitudes after a longer time period for higher forcing amplitudes F_0.

4. Conclusions

In this paper, linear and nonlinear transient vibrations of a fractional cantilever beam with an attached eccentric mass element are presented. The fractional Kelvin-Voigt viscoelastic material model is assumed as the beam material. Nonlinear and linear equations of motion of the beam are derived using Hamilton's principle. The characteristic equation, modal frequencies, eigenfunction and orthogonality condition are obtained for linear beam vibrations. The achieved equations of motion are solved numerically. Numerical calculations are carried out for selected beam parameters. Transient responses of the beam to the harmonic and linearly time-varying

increasing frequency of a sinusoidal excitation are calculated. The beam responses to the harmonic excitation are calculated for linear and nonlinear equations of motion. Comparing the determined linear and nonlinear responses, it can be seen that the maximum amplitudes of the linear responses increase monotonically until their values stabilize, whereas the maximum amplitudes of the nonlinear responses oscillate until their values stabilize.

The obtained nonlinear responses to time-varying frequency of the sinusoidal excitation reveal that the maximum vibration amplitudes decrease monotonically after reaching their maximum value. The decrease is faster for higher orders of the fractional derivative and greater dimensionless damping coefficients.

For all obtained results, it can be stated that the maximum amplitudes of vibrations decrease as the order of the fractional derivative increases in all performed calculations, which was expected.

The carried out researches show that he effect of eccentricity on natural frequencies is approximately linear. Thus, in my opinion, the eccentricity should be taken into account in some calculations if η is greater than 0.1.

In further investigations, actual parameters of the fractional Kelvin-Voigt model corresponding to the system analyzed should be determined by conducting appropriate experimental examinations.

References

1. Caputo M., 1967, Linear models of dissipation whose Q is almost frequency independent-II, *Geophysical Journal of Royal Astronomical Society*, **13**, 529-539
2. Chapra S.C., Canale R.P., 2010, *Numerical Methods for Engineers*, McGraw Hill, Boston
3. Diethelm K., 2010, *The Analysis of Fractional Differential Equations*, Lecture Notes in Mathematics, Vol. 2004, Springer, Berlin
4. Diethelm K., Ford N.J., Freed A.D., Luchko Y., 2005, Algorithms for the fractional calculus: A selection of numerical methods, *Computer Methods in Applied Mechanics and Engineering*, **194**, 743-773
5. Erturk A., Inman D.J., 2011, *Piezoelectric Energy Harvesting*, John Wiley and Sons, The Atrium
6. Freundlich J., 2016, Dynamic response of a simply supported viscoelastic beam of a fractional derivative type to a moving force load, *Journal of Theoretical and Applied Mechanics*, **54**, 4, 1433-1445
7. Freundlich J., 2019, Transient vibrations of a fractional Kelvin-Voigt viscoelastic cantilever beam with a tip mass and subjected to a base excitation, *Journal of Sound and Vibration*, **438**, 99-115
8. Freundlich J., 2021, Transient vibrations of a fractional Zener viscoelastic cantilever beam with a tip mass, *Meccanica*, **56**, 1971-1988
9. Gürgöze M., Zeren S., 2011, The influences of both offset and mass moment of inertia of a tip mass on the dynamics of a centrifugally stiffened visco-elastic beam, *Meccanica*, **46**, 1401-1412
10. Mainardi F., Spada G., 2011, Creep, relaxation and viscosity properties for basic fractional models in rheology, *The European Physical Journal Special Topics*, **193**, 133-160
11. Malendowski M., Sumelka W., Gajewski T., Studziński R., Peksa P., Sielicki P.W., 2023, Prediction of high-speed debris motion in the framework of time-fractional model: theory and validation, *Archives of Civil and Mechanical Engineering*, **23**, 46, 1-21
12. Markiewicz M., 1995 Optimum dynamic characteristics of Stockbridge dampers for dead-end spans, *Journal of Sound Vibration*, **188**, 243-256
13. Meirovitch L., 1967, *Analytical Methods in Vibrations*, Macmillan, New York

14. Podlubny I., 1999, *Fractional Differential Equations*, Academic Press, San Diego

15. Rama Bhat B., Wagner H., 1976, Natural frequencies of a uniform cantilever with a tip mass slender in the axial direction, *Journal of Sound and Vibration*, **45**, 304-307

16. Rossikhin Y.A., Shitikova M.V., 1997, Application of fractional derivatives to the analysis of damped vibrations of viscoelastic single mass systems, *Acta Mechanica*, **120**, 109-125

17. Rossikhin Y.A., Shitikova M.V., 2009, Application of fractional calculus for dynamic problems of solid mechanics: Novel trends and recent results, *Applied Mechanics Reviews*, **63**, 1, 010801-1--010801-51

18. Seidel H., Csepregi L., 1984, Design optimization for cantilever-type accelerometers, *Sensors and Actuators*, **6**, 2, 81-92

19. Shen Y., Hua J., Hou Q., Xia X., Liu Y., Yang X., 2022, Performance analysis of the fractional-order vehicle mechatronic ISD suspension with parameter perturbation, *Journal of Theoretical and Applied Mechanics*, **60**, 1, 141-152

20. Sumelka W., 2016, On geometrical interpretation of the fractional strain concept, *Journal of Theoretical and Applied Mechanics*, **54**, 2, 671-674

21. Sumelka W., Łuczak B., Gajewski T., Voyiadjis G.Z., 2020, Modelling of AAA in the framework of time-fractional damage, *International Journal of Solids and Structures*, **206**, 30-42

22. Suzuki J.L., Kharazmi E., Varghaei P., Naghibolhosseini M., Zayernouri M., 2021, Anomalous nonlinear dynamics behavior of fractional viscoelastic beams, *Journal of Computational and Nonlinear Dynamics*, **16**, 111004-1-11

23. Tayel I.M., Hassan A.F., 2019, Heating a thermoelastic half space with a surface absorption pulsed laser using fractional order theory of thermoelasticity, *Journal of Theoretical and Applied Mechanics*, **57**, 2, 489-500

24. Torvik P.J., Bagley R.L., 1984, On the appearance of the fractional derivative in the behavior of real materials, *Journal of Applied Mechanics*, **51**, 294-298

25. Yang H., 2017, Vibration control for a cantilever beam with an eccentric tip mass using a piezoelectric actuator and sensor, *International Journal of Acoustics and Vibration*, **22**, 1, 84-91

Manuscript received December 18, 2023; accepted for print March 20, 2024

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
62, 2, pp. 435-455, Warsaw 2024
https://doi.org/10.15632/jtam-pl/186718

ERROR MEASURES AND SOLUTION ARTIFACTS OF THE HARMONIC BALANCE METHOD ON THE EXAMPLE OF THE SOFTENING DUFFING OSCILLATOR[1]

Hannes Dänschel

Technische Universität Berlin, Institut für Mechanik, Berlin, Germany

corresponding author Hannes Dänschel, e-mail: hannes.daenschel@tu-berlin.de

Lukas Lentz

Hochschule Trier, Institut für Betriebs- und Technologiemanagement, Trier, Germany

Utz von Wagner

Technische Universität Berlin, Institut für Mechanik, Berlin, Germany

The Harmonic Balance Method (HBM) is one of the most often applied semi-analytic approximation methods in nonlinear dynamics. In earlier publications, the two coauthors already observed for the softening Duffing oscillator and other systems that especially low ansatz order HBM solutions may contain larger errors for some solution branches, and called this artifacts. In the present work, this problem is studied systematically with a new implementation of the method and applied again to the example of the softening Duffing oscillator. In conjunction with a mathematical definition for HBM artifacts we discuss and present possible *a posteriori* and *a priori* HBM error measures.

Keywords: Harmonic Balance Method (HBM), artifact solutions, softening Duffing oscillator

1. Introduction

In 1918, German engineer Georg Duffing published his seminal work (Duffing, 1918) exploring the dynamics of forced nonlinear oscillations. The considered system, nowadays called the Duffing oscillator, is represented by the second-order differential equation

$$x''(t) + \delta x'(t) + \alpha x(t) + \beta x^2(t) + \gamma x^3(t) = \widehat{u}\cos(\Omega t) \tag{1.1}$$

This equation describes the displacement x of a system subjected to a linear damping force (parameter δ) as well as linear (α), quadratic (β) and cubic (γ) restoring forces, along with a harmonic driving force defined by its amplitude $\widehat{u}$ and frequency Ω. The first and second order time derivatives of x are denoted by x' and x'', respectively. Thereby and in the following, all parameters and variables, including time, are considered to be dimensionless. Whilst Eq. (1.1) may be interpreted as a nonlinear extension of the standard harmonic oscillator (recovered when $\beta = 0$ and $\gamma = 0$), its dynamics is vastly more complex (Ueda, 1991). In the book by Kovacic and Brennan (2011), many details about history, applications, solution methods and phenomena of the Duffing equation can be found. Due to its simple form and the ability to display a plethora of nonlinear phenomena like multiple coexisting solutions, subharmonic and superharmonic components in the system response, bifurcations (Holmes and Rand, 1976) as well as chaotic behavior for certain parameter choices (Novak and Frenlich, 1982), the Duffing

[1]Paper presented during PCM-CMM 2023, Gliwice, Poland

oscillator rapidly became a model of significant theoretical and practical interest and was covered in introductory textbooks on nonlinear dynamics, e.g. Nayfeh and Mook (1979) or Strogatz (1994).

In the context of analyzing the Duffing oscillator, it is important to note that closed-form solutions in general are not available. Therefore, it becomes necessary to apply alternative methods, such as numerical integration or approximate analytical techniques, to explore the system behavior. While applying numerical integration, one starts from a given set of initial conditions and may end up in an asymptotically stable stationary periodic solution, quasiperiodic, chaotic or in general irregular behavior, or drifting to $\pm\infty$. Compared to this, semi-analytic approximation methods like Lindstedt-Poincaré perturbation analysis or the Harmonic Balance Method (HBM) applied here, calculate stationary solutions without transients, while other methods like Multiple (Time) Scales are also able to calculate transient behavior (Hagedorn, 1981). Nonlinear systems may have multiple stationary periodic solutions (some of them being stable and some unstable) for one excitation frequency, which is easily recognized, when methods are directly applied to calculate them. To get the variety of multiple stationary solutions while applying numerical integration, initial conditions have to be varied.

In scenarios where the focus is solely on stationary periodic solutions, the HBM can be applied with great benefit. It was introduced in Urabe (1965) as an application of the Galerkin method with harmonic ansatz functions, and nowadays is widely known under the name HBM. In the HBM, a finite Fourier series representation is used to approximate the exact solution of the nonlinear system under consideration.

In general, in the HBM, the accuracy of a solution can be improved by increasing the truncation order of the series. Below certain truncation orders, the solution behavior may differ largely from the real solutions. Some examples of such anomalous solutions produced by the HBM at lower approximation orders are e.g. documented in previous articles of the coauthors (von Wagner and Lentz, 2016, 2018, 2019) or in the book of Krack and Gross (2019). Corresponding error estimates were first derived by Urabe (1965) and the associated error bounds were later improved upon by García-Saldaña and Gasull (2013), Kogelbauer Brennan (2021) and Woiwode and Krack (2023). For further considerations, we refer to the book by Krack and Gross (2019).

As considered e.g. in von Wagner and Lentz (2016, 2018), applying the HBM to the softening Duffing oscillator results in the occurence of high amplitude solutions for low excitation frequencies with the shape of a "nose" with large residua for small ansatz orders. The increasing of the ansatz order results in a significant change of the shape and frequency range of occurence of these solutions. This was called *artifact behavior* by the authors but a comprehensive investigation of a rigorous definition and possible critera of their *a priori* or *a posteriori* detection is yet missing. Therefore, in the present work this problem is studied systematically by discussing error measures for the error in the HBM. Hereby, the object of study is again the softening Duffing oscillator where we restrict the problem to one with solutions with a vanishing mean value with the consequence of a vanishing constant, and even terms in the HBM ansatz. As shown in von Wagner and Lentz (2016, 2018) solutions with the non-vanishing mean value exist for the softening Duffing oscillator but are inconspicuous with respect to artifacts. Instead, we consider mainly the already mentioned "nose" shaped solution branch.

The paper is structured as follows. In Section 2, we provide a description of the HBM and its implementation performed by the first author. A test case is considered showing the problems of HBM as discussed in the following. In Section 3, a definition for the artifact solution is provided and then error measures based on numerical and geometrical considerations are introduced, and in Section 4 applied to the considered case of the softening Duffing oscillator. The paper ends with corresponding conclusions in Section 5.

2. Harmonic Balance Method (HBM)

In this Section, we present the theoretical basics required to obtain the frequency response of the softening Duffing system (1.1) by means of the HBM and numerical continuation methods. Let a range of excitation frequencies $\Omega \in \mathbb{F} := [\underline{\Omega}, \overline{\Omega}]$, the system parameters $\delta, \alpha, \beta, \gamma$ and the harmonic excitation $\widehat{u}\cos(\Omega t)$ be given. We denote the system frequency response as the set

$$\Gamma(\mathbb{F}) := \Big\{ (\Omega, \|x\|) \in \mathbb{R}^2 \mid x(t) = x(t+T) \text{ solves (1.1)}, \ \Omega \in \mathbb{F} \Big\} \tag{2.1}$$

with the period $T = 2\pi/\Omega$ and a later to be specified solution amplitude $\|x\|$. Computing an approximation of (2.1) requires to find approximations to x by means of the HBM over samples of the frequency range $\mathbb{F}$.

2.1. Fundamentals

The HBM is a mean weighted residual method that comprises of two approximation steps. As preliminaries, consider a time domain $\mathbb{T} := [0, T]$, an *ansatz* or *approximation order* $n \in \mathbb{N}$ as well as a real Fourier space

$$\mathcal{F}_n(\mathbb{T}, \Omega) := \Big\{ x_n \ : \ \mathbb{T} \to \mathbb{R} \mid x_n(t) = c_0\phi_0(t) + \sum_{j=1}^{n} \big(c_{2j-1}\phi_{2j-1}(t) + c_{2j}\phi_{2j}(t) \big) \Big\} \tag{2.2}$$

with the basis functions $\phi_0(t) = 1$, $\phi_{2j-1}(t) = \cos(j\Omega t)$ and $\phi_{2j}(t) = \sin(j\Omega t)$, $j = 1, \dots, n$. For convenience, we define the vector of Fourier coefficients $\mathbf{c}_n := [c_j]_{j=0}^{2n} \in \mathbb{R}^{2n+1}$ which also allows to identify $x_n \in \mathcal{F}_n$ with $\mathbf{c}_n \in \mathbb{R}^{2n+1}$. Finally, consider the residual function of the Duffing system (1.1)

$$r(t, x) = x''(t) + \delta x'(t) + \alpha x(t) + \beta x^2(t) + \gamma x^3(t) - \widehat{u}\cos(\Omega t) = 0 \tag{2.3}$$

The first step of the HBM is inserting the ansatz $x \approx x_n \in \mathcal{F}_n$ into the Duffing residual from which we obtain $r(t, x) \approx r(t, x_n) = r(t, \mathbf{c}_n)$. Since residual (2.3) is a third-degree polynomial in the trigonometric polynomial x_n, and the excitation $u \in \mathcal{F}_1$ is a simple harmonic, the convolution theorem yields that $r(t, \mathbf{c}_n)$ can be expressed as a truncated Fourier series of the order $3n$, i.e.

$$r(t, \mathbf{c}_n) = R_n(t, \mathbf{c}_n) := R_0(\mathbf{c}_n) + \sum_{j=1}^{3n} \big(R_{2j-1}(\mathbf{c}_n)\cos(j\Omega t) + R_{2j}(\mathbf{c}_n)\sin(j\Omega t) \big) \tag{2.4}$$

The second approximation step of the HBM requires that the first $2n+1$ Fourier coefficients of (2.4) vanish, i.e. $R_i(\mathbf{c}_n) = 0$. This is imposed by the $2n+1$ conditions

$$\langle R_n(\cdot, \mathbf{c}_n)\phi_i \rangle_{\mathcal{F}_n} = \frac{1}{T} \int_0^T R_n(t, \mathbf{c}_n)\phi_i(t) \, dt = 0 \qquad \forall i = 0, 1, \dots, 2n \tag{2.5}$$

In fact, by the definition of Fourier coefficients the equality $R_i(\mathbf{c}_n) = \langle R_n(\cdot, \mathbf{c}_n)\phi_i \rangle_{\mathcal{F}_n} = 0$ holds for all $i = 0, 1, \dots, 2n$ (Herman 2016). The conditions (2.5) basically ensure that the error introduced in the residual only comprises of higher order harmonics since for all $t \in \mathbb{T}$ it holds that

$$\begin{aligned} R_n(t, \mathbf{c}_n) = & \underbrace{R_0(\mathbf{c}_n) + \sum_{j=1}^{n} \big(R_{2j-1}(\mathbf{c}_n)\cos(j\Omega t) + R_{2j}(\mathbf{c}_n)\sin(j\Omega t) \big)}_{=0} \\ & + \sum_{j=n+1}^{3n} \big(R_{2j-1}(\mathbf{c}_n)\cos(j\Omega t) + R_{2j}(\mathbf{c}_n)\sin(j\Omega t) \big) \end{aligned} \tag{2.6}$$

The $2n+1$ equations (2.5) define the algebraic equation system

$$F_n(\mathbf{c}_n, \Omega) := [R_j(\mathbf{c}_n, \Omega)]_{j=0}^{2n} = 0 \tag{2.7}$$

that can be solved for $\mathbf{c}_n$. In order to compute the frequency response $\Gamma(\mathbb{F})$, we explicitly include Ω as a parameter in (2.7). Finally, if $\mathbf{c}_n$ solves (2.7) we refer to $\mathbf{c}_n$, x_n and R_n as *HBM coefficient vector*, *HBM solution* and *HBM residual*, respectively.

2.2. Solvers

In the following, we discuss how to compute the frequency response $\Gamma(\mathbb{F})$ by solving the parameter-dependent algebraic system (2.7).

2.2.1. Determining Fourier coefficients

Solving the algebraic system (2.7) in order to obtain the Fourier coefficients $\mathbf{c}_n$ one requires to evaluate F_n for which the integrals (2.5) must be determined. Since the Duffing nonlinearities are polynomials in x, the evaluation can be done by obtaining the integral closed form as well as via the discrete convolution or discrete Fourier transform (Krack and Gross, 2019; Woiwode *et al.*, 2020). However, since we also want to investigate the influence of the algebraic structure on the solution artifacts we opted for an equivalent approach of obtaining an algebraic expression of the truncated Fourier series of the residual (2.4) in the Fourier coefficients $\mathbf{c}_n$. The algorithmic implementation used in this work is a pure Python implementation without the use of computer algebra tools. A publication about the associated theoretical details as well as the source code is planned.

2.2.2. Newton's method

The next step is to solve (2.7). Let $\Omega \in \mathbb{F}$ be fixed, an approximation order $n \in \mathbb{N}$ be given and assume the Jacobian of F w.r.t. $\mathbf{c}_n$ is regular. Then, for any initial guess $\mathbf{c}_n^0 \in \mathbb{R}^{2n+1}$ "close enough" to $\mathbf{c}_n$, the algebraic system (2.7) can be solved iteratively by Newton's method for an approximated solution $\mathbf{c}_n \approx \mathbf{c}_n^k \in \mathbb{R}^{2n+1}$ subject to a prescribed iteration error tolerance $\varepsilon > 0$ s.t. for some $k \in \mathbb{N}$, we have $\|\mathbf{c}_n^k - \mathbf{c}_n^{k-1}\| \leqslant \varepsilon$ (Deuflhard, 2011).

2.2.3. Displaying the results

For error measures and visualization of the HBM results, the amplitude of $x_n \leftrightarrow \mathbf{c}_n$ must be measured. One option of computing the amplitude of x_n is the maximum-norm $\|x_n(\mathbf{c}_n)\|_\infty$. However, in order to compute this in a robust manner one needs to compute the roots of x_n'. Determination of the derivative x_n' is trivial. The roots of the trigonometric polynomial x_n' can be interpreted as the eigenvalues of the associated Frobenius companion matrix (Edelman and Murakami, 1995). However, obtaining these eigenvalues is accompanied by typical numerical challenges of this type of problem (De Terán *et al.*, 2013). Instead, we use the readily available Euclidean norm of the HBM coefficient vector $\|\mathbf{c}_n\|_2$ to measure the system's amplitude.

2.2.4. Numerical continuation

At this point, we want to discuss how to compute an approximation of the frequency response $\Gamma(\mathbb{F})$. In principal, in analogy to (2.1) we could formulate the problem of computing an approximation to $\Gamma(\mathbb{F})$ by the set

$$\Big\{(\Omega, \|x_n(\mathbf{c}_n)\|_\infty) \in \mathbb{R}^2 \mid x_n(\mathbf{c}_n, t) = x_n(\mathbf{c}_n, t+T) \text{ solves (1.1)}, T > 0, \Omega \in \mathbb{F}\Big\}$$

However, the implication

$$\mathbf{c}_n \text{ solves (2.7)} \;\Rightarrow\; x_n(\mathbf{c}_n, t) = x_n(\mathbf{c}_n, t+T) \text{ solves (1.1) for } T > 0$$

and the choice of $\|\mathbf{c}_n\|_2$ over $\|x_n(\mathbf{c}_n)\|_\infty$ as an amplitude measure suggests the alternative problem: Compute an approximation of $\Gamma(\mathbb{F})$ by an *approximated frequency response* that is the set

$$\Gamma_n(\mathbb{F}) := \left\{(\Omega, \|\mathbf{c}_n\|_2) \in \mathbb{R}^2 \mid \mathbf{c}_n \text{ solves (2.7)},\ \Omega \in \mathbb{F}\right\} \tag{2.8}$$

The nonlinearity of the Duffing system allows for multiple *solution branches* of the (approximated) frequency response $B_n^i \subset \Gamma_n(\mathbb{F})$, $i = 1, 2, \ldots$, where $\Gamma_n(\mathbb{F}) = \{B_n^1, B_n^2, \ldots\}$. With this, computing $\Gamma_n(\mathbb{F})$ reduces to finding each solution branch B_n^i individually. In advanced implementations of the HBM this is typically done by *numerical continuation methods* (Krack and Gross, 2019; Woiwode *et al.*, 2020). The basic idea behind these methods is simple: Fix the parameter Ω, solve $F_n(\mathbf{c}_n, \Omega) = 0$ for $\mathbf{c}_n$ via Newton's method by starting at $\mathbf{c}_n^0$, compute the increment $\Omega \leftarrow \Omega + \Delta\Omega$ for an "optimal" choice of $\Delta\Omega$, perform the update $\mathbf{c}_n \leftarrow \mathbf{c}_n^0$ and repeat. Here, determining an "optimal" choice of $\Delta\Omega$ depends on F as well as a prescribed error tolerance ε. Additionally, the solvability of (2.7) is only given if the Jacobian of F w.r.t. $\mathbf{c}_n$ is regular. Both topics are addressed by the specific algorithmic implementation of a numerical continuation method. A popular choice for these methods is the *pseudo-arclength method* since it can follow turning points of the solution branch and it has a robust implementation in the code `AUTO` (Deuflhard, 2011). However, it relies on empirically-based control of the stepsize $\Delta\Omega$ which happened to fail on several of the authors' examples. A noteworthy alternative is the *asymptotic numerical method* (ANM). Woiwode *et al.* (2020) provide a thorough comparison of the pseudo-arclength method and the ANM. However, in order to avoid the drawbacks of the pseudo-arclength method we opted to use the *global quasi-Gauss-Newton method* (GQGNM) as proposed by Deuflhard *et al.* (1987). Similar to the pseudo-arclength method, the GQGNM constitutes a predictor-corrector scheme in which, first, starting at the current point, a prediction step is made, scaled by a stepsize s, in the direction of the solution branch tangent. The thereby introduced error is then corrected via a quasi-Gauss-Newton iteration s.t. a prescribed iteration tolerance $\varepsilon > 0$ is fulfilled. The GQGNM employs an error estimate-based control of the stepsize s and can deal with turning points. To different capabilities it is implemented in the codes `ALCON1` and `ALCON2` (Deuflhard *et al.*, 1987; Deuflhard, 2011) in Fortran. As to the best knowledge of the authors, a Python version of said codes is not publicly available. However, since in this work the evaluation of F is implemented in a Python routine, we implemented our own version of `ALCON1` in Python of which the source code is planned to be published as well.

2.2.5. Generating initial guesses

As it is often the case in nonlinear dynamics, the task of finding "good" initial guesses for Newton's method in order to find all system frequency responses can be challenging. Fortunately, the Duffing system (1.1) allows for a systematic generation of certain initial guesses $\mathbf{c}_n^0 \in \mathbb{R}^{2n+1}$ for arbitrary approximation orders n in order to compute certain branches of its frequency response. The required approach involves two steps:

- Let $n = 1$ and determine all solutions $x_{1,i}$, $i = 1, 2, 3$, analytically at $\Omega = 0$. From this, obtain the associated HBM vectors $\mathbf{c}_{1,i}$. Then, starting at $\mathbf{c}_{1,i}$ for each $i = 1, 2, 3$ compute the associated branch $B_{1,i} \subset \Gamma_1(\mathbb{F})$ by the above introduced global quasi-Gauss-Newton method.
- Next, increase the ansatz order to $N = n + \Delta n$. Then compute $\mathbf{c}_{N,i}$ at $\Omega = 0$ by solving (2.7) via Newton's method and take $\mathbf{c}_{N,i}^0 = \left[\mathbf{c}_{n,i}^{\mathrm{T}}, 0^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{2N+1}$ as the initial guess for each

$i = 1, 2, 3$. Then compute the branches $B_{N,i} \subset \Gamma_N(\mathbb{F})$ accordingly. In fact, since two of the three solutions $\mathbf{c}_{n,i}$ are elements of the same branch, only two instead of three solution branches have to be computed.

- Repeat the previous step until each branch B_i is "sufficiently well" approximated by the approximated branch $B_{n,i}$ for some n.

Remark. The approach of computing the approximated frequency responses Γ_n as described above appears to be quite robust. In particular, the employed GQGNM required barely any tweaks of the user-adjustable parameters. However, the approach of generating initial values as described above does not yield *all* existing frequency responses of the Duffing system (1.1), cf, von Wagner and Lentz (2016). It only allows for a robust computation of the solution types already occurring for $n = 1$. In order to not over-complicate the problem, we did not endeavor to compute additonal solutions.

2.3. Test case and reference solution

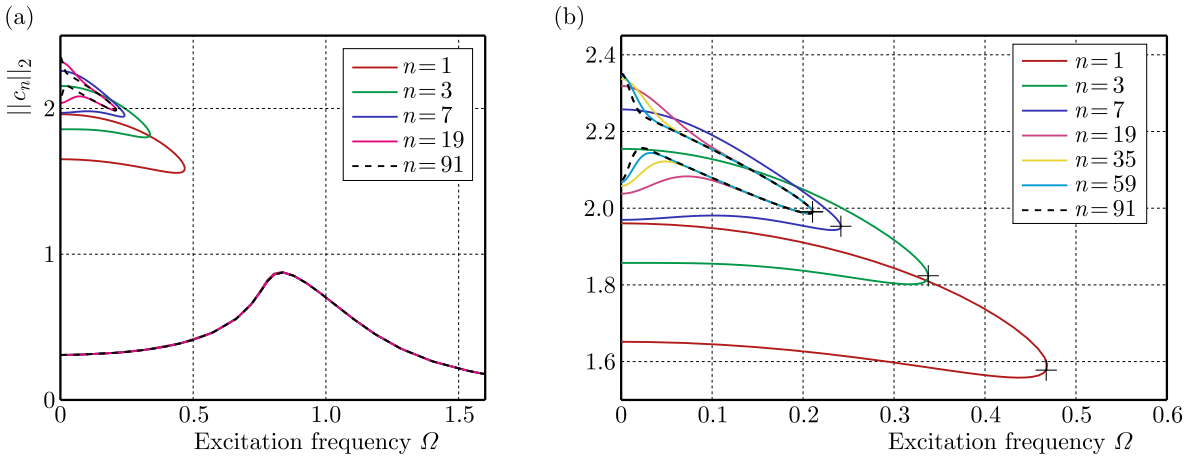


Fig. 1. Frequency response of the Duffing equation (1.1) with the test case parameters (2.9) for several approximation orders n as obtained by the solvers described in Section 2.2. The system exhibits two types of solution branches: A branch with small- respectively medium ("standard") and a branch with large-amplitude ("nose") responses. The nose solutions differ strongly in frequency range and amplitude for different ansatz orders n: (a) fequency response $\Gamma_n(\mathbb{F})$ for $\mathbb{F} = [0, 1.6]$, (b) "nose" branch of frequency response $\Gamma_n([0, 0.6])$

We consider the parameter set

$$\begin{aligned} &\delta = 2D\omega = 0.4 \qquad \alpha = \omega^2 = 1 \qquad \beta = 0 \\ &\gamma = -0.4 \qquad \widehat{u} = 0.3 \qquad \Omega \in [0, 1.6] \end{aligned} \tag{2.9}$$

with $D = 0.2$ and $\omega = 1$ as the *test case* of the Duffing system (1.1) for this study. As already mentioned in the introduction, all parameters are considered to be dimensionless which also holds for the displacement x and time t. Solver-wise we used an iteration error tolerance of $\varepsilon = 10^{-14}$ throughout this study. Corresponding results are shown in Fig. 1 for different ansatz orders n. The system exhibits the — for the softening case — well known types of solution branches: One small-, respectively medium- and two large-amplitude responses which we refer to as "standard" and "nose" branches or responses. For the resonance peak of the standard response at $\Omega \approx \omega = 1$, there are for the parameter set (2.9) no multiple solutions due to moderate damping. The standard response covers the entire considered frequency range $\mathbb{F} = [0, 1.6]$ with amplitudes in the range of $\|\mathbf{c}_n\|_2 \in [0.19, 0.87]$. Our special focus in the following is on the

nose responses, however. These nose responses cover only the frequency range from zero to the characteristic turning point marked by +, which differs largely for different ansatz orders n. This behavior has been denoted as "artifacts" by the second and third author of the present paper and investigated in several papers, e.g. von Wagner and Lentz (2016, 2018, 2019). In these prior publications certain solutions are considered as artifacts that exceed a maximum amplitude threshold w.r.t. the neglected higher order terms of the HBM method, i.e., in the Duffing equation, the terms with harmonics of order $n+1$ to $3n$.

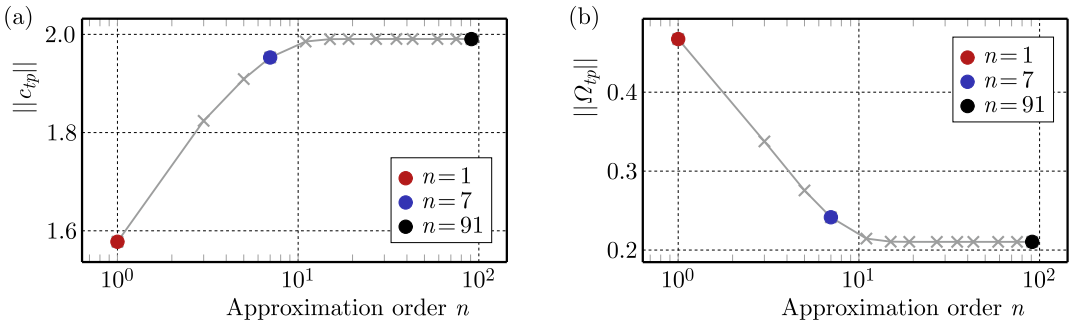


Fig. 2. Convergence of the amplitude $\|c_{tp}\|$ and excitation frequency Ω_{tp} of the turning point

In contrast, in the present paper we consider new suitable aspects for defining and identifying artifact solutions as presented in Sections 3 and 4. Thereby, several numerical and geometrical measures, e.g. the position of the turning point of the nose response and its convergence w.r.t. the ansatz order n of the associated HBM solution x, are investigated. As will be seen, not all of the examined measures are useful with respect to the task of determining artifacts. As the first attempt, we consider in Fig. 2a and 2b for our test case the convergence of the nose solutions turning point denoted as $(\Omega_{tp}, \|c_{tp}\|_2)$. It can be observed that both the solution amplitude $\|c_{tp,n}\|_2$ and the excitation frequency $\Omega_{tp,n}$ appear to converge for increasing ansatz orders n. Nevertheless, as can be observed in Fig. 1, convergence of the turning point does not necessarily imply convergence of all other solution points of the nose branch, where in general larger ansatz orders are necessary. The iteration error of the ansatz orders $n = 75, 91$ yields $\big|\|c_{tp,91}\|_2 - \|c_{tp,75}\|_2\big| \approx 6.67 \cdot 10^{-7}$ and $|\Omega_{tp,91} - \Omega_{tp,75}| \approx 2.48 \cdot 10^{-4}$ which we deem to be small. Hence we assume that the HBM solution x_{91} converged sufficiently close to the exact solution x of (1.1) at least at the turning point. Although for $n = 19$ the iteration error at the turning point is similiar to the one for $n = 91$, sufficient convergence is not yet achieved in a large part of the frequency range of the nose branch. This is why we consider x_{91} over a potential lower order solution as the *reference solution* with the associated ansatz order $n = 91$ of the test case (2.9).

Next, in Fig. 3, the frequency response of the ansatz order $n = 1$ is compared to the reference response of the ansatz order $n = 91$ in more detail. The frequency values of the turning points of the nose branches for $n = 91$ and $n = 1$ are found to be approximately $\Omega_{tp,91} = 0.21$ and $\Omega_{tp,1} = 0.47$, respectively. The two turning points at $\Omega_{tp,91}$ and $\Omega_{tp,1}$ can be considered to divide the entire frequency range into the sub-intervals $\mathbb{F}_A := [0, \Omega_{tp,91}]$, $\mathbb{F}_B := [\Omega_{tp,91}, \Omega_{tp,1}]$ and $\mathbb{F}_C := [\Omega_{tp,1}, 1.6]$ s.t. $\mathbb{F} = \mathbb{F}_A \cup \mathbb{F}_B \cup \mathbb{F}_C$. Now note that the reference nose response only covers $\mathbb{F}_A$ but the lower order nose response of $n = 1$ covers $\mathbb{F}_A$ and also $\mathbb{F}_B$. Apparently, the solutions of the response of $n = 1$ for all frequencies in $\mathbb{F}_B$ "vanish" upon an increase of the ansatz order to $n = 91$. From this, we conclude that the solution of the frequency response of $n = 1$ in the frequency range $\mathbb{F}_B$ are *artifact solutions* as defined in more detail in Section 3.

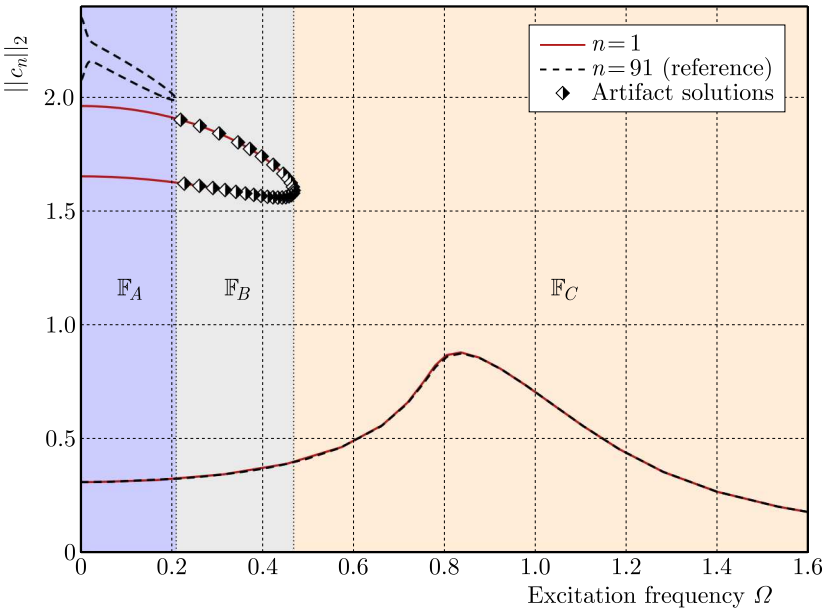


Fig. 3. Approximated frequency response $\Gamma_n(\Omega)$ for ansatz orders $n = 1, 91$ of the case (2.9). Solutions in the range $\mathbb{F}_B$ are denoted as artifact solutions, cf. Definition 1 in Section 3

3. Error measures

As can be seen from the results in Fig. 1 and 3, HBM solutions of low ansatz orders exhibit, especially for the nose solutions, a deviation from the reference solution ($n = 91$). This deviation is considered to be an error due to HBM and is divided into two types of errors. The first refers to the amplitude error which we refer to as *quantitative error* which can be measured, e.g., by the convergence error $\|\mathbf{c}_n - \mathbf{c}_{n_{ref}}\|$. The second error type is the artifact behavior which we refer to as *qualitative error*. In order to be able to measure the qualitative error, a mathematical definition of the artifact behavior is required. The definition is presented and discussed in the following. After that we discuss potential qualitative error measures based on the residual as well as algebraic, geometric and solver-related properties.

3.1. Artifact definition

In the following, we provide and discuss a mathematical definition of the artifact behavior which we base on the turning points of the solution branches. We start by providing the required mathematical lingua. Let $n, n_{ref} \in \mathbb{N}$ be two HBM ansatz orders with $n < n_{ref}$ as well as $B_n \subset \Gamma_n(\mathbb{F})$ and $B_{n_{ref}} \subset \Gamma_n(\mathbb{F})$ two computed solution branches, respectively. Again, we refer to the solutions of the ansatz order n_{ref} as reference (solutions). Recall that $B_n = \{P_1, \dots, P_N\}$ and $B_{n_{ref}} = \{Q_1, \dots, P_M\}$ are *ordered* point sets with points $P_i = (\Omega_n^i, (\|\mathbf{c}_n\|_2)^i) \in \mathbb{R}^2$, $i = 1, \dots, N$, and $Q_j = (\Omega_{n_{ref}}^j, (\|\mathbf{c}_{n_{ref}}\|_2)^j) \in \mathbb{R}^2$, $j = 1, \dots, M$, where $N, M \in \mathbb{N}$ are the number of points of the solution branches B_n and $B_{n_{ref}}$, respectively. We assume that B_n and $B_{n_{ref}}$ each exhibit a turning point denoted as $P_{tp} \in B_n$ and $Q_{tp} \in B_{n_{ref}}$ and we denote the associated excitation frequency as $\Omega_{n,tp}$ and $\Omega_{n_{ref},tp}$, respectively. In order to compare the similarity of the turning points P_{tp} and Q_{tp}, we consider their local curvature w.r.t. the frequency component. For this, let X be a turning point on a curve $B \in \mathbb{R}^2$ and let $\mathcal{B}_s(X) \subset B$ denote an arclength-parameterized neighborhood[1] of B at X with arclength $s > 0$. With this, we define the *signed normalized curvature w.r.t. Ω at X* as

[1]That is, all points Y that lie on B and that are closer to X than the arclength $s > 0$.

$$\kappa_\Omega(X) := \begin{cases} +1 & \forall Y \in \mathcal{B}_s(X): \quad (Y)_\Omega > (X)_\Omega \\ -1 & \forall Y \in \mathcal{B}_s(X): \quad (Y)_\Omega < (X)_\Omega \end{cases} \tag{3.1}$$

where $(X)_\Omega, (Y)_\Omega$ denote the Ω-component of $X, Y \in B$, respectively. In (3.1), $\kappa_\Omega(X) = \pm 1$ basically means that at the turning point X the curve B "opens" to the right (resp. left). With this, we can present the following

Definition 1 (Artifact solution). *Let two solution branches B_n and $B_{n_{ref}}$ be given. Assume they each exhibit a single turning point $P_{tp} \in B_n$ and $Q_{tp} \in B_{n_{ref}}$. If $\kappa(P_{tp}) = \kappa(Q_{tp}) = \pm 1$ and $\Omega_{n_{ref},tp} \gtrless \Omega_{n,tp}$ then all points $P \in B_n$ with frequency components Ω in the frequency range $[\Omega_{n,tp}, \Omega_{n_{ref},tp}]$ (respectively $[\Omega_{n_{ref},tp}, \Omega_{n,tp}]$) are called* artifact solutions.

Two possible situations for artifact solutions to occur as given in Definition 1 are depicted in Fig. 4a and 4b.

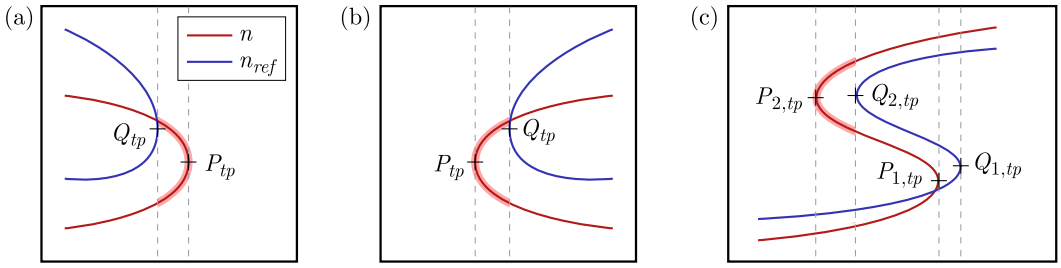


Fig. 4. Three representative cases of artifact solutions (▬) on solution branches as identified by applying Definition 1: (a) and (b) single turning point per solution branch B_n and $B_{n_{ref}}$; (c) multiple turning points per solution branch B_n and $B_{n_{ref}}$, however artifact solutions are only existent between the turning points $P_{2,tp}$ and $Q_{2,tp}$

In case both solution branches B_n and $B_{n_{ref}}$ exhibit multiple turning points then the above definition may be applied in succession according to the following scheme:

1. Assume that B_n and $B_{n_{ref}}$ exhibit the same number $H \in \mathbb{N}$ of turning points denoted as $P_{tp,1}, \ldots, P_{tp,H} \in B_n$ and $Q_{tp,1}, \ldots, Q_{tp,H} \in B_{n_{ref}}$, respectively. Further assume that the aforementioned turning points ordering coincides with the ordering of the points of B_n and $B_{n_{ref}}$, i.e. $B_n = \{\ldots, P_{tp,1}, \ldots, P_{tp,H}, \ldots\}$ and $B_{n_{ref}} = \{\ldots, Q_{tp,1}, \ldots, Q_{tp,H}, \ldots\}$, respectively.
2. If now $\kappa(P_{tp,i}) = \kappa(Q_{tp,i})$ for all $i = 1, \ldots, H$, then Definition 1 can be applied to each pair of the turning point $(P_{tp,i}, Q_{tp,i})$ successively in order to identify artifact solutions.

Figure 4c shows an exemplary case of two turning points per solution branch where the above scheme can be applied. Although there are two turning points per branch, there is only a single frequency region in which solution artifacts occur. For an algorithmic detection of artifact solutions based on Definition 1, a robust detection of turning points is critical. This can be done within the employed GQGNM by means of a readily available cubic Hermite interpolation (Deuflhard *et al.*, 1987). An alternative approach of robust computation of the turning points would be available upon implementation of the ANM (Woiwode *et al.*, 2020).

The objective in the following is to consider a number of error measures in general and at its best to find a way to distinguish between artifacts and other types of errors, and to avoid both of them. To this end, various methods of measuring the error of an HBM solution will be introduced and applied to the test case described in Subsection 2.3. Of course, other classifications of errors are possible, e.g. errors due to the HBM itself, comparing the HBM solution with the exact solution and numerical errors while applying the HBM. These errors

have been studied in detail in e.g. Urabe (1965), Kogelbauer and Breunung (2021), García--Saldaña and Gasull (2013), Woiwode and Krack (2023). These studies do not investigate the qualitative error type, i.e. the artifact behavior described above is not considered. However we want to point out that in the work of Woiwode and Krack (2023), the suggested n-adaptive error measure appears to us to be a potential tool of detecting artifacts. Within their approach of a numerical continuation method, the HBM ansatz order is adaptively refined or coarsened based on an error measure. The adaptive switching of the ansatz order could possibly speed up the detection of turning points which is a mandatory step to successfully apply the solution artifact Definition 1. To our understanding, this approach could, in principle, avoid the computation of possible artifact solutions, although a confirmation of this hypothesis would require further research.

3.2. Residual

An obvious way to measure the error of a HBM solution is to consider the terms neglected in equation (2.6), which are evaluated in the following. As can be inferred from the definition of this expression, the value of this residual must be zero if the HBM solution exactly satisfies the underlying differential equation. The residual thus represents a measure of the non-fulfillment of the differential equation, but does not provide direct information about the extent to which the HBM solution deviates from the exact solution. Nevertheless, it has been shown in previous works, e.g. Ferri and Leamy (2009), von Wagner and Lentz (2018, 2019), Lentz and von Wagner (2020), that the residual can be used to determine whether the approximation order of an HBM solution needs to be further increased to achieve the HBM solution that accurately represents the exact solution. A drawback of this procedure, as shown e.g. in von Wagner and Lentz (2018, 2019), Lentz and von Wagner (2020) is that the residual is not suitable for distinguishing between the quantitative and qualitative error type. A high value merely indicates that the examined HBM solution has a high error. Therefore, the solution might be an artifact, or it could be a solution that exists but provides a poor approximation of the exact amplitude due to an insufficient order of approximation. Therefore, as the residual was considered in several earlier publications of the authors, it is not further considered in the subsequent analysis in the present paper.

3.3. Algebraic measures

Another approach to investigate the error associated with a solution involves a direct examination of the underlying algebraic system of equations. The approach consists of searching within the algebraic system equations F_n for indications that finding a solution will be problematic. If such indications are present, it is reasonable to assume that the solutions found are flawed. Therefore, various methods will be enumerated in the following, which can be used to estimate the quality of a solution based on the solved system of equations. Since the properties of the algebraic equation system are largely determined by the linear component, the Jacobian matrix — in short *Jacobian* — at the solution point is used for this purpose. Since the equation system can be considered as a function of the coefficients or the excitation frequency, the following three Jacobian can be defined

$$J_{\mathbf{c},n}(\mathbf{c},\Omega) := D_{\mathbf{c}}F_n(\mathbf{c},\Omega) \qquad J_{\Omega,n}(\mathbf{c},\Omega) := D_{\Omega}F_n(\mathbf{c},\Omega)$$
$$J_n(\mathbf{c},\Omega) := \left[J_{\mathbf{c},n}(\mathbf{c},\Omega), J_{\Omega,n}(\mathbf{c},\Omega)\right]$$

Here, $J_{\mathbf{c},n}(\mathbf{c},\Omega)$ denotes the Jacobian of F_n w.r.t. $\mathbf{c}$, $J_{\Omega,n}(\mathbf{c},\Omega)$ denotes the Jacobian of F_n w.r.t. Ω and $J_n(\mathbf{c},\Omega)$ denotes the full Jacobian of F_n.

3.3.1. Condition number

Before starting to assess the qualitative error of solutions, we investigate the ill- or well--posedness of the problem of solving the system $F_n(\mathbf{c}_n, \Omega) = 0$ in the scope of Newton's method as required in the solvers described in Section 2.2. For $A = J_{\mathbf{c},n}(\mathbf{c}, \Omega)$ or $A = J_n(\mathbf{c}, \Omega)$, this is measured by the *condition number of* A

$$\operatorname{cond}_2(A) := \|A\|_2 \|A^{-1}\|_2 \tag{3.2}$$

where $\|\cdot\|_2$ is the matrix norm induced by the Euclidean vector norm. An upper bound for the relative error amplification made during solving of the linear equation system within Newton's method is given by the factor $\operatorname{cond}_2(A) \cdot \delta$, where δ is the relative error in $\mathbf{c}_n$ (Deuflhard and Hohmann, 2019). Since in Newton's method linear systems are solved iteratively, the cummulative relative error is proportional to $\operatorname{cond}_2(A) \cdot \delta \cdot M$, where M is the number of iterations. In our case, $\delta = \varepsilon = 10^{-14}$, and typically $M = 12, \ldots, 20$. Consequently, this problem is said to be *well-* or *ill-conditioned*, if $\operatorname{cond}_2(A)$ is small or large, respectively. Furthermore, A being singular is equivalent to $\operatorname{cond}_2(A) = \infty$. This is the case for the turning point of the nose solution branch at which $J_{\mathbf{c},n}(\mathbf{c}, \Omega_{tp})$ is singular. Hence, in numerical practice, large condition numbers can be used as an indicator for singularities of the associated matrix. To illustrate this concept, Fig. 5 depicts values of the condition number for the range $(c_1, c_2) \in [-2.4, 2.4]^2$ and for excitation frequencies $\Omega = 0.0,\ 0.2,\ 0.4,\ 0.6$. Additionally, the standard branch solution (◓) as well as the two nose branch solutions (larger amplitude: ⬗, smaller amplitude: ⬖) existing at each frequency are marked. Based on these graphs, it is possible to assess the values that the condition number of the Jacobian matrix takes in the vicinity of the solutions. As can be seen, solutions with a large amplitude (i.e. ⬗, ⬖) are located in a region with high values, while the solutions with a low amplitude (i.e. ⬗, ⬖) is in a region with low values.

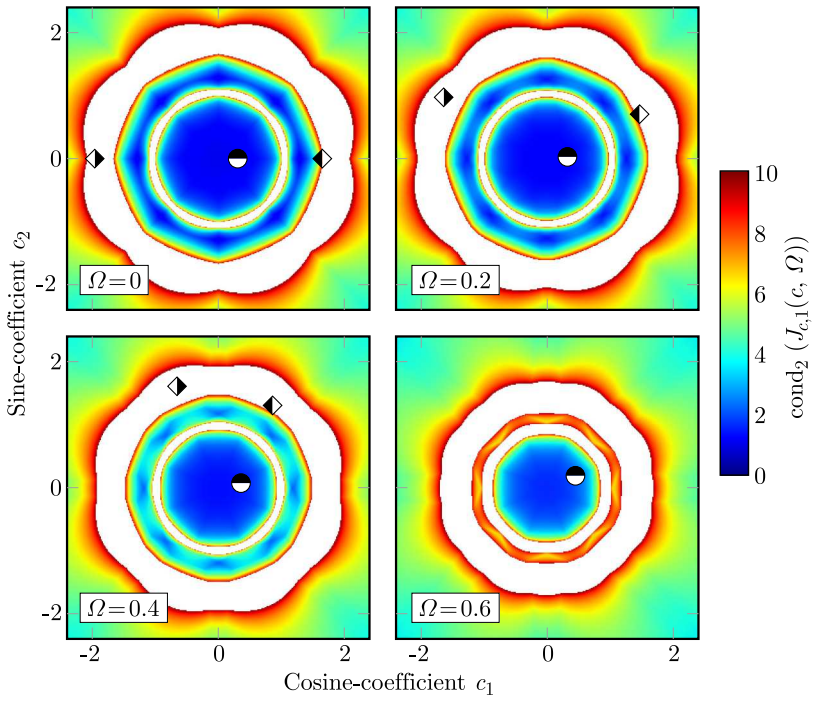


Fig. 5. Condition number $\operatorname{cond}_2\big(J_{\mathbf{c},1}(\mathbf{c}, \Omega)\big)$ for $n = 1$ for four different excitation frequencies of test case (2.9). Values greater than ten are color-coded to white. The symbols ⬗, ⬖ and ◓ denote the two nose solutions and the standard solution, respectively

3.3.2. Jacobian angle

Another possibility is to consider the angles between the columns of the Jacobian. The motivation for this is that these angles can be used as a measure of the linear independence of the linearized equations. For example, an angle of 90° means that the equations are linearly independent, while an angle of 0° means that the equations are linearly dependent. With regard to the solvability of the equation system, it is therefore assumed that small angles may indicate difficulties in computing the solution. For a more precise definition, let $(\mathbf{J}_{\mathbf{c},n}(\mathbf{c},\Omega))_i \in \mathbb{R}^{1,2n+1}$ denote the i-th row vector of $J_{\mathbf{c},n}(\mathbf{c},\Omega)$ for all $i = 0, 1, \ldots, 2n$. In particular, for $n = 1$ and $c_0 := 0$ let $\mathbf{J}_1, \mathbf{J}_2 \in \mathbb{R}^{1,2}$ denote the first and second row vector of the Jacobian matrix $J_{\mathbf{c},1}(\mathbf{c},\Omega)$. Then

$$\theta := \arccos \frac{\mathbf{J}_1 \mathbf{J}_2^{\mathrm{T}}}{\|\mathbf{J}_1\| \|\mathbf{J}_2\|} \in \left[0, \frac{\pi}{2}\right] \tag{3.3}$$

is referred to as the Jacobian angle. To illustrate this concept as well, the same method as described above is employed. The corresponding graphs can be found in Fig. 6. These graphs

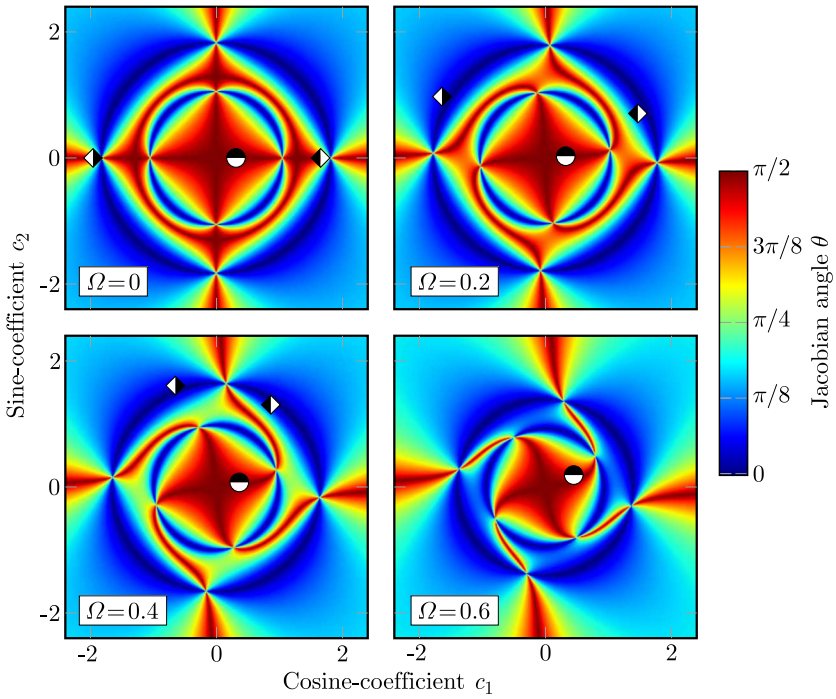


Fig. 6. Jacobian angle θ for $n = 1$ for four different excitation frequencies of the test case (2.9), symbols as in Fig. 5

also demonstrate that nose solutions with large amplitude (◗, ◖) are located in regions with poor solution properties, characterized by small Jacobian angles, while the standard solution with a small amplitude (◓) is situated in a region with good solution properties, characterized by a large Jacobian angle. In order to extend the concept of the Jacobian angle to ansatz orders $n > 1$, we consider to find the minimal Jacobian angle over all pair-wise distinct row vectors $\mathbf{J}_i \in \mathbb{R}^{1,2n+1}$ which we define as

$$\theta_{min} := \min_{i \neq j = 0,1,\ldots,2n} \arccos \frac{\mathbf{J}_i \mathbf{J}_j^{\mathrm{T}}}{\|\mathbf{J}_i\| \|\mathbf{J}_j\|} \in \left[0, \frac{\pi}{2}\right] \tag{3.4}$$

and refer to it as the *minimal Jacobian angle*. Here, we consider the minimum as the measure of choice since of all row vectors of the two associated to the minimal Jacobian angle are the pair closest to be linearly dependent. And, of course, for $n = 1$, the two definitions (3.3) and (3.4) are equivalent.

3.3.3. Number of solutions

As the final algebraic measure, we consider the number of real solutions of x of (1.1) for a given excitation frequency Ω. In the context of the HBM, this requires to investigate *the number of real solutions of* F_n for a given ansatz order n and excitation frequency Ω, which we denote as $\#F_n$. In general, there exists no *a priori* way of determining $\#F_n$ expect for computing *all* real solutions $\mathbf{c}_{n,i}$, $i = 1, \dots, \#F_n$. However, this is impractical since Bézout's theorem provides $\#F_n \leqslant \prod_{i=0}^{2n} \deg(R_i)$ as the upper bound for the number of solutions of F_n where $\deg(R_i)$ is the degree of the multivariate polynomial R_i which is the i-th equation of F_n (Basu *et al.*, 2006). Fortunately, in this work we only consider the standard and nose solution branch as a subset of the entire frequency response of the Duffing system. This reduces the complexity of the problem of determining $\#F_n$ drastically to simply counting the number of solutions of $\Gamma_n(\{\Omega\})$ for each $\Omega \in \mathbb{F}$, i.e. $\#F_n(\Omega) = |\Gamma_n(\{\Omega\})|$, where $|\cdot|$ denotes the cardinality of $\Gamma_n(\{\Omega\})$. Considering only the standard and nose branch, this yields $\max_{\Omega\in\mathbb{F}} |\Gamma_n(\{\Omega\})| = 3$. With this, the number of solutions can be compared for different ansatz orders and different frequencies. In fact, this procedure can be applied to any combination of two ansatz orders $n_1 < n_2$ for which the difference $|\Gamma_{n_2}(\{\Omega\})| - |\Gamma_{n_1}(\{\Omega\})|$ needs to be determined for every $\Omega \in \mathbb{F}$. Intuitively, the main disadvantage of this approach is that the two frequency responses Γ_{n_1} and Γ_{n_2} need to be computed beforehand, i.e. it is not an *a priori* measure that identifies artifacts for a requested ansatz order — it needs the frequency response of the second, higher ansatz order as a reference.

3.4. Geometric measures

Next, we want to investigate two geometric measures. In order to be able to compare the "resemblance" of two solution branches of the frequency response, an adequate measure is required. We already discussed the convergence of the nose branch turning point in Section 2.3 and its disadvantage of not being able to capture the entirety of the solution branches. Instead, we consider two normalized distance measures that measure the distance between the solution branch of the ansatz order n and the corresponding reference solution branch of the ansatz order n_{ref}.

3.4.1. Arclength distance

First, straightforwardly, consider the *arclength* or *length* $L(B) > 0$ of the solution branch $B \in \Gamma(\mathbb{F})$. The approximated solution branch B_n can be interpreted as a polygonal curve — or polyline — represented by N points of the ordered set $\{(\Omega, \|\mathbf{c}_n\|_2)_i\}_{i=1}^N \subset \mathbb{R}^2$. A simple approximation $L(B) \approx L(B_n)$ for the approximated solution branch $B_n \subset \Gamma_n(\mathbb{F})$ is obtained by

$$L(B_n) := \sum_{i=0}^{N-1} \|d_{i+1} - d_i\| \qquad d_i := \begin{bmatrix} \Omega \\ \|\mathbf{c}_n\|_2 \end{bmatrix}_i \in \mathbb{R}^2 \tag{3.5}$$

i.e. the line segments of the polygonal curve B_n. With this, we introduce the *arclength distance* between the solution branch of the ansatz order n and n_{ref} as

$$d_L(B_n, B_{n_{ref}}) := |L(B_n) - L(B_{n_{ref}})| \tag{3.6}$$

Furthermore, in order to be able to compare multiple arclength distances, we introduce the *normalized arclength distance*

$$\overline{d}_L(B_n, B_{n_{ref}}) := \frac{d_L(B_n, B_{n_{ref}})}{L(B_{n_{ref}})} \tag{3.7}$$

where $L(B_{n_{ref}})$ is the arclength of the reference solution branch. The arclength is computationally inexpensive and, therefore, $d_L, \overline{d}_L$ readily available. However, both variants are not invariant under translation and rotation of the polygonal curves $B_n, B_{n_{ref}}$.

3.4.2. Fréchet distance

An improved but computationally more expensive distance measure is the *Hausdorff distance*. However, it does not consider the course of two compared curves. Fortunately, the so-called Fréchet distance circumvents the disadvantages of both distance measures at the expense of higher computational costs. Let a, b be parametrizations of two polylines A, B, respectively. Then the *Fréchet distance* between A and B is defined as

$$d_F(A,B) := \inf_{a,b} \max_{t\in[0,1]} \{\|A(a(t)) - B(b(t))\|_2\} \tag{3.8}$$

This distance measure captures the similarity between A and B while it takes into account the ordering and position of the curves points. An intuitive understanding of (3.8) might be obtained by the following analogy (Alt and Godau, 1995): "A person is walking a dog on a leash: the person can move on one curve, the dog on the other; both may vary their speed, but backtracking is not allowed. Then the Fréchet distance of the two curves is the minimal required length of the leash". In practice, we are actually interested in computing an approximation of the Fréchet distance for two approximated solution branches B_{n_1}, B_{n_2}. A "good" approximation of $d_F(B_{n_1}, B_{n_2})$ is given by the so-called *discrete Fréchet distance* (DFD) (Alt and Godau, 1995) which we denote by $d_{DF}(B_{n_1}, B_{n_2})$. In order to compute the DFD, we utilize the Python code `discrete-frechet` by Figueira (2023). Finally, we introduce the *normalized DFD*

$$\overline{d}_{DF}(B_n, B_{n_{ref}}) := \frac{d_{DF}(B_n, B_{n_{ref}})}{L(B_{n_{ref}})} \tag{3.9}$$

in order to be able to compare it to the normalized arclength distance.

3.5. Solver measures

In addition to the aforementioned residual, algebraic and geometric measures, we also investigate measures obtainable from the employed solvers in order to test whether information available by the solvers can indicate artifact behavior or not. For this, we consider the *number of correction steps k per prediction step* as well as the *computation time per prediction step*. Since both measures correlate strongly, we only present the number of correction steps k as a representative quantity of the solver behavior. Additionally, in order to benchmark the performance of the employed numerical continuation method we also provide data on the following solver-related quantities:

- The *computation time* in seconds t_{comp} required to obtain each solution branch per given ansatz order n.
- The *number of prediction steps* k_{pred} of each solution branch per given ansatz order n, i.e. the number of points that constitute each solution branch.
- The *total number of correction steps* k_{corr} in order to compute each solution branch per given ansatz order n, i.e. the sum of the number of correction steps over all prediction steps.
- The *average number of required correction steps* $\overline{k}_{corr} = k_{corr}/k_{pred}$ per prediction step.

4. Error measures applied to test case

In this Section, the error measures described in Section 3 are applied to the test case given in Section 2.3. This is intended to assess the extent to which these error measures are suitable for identifying artifacts and errors in general. As clarified in Section 2.3, regarding the test case, it is known that all nose solutions, i.e. those with large amplitudes, for an excitation frequency

$\Omega > 0.21$ are artifacts. Hence, an error measure that is suitable for distinguishing artifacts from regular solutions is expected to yield significantly different values for solutions with large amplitudes for excitation frequencies $\Omega > 0.21$ ($\mathbb{F}_B$ in Fig. 3) compared to excitation frequencies $\Omega \leqslant 0.21$ ($\mathbb{F}_A$ Fig. 3). Consequently, we expect a discontinuity at $\Omega = 0.21$ of such an error measure. Recall that in the present work, the focus is on ansatz functions x_n with a vanishing mean value, i.e. $c_0 = 0$, and thus $\mathbf{c}_n \in \mathbb{R}^{2n}$.

4.1. Algebraic measures

4.1.1. Condition number

The first algebraic measure results we present are the condition numbers $\mathrm{cond}_2(J_{\mathbf{c},n}(\mathbf{c}_n, \Omega))$ and $\mathrm{cond}_2(J_n(\mathbf{c}_n, \Omega))$ given in Fig. 7a and 7b, respectively. Both figures plot the respective condition number over the excitation frequency for the standard and nose response for ansatz orders $n = 1, 7, 91$. We first discuss Fig. 7a. First of all, the expected increase of the condition number for an increase of the ansatz order can be observed. For $n = 1$ the nose response exhibits a condition number of one to three orders of magnitude larger than the condition number of the standard response. In particular, towards the turning point of the nose the condition number rapidly increases towards values of $\mathrm{cond}_2(J_{\mathbf{c},1}(\mathbf{c}_1, \Omega_{tp,1})) \approx 10^3$. Similar behavior can be observed for the ansatz order $n = 7, 91$ with a maximum condition number towards the turning point at around $10^5, 10^7$, respectively. For the largest condition number associated with $n = 91$, we have $\mathrm{cond}_2(J_{\mathbf{c},91}(\mathbf{c}_{91}, \Omega_{tp,91})) \cdot \varepsilon \approx 10^7 \cdot 10^{-14} = 10^{-7} \ll 1$ for a single Newton step. Consequently, with a typical value of around $N = 15$ Newton steps until convergence we may extrapolate to $\mathrm{cond}_2(J_{\mathbf{c},91}(\mathbf{c}_{91}, \Omega_{tp,91})) \cdot \varepsilon \cdot N = 5 \cdot 10^{-6} \ll 1$. From this we conclude that, from the numerical practical standpoint, the problem of solving the linear system within Newton's method is still considered to be well-conditioned. However, note that in case the numerical continuation method reaches an excitation frequency that is numerically close to the turning point of the system, the condition number becomes unbounded and the problem ill-conditioned.

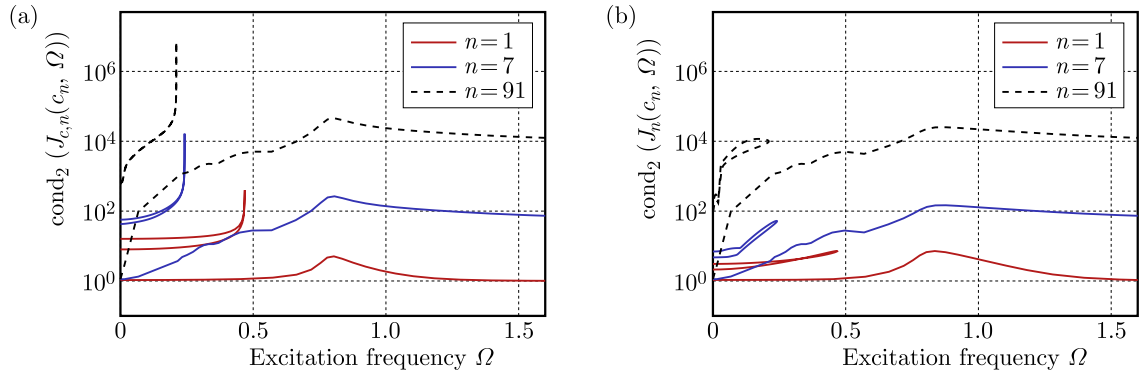


Fig. 7. Condition numbers $\mathrm{cond}_2(J_{\mathbf{c},n}(\mathbf{c}_n, \Omega))$ and $\mathrm{cond}_2(J_n(\mathbf{c}_n, \Omega))$ of the system F_n for approximation orders $n = 1, 3, 91$ for the test case (2.9)

Next, we discuss Fig. 7b. It shows a similar qualitative behavior for the condition number of the extended Jacobian matrix $\mathrm{cond}_2(J_n(\mathbf{c}_n, \Omega))$. However, the largest condition number values at the turning points of the nose branch of ansatz orders $n = 1, 7, 91$ are approximately 7.21, 52.4 and $1.16 \cdot 10^4$, respectively. A comparison of the condition number of $J_{\mathbf{c},n}$ to J_n yields that the values of the condition number of the extended Jacobian are three orders of magnitude smaller. This can be explained by the additional information due to the existence of the additional matrix column $J_{\Omega,n}$, i.e. additionally considering the derivative w.r.t. the excitation frequency improves the conditioning of the original problem. This is in fact used by the pseudo-arclength method or the GQGNM, as presented in Section 2.2. Interestingly, near the standard branch

resonance peak at $\Omega = 0.8$, the condition number of the extended Jacobian $\mathrm{cond}_2(J_n(\mathbf{c}_n, \Omega))$ for $n = 7, 91$ is three (respectively two) times larger. However, upon returning to the question of artifact solutions, for $n = 1, 7$ in the frequency range around the turning point $\Omega_{tp,91} = 0.21$ no noticeable change in the condition numbers $\mathrm{cond}_2(J_{\mathbf{c},n}(\mathbf{c}_n, \Omega))$ and $\mathrm{cond}_2(J_n(\mathbf{c}_n, \Omega))$ can be observed. This is why we do not consider either of these two condition numbers to be indicative of artifact behavior.

4.1.2. Jacobian angle

The second algebraic measure result we discuss is the minimal (extended) Jacobian angle $\theta_{min}(J_{\mathbf{c},n})$ (resp. $\theta_{min}(J_n)$) given in Fig. 8a and 8b as an extension of the Jacobian angle for an ansatz order $n \geqslant 1$.

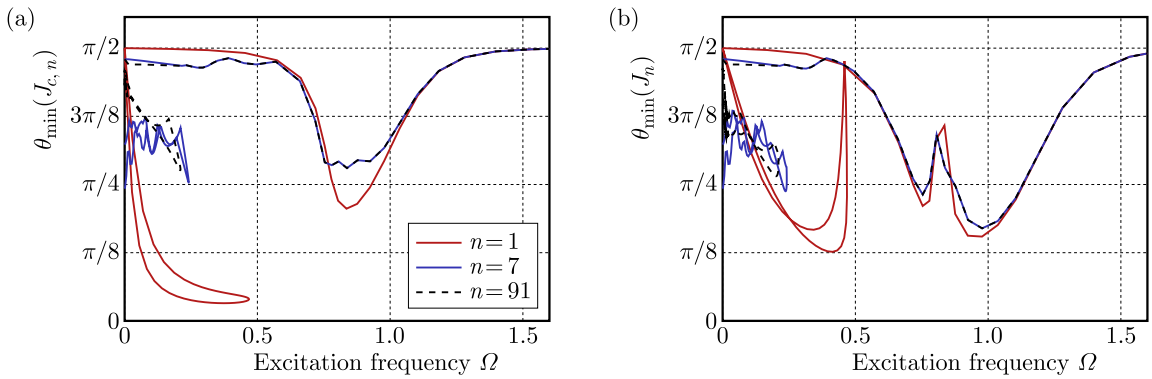


Fig. 8. Minimal (extended) Jacobian angle $\theta_{min}(J_{\mathbf{c},n})$ (resp. $\theta_{min}(J_n)$) per excitation frequency for ansatz orders $n = 1, 7, 91$ for the test case (2.9)

The two figures plot the angles over the excitation frequency for the standard and nose response for ansatz orders $n = 1, 7, 91$. We start by discussing Fig. 8a. The curves are mostly close to $\pi/2$ and have minimal values around $\pi/4$ close to the small amplitudes resonance peak. The minimum angles over all frequencies can be found close to the respective standard branch resonance peak with values of 41%, 56% and 55% of $\pi/2$ for $n = 1, 7, 91$, respectively. As could be expected, the minimal Jacobian angle for standard branches for $n = 7$ appears to be mostly converged against the reference solution branch of $n_{ref} = 91$. On the other hand, the nose branch minimal angles are 6.3%, 50% and 55% for $n = 1, 7, 91$, respectively, and are observed to be close to the respective turning points. Note, that there is a noticeable difference in the angles of roughly 30° when comparing the nose branch of ansatz orders $n = 1$ to $n = 7, 91$. Next, we turn our focus to Fig. 8b. The curves look mostly similar, expect for two noticeable outliers. First, for all three ansatz orders the standard solution branches exhibit a similar but somewhat remarkable increase of the angle close to the resonance peak of the standard frequency response curves. However, in contrast, when comparing the respective nose solution branches only for $n = 1$, a noticeable increase at the turning point can be observed. Similar to the case of the condition number of the extended Jacobian, we attribute these two frequency-wise "local" increases of the Jacobian row vector angles to including the additional column vector of $J_{\Omega,n}$ in the extended Jacobian. However, it would require further investigation in order to answer why this is only observed in such a local manner for the standard solution branch. Upon returning to the original question of artifact detection capabilities, we conclude that neither of the two discussed measures is suitable for detecting artifacts since for $n = 1$ or $n = 7$ there is no observable change of the minimal (extended) Jacobian angle at $\Omega = 0.21$, i.e. the frequency value of the turning point of the reference solution branch.

4.1.3. Number of solutions

Next, we investigate the number of solutions, as defined in Section 3.3, as a potential artifact measure. For this, consider the frequency range partition $\mathbb{F} = \mathbb{F}_A \cup \mathbb{F}_B \cup \mathbb{F}_C$, as introduced in Section 2.3, which we obtained by identifying the turning points of the frequency response curves of the ansatz order $n = 1$ and the reference ansatz order $n_{ref} = 91$. Upon counting the number of solutions over each frequency range, $\mathbb{F}_A, \mathbb{F}_B, \mathbb{F}_C$ yields for $n = 1$

$$|\Gamma_1(\mathbb{F}_A \cup \mathbb{F}_B)| = 3 \qquad \text{and} \qquad |\Gamma_1(\mathbb{F}_C)| = 1$$

and for $n_{ref} = 91$

$$|\Gamma_{91}(\mathbb{F}_A)| = 3 \qquad \text{and} \qquad |\Gamma_{91}(\mathbb{F}_B \cup \mathbb{F}_C)| = 1$$

Comparing the number of solutions of both ansatz orders, yields the differences

$$|\Gamma_{91}(\mathbb{F}_A)| - |\Gamma_1(\mathbb{F}_A)| = 0 \qquad |\Gamma_{91}(\mathbb{F}_C)| - |\Gamma_1(\mathbb{F}_C)| = 0 \qquad |\Gamma_{91}(\mathbb{F}_B)| - |\Gamma_1(\mathbb{F}_B)| = 2$$

That is, on $\mathbb{F}_A$ and $\mathbb{F}_C$ the number of solutions matches, but not on $\mathbb{F}_B$ since there is a difference of two. This is exactly where the above introduced artifact solutions can be observed. However, this approach has two disadvantages. First, it is an *a postiori* measure, i.e. computation of two frequency responses of different ansatz orders is required. Second, the way we presented the counting of solutions requires to count for *all* frequencies $\Omega \in \mathbb{F}$ which is, of course, not feasible in finite precision. Instead, either a sampling of the frequency range $\mathbb{F}$ or a comparison of the frequency components of all points of the two sets $\Gamma_1(\mathbb{F})$ and $\Gamma_{91}(\mathbb{F})$ subject to a given frequency tolerance $\varepsilon_\Omega > 0$ would be required. However, this approach is not likely to be numerically robust, which is why we consider it to be of rather academic nature.

4.2. Geometric measures

In this part, we discuss if the two normalized distance measures introduced in Section 3.4 applied to the test case can be used as artifact solution identifiers. Since both the normalized arclength distance and the normalized discrete Fréchet distance require the arclength of the solution branches we start by considering convergence of the arclength, as depicted in Fig. 9a. It shows the arclength of the standard and nose branch over the approximation order n. Apparently, the arclength of the standard branch B_n^s converged quite quickly to a value of $L(B_{91}^s) = 2.16$ with an error $|L(B_{91}^s) - L(B_{75}^s)| < \varepsilon = 10^{-14}$. In contrast, the arclength of the nose branch B_{91}^n converged noticeably slower to a value of $L(B_{91}^n) = 0.77$ with an error $|L(B_{91}^n) - L(B_{75}^n)| \approx 1.03 \cdot 10^{-2}$.

Next, we discuss the results of the normalized arclength distance $\overline{d}_L$ and the normalized discrete Fréchet distance $\overline{d}_{DF}$ plotted over the approximation order n, as depicted in Fig. 9b. For both distance measures, it can be observed that, again, the standard branch converges noticeably faster than the nose branch. For this reason, we choose a linear scale of the diagram for both distance measures in order to be able to better compare the qualitative behavior of each of the curves. Upon comparing the standard branch of ansatz orders $n = 75, 91$, the distance measures yield $\overline{d}_L(B_{75}^s, B_{91}^s) < 10^{-14}$ and $\overline{d}_{DF}(B_{75}^s, B_{91}^s) = 1.14 \cdot 10^{-13}$. However, for the nose branch, the two distance measures yield noticeably larger errors of $\overline{d}_L(B_{75}^n, B_{91}^n) = 1.32 \cdot 10^{-2}$ and $\overline{d}_{DF}(B_{75}^n, B_{91}^n) = 8.08 \cdot 10^{-3}$. Note that for $n = 1$, the value of the two distance measures of the nose branch are noticeably larger compared to the values of the standard branch, i.e. $\overline{d}_L(B_1^n, B_{91}^n) = 0.49$ versus $\overline{d}_L(B_1^s, B_{91}^s) = 0.003$ and $\overline{d}_{DF}(B_1^s, B_{91}^s) = 0.65$ versus $\overline{d}_{DF}(B_1^s, B_{91}^s) = 0.03$. This amounts to a difference of roughly one order of magnitude. Since both distance measures are normalized by the arclength of the respective reference solution branch, this difference is noticeable. However, it is not yet clear if this is characteristic behavior

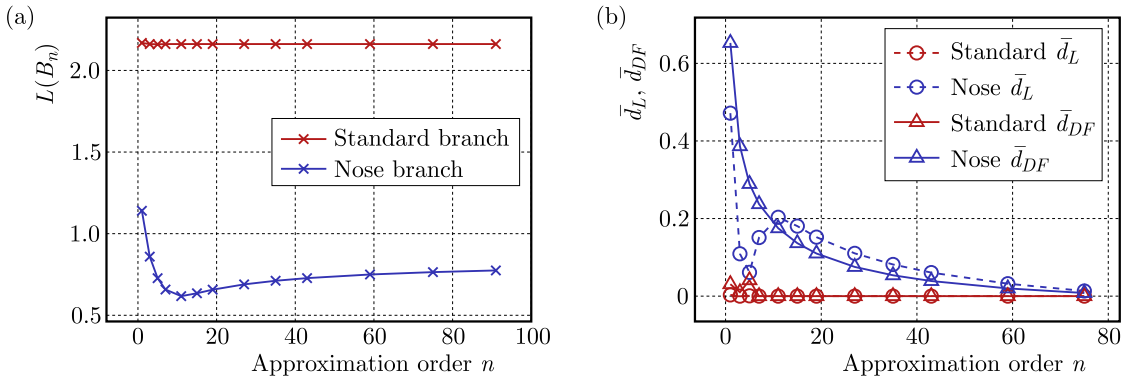


Fig. 9. Geometric measures for the test case (2.9): (a) arclength $L(B_n)$, (b) normalized arclength distance $\overline{d}_L(B_n, B_{n_{ref}})$ and normalized discrete Fréchet distance $\overline{d}_{DF}(B_n, B_{n_{ref}})$

of the artifact solutions. At this point, further studies for different parameter sets for the Duffing system might provide deeper insight. Additionally, consideration of rather large deviations in the amplitudes for frequencies $\Omega < \Omega_{1,tp}$ (cf. Fig. 1) suggests that this is also contributing to somewhat large normalized distance measures of the nose. One would have to filter out the amplitude part of the errors w.r.t. the normalized distances in order to better assess if these distance measures are suitable artifact solution identifiers. A possible way to get the frequency part of the difference of two curve points $A - B$ w.r.t. the Fréchet distance would be to modify the Euclidean norm $\|A - B\|_2$ to a weighted norm, i.e. $\|\mathbf{W}(A - B)\|_2$ with the diagonal matrix $\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix}$ for $0 < \epsilon \leqslant 1$. Computing either of the above distance measures might be more expensive than identifying artifact solutions by the turning points of two compared solution branches within Definition 1. However, a conclusive complexity analysis is yet missing.

4.3. Solver measures

In this last part, we seek to investigate if the solver-related quantities allow for a detection of artifact solutions. For this, we focus on the number of correction steps k per prediction step of the employed numerical continuation method which is depicted in Fig. 10. This figure shows the

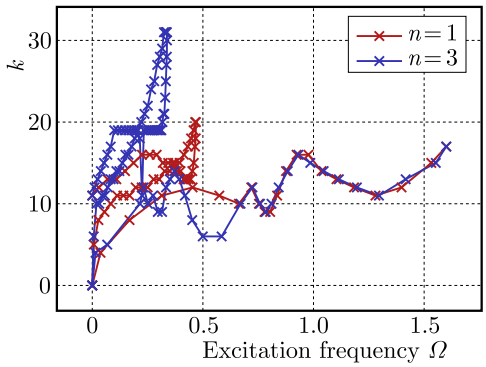


Fig. 10. Number of correction steps k per excitation frequency for the test case (2.9)

number of correction (Newton iteration) steps k over the entire frequency range of test case (2.9). Here, we only plotted the curves for the ansatz order $n = 1, 3$ to not clutter the diagram. For the ansatz order $n = 1, 3$, the nose branch exhibits values of k in the range 0 to 20 and 0 to 31, and the standard branch values of 0 to 17 and 0 to 18, respectively. Since the employed numerical continuation method starts at $\Omega = 0$ with pre-computed initial guesses $\mathbf{c}_n^0$, the canonical values

of $k = 0$ at this frequency can be observed. Additionally, the smallest, largest and average number of correction steps over all ansatz orders $n = 1, 3, \ldots, 91$ and excitation frequencies are 0, 15.2 and 41, respectively. Similar to the diagrams of the condition number in Fig, 7a and 7b, there is a noticeable peak in the number of correction steps at the nose branch turning point. However, in the characteristic frequency range around the turning point of the reference solution at $\Omega = 0.21$, neither for $n = 1$ nor for $n = 3$, there is a noticeable change of the number of correction steps. Hence, this measure is also not indicative for the studied artifact solutions. We end this Section by presenting the solver-related quantities computation time t_{comp} in seconds as well as the number of prediction steps k_{pred}, the total number of correction steps k_{corr} and the average number of correction steps $\overline{k}_{corr}$, all per solution branch and against the approximation order n in Fig. 11. All computations were performed on a 64 bit `Ubuntu 22.04.03 LTS` operating

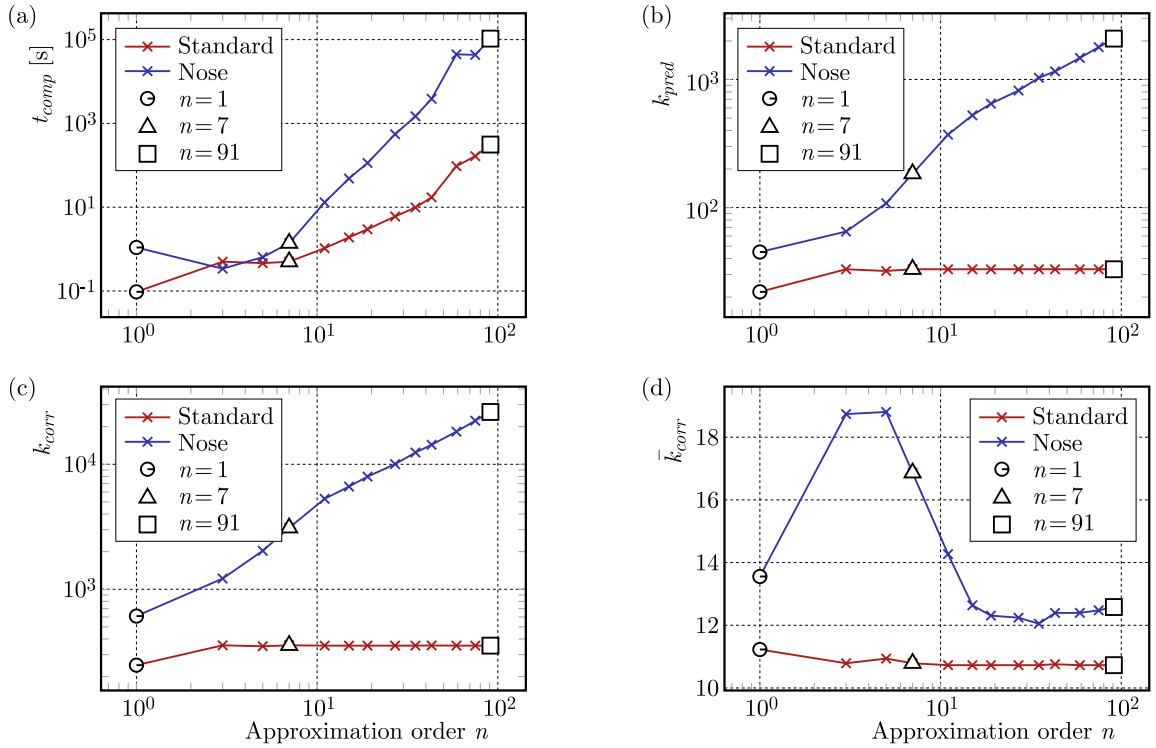


Fig. 11. Computation time t_{comp}, number of prediction steps k_{pred}, total number of correction steps k_{corr} and average number of correction steps $\overline{k}_{corr}$ for the test case (2.9)

system with an AMD Ryzen 7 Pro 4750U CPU and 32 GB of RAM. For the computation time, an expected exponential increase upon the increase of the ansatz order n is observed. The noticeable outliers of the nose branch at $n = 1, 59$ are attributed to an additional computational load of the operating system. The number of prediction steps k_{pred} and the total number of correction steps k_{corr} of the nose branch exhibit an almost exponential increase as well. However, for the standard branch, the number of prediction steps k_{pred} and the total number of correction steps k_{corr} already converged for $n \geqslant 3$ around values of 31 and 350, respectively. Finally, consider the average number of correction steps $\overline{k}_{corr}$. For the nose and standard branch, this value converges to values around 12.4 and 11.2, respectively. Interestingly, for lower ansatz orders, both solution branches exhibit a larger average number of correction steps compared to the value for the reference ansatz order of $n = 91$. In particular, over the course of the ansatz order increase from $n = 7$ to $n = 91$, the nose branch exhibits a decrease of $\overline{k}_{corr}$ by about a third.

5. Conclusions

In this work, we discuss several qualitative error measures in order to characterize the so-called artifact behavior that occurs during computation of HBM solutions for the softening Duffing oscillator. In particular, we provide a mathematical definition of artifact solutions in which the turning points of two solution branches of different ansatz orders are compared. This allows for an *a posteriori* identification of artifact solutions based solely on a robust computation of turning points. Additionally, of the residual, geometric, algebraic and solver-related error measures, investigating only the approach of counting the number of computed solutions, yields the desired discontinuity at the frequency value $\Omega = 0.21$ of the turning point of the reference solution branch. However, this *a posteriori* approach is of rather academic nature and expected to be not robust in the numerical practice. Furthermore, unfortunately, none of the examined error measures potentially showed to be *a priori* indicative of artifact behavior but only *a posteriori*. A possible explanation for the lack of an artifact-related characteristic behavior of the investigated measure lies in the fact that up to now, static error measures have been considered. This means that the value of quantities under examination was always evaluated for a specific order of development only. What remained unconsidered is the dependence of the error measures on the rate of change of the truncation order. Additionally, further studies are required to connect the concept of artifact solutions with the existing error measures such as, e.g., Urabe (1965), Kogelbauer and Breunung (2021), Woiwode and Krack (2023). Consequently, the following question may be raised: Do artifact solutions exist for truncation orders that can be deemed "sufficiently large" as by the measures of the aforementioned authors? Furthermore, the authors plan to publish further studies on their Python codes for the HBM algebraic system generation and the employed numerical path continuation solver. Among others, the presented definition of solution artifacts as a theoretical foundation as well as the Fréchet distance between solution branches of different truncation orders should be further studied as *a posteriori* artifact identifiers. In this context, application to different Duffing parameter sets as well as other nonlinear systems needs investigation to further assess the robustness in numerical practice.

Acknowledgements

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number LE 4694/1-1. The authors wish to thank Volker Mehrmann and the members of his group for their advice during many helpful discussions.

References

1. Alt H., Godau M., 1995, Computing the Fréchet distance between two polygonal curves, *International Journal of Computational Geometry and Applications*, **5**, 75-91
2. Basu S., Pollack R., Roy M., 2006, *Algorithms in Real Algebraic Geometry*, 2nd ed., Springer Berlin, Heidelberg
3. De Terán F., Dopico F.M., Pérez J., 2013, Condition numbers for inversion of Fiedler companion matrices, *Linear Algebra and its Applications*, **439**, 4 944-981
4. Deuflhard P., 2011, *Newton Methods for Nonlinear problems: Affine Invariance and Adaptive Algorithms*, Springer Series in Computational Mathematics, Springer Berlin, Heidelberg
5. Deuflhard P., Fiedler B., Kunkel P., 1987, Efficient numerical pathfollowing beyond critical points, *SIAM Journal on Numerical Analysis*, **24**, 4, 912-927
6. Deuflhard P., Hohmann A., 2019, *Eine algorithmisch orientierte Einführung*, De Gruyter, Berlin, Boston

7. Duffing G., 1918, *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*, Samlung Vieweg
8. Edelman A., Murakami H., 1995, Polynomial roots from companion matrix eigenvalues, *Mathematics of Computation*, **64**, 763-776
9. Ferri A.A., Leamy M.J., 2009, Error estimates for harmonic-balance solutions of nonlinear dynamical systems, *Collection of Technical Papers – AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, May, 1-15
10. Figueira J.P., 2023, *Discret-frechet*, https://github.com/joaofig/discrete-frechet
11. García-Saldaña J.D., Gasull A., 2013, A theoretical basis for the Harmonic Balance Method, *Journal of Differential Equations*, **254**, 1, 67-80
12. Hagedorn P., 1981, *Non-linear Oscillations*, Clarendon Press, Oxford and New York
13. Herman R.L., 2016, *An Introduction to Fourier Analysis*, CRC Press
14. Holmes P.J., Rand D.A., 1976, The bifurcations of Duffing's equation: An application of catastrophe theory, *Journal of Sound and Vibration*, **44**, 2, 237-253
15. Kogelbauer F., Breunung T., 2021, When does the method of harmonic balance give a correct prediction for mechanical systems, *Applicable Analysis*, **102**, 2, 425-443
16. Kovacic I., Brennan M.J., 2011, *The Duffing Equation: Nonlinear Oscillators and their Phenomena*, Wiley
17. Krack M., Gross J., 2019, *Harmonic Balance for Nonlinear Vibration Problems*, Springer
18. Lentz L., von Wagner U., 2020, Avoidance of artifacts in harmonic balance solutions for nonlinear dynamical systems, *Journal of Theoretical and Applied Mechanics*
19. Nayfeh A.H., Mook D.T., 1979, *Nonlinear Oscillations*, Wiley
20. Novak S., Frehlich R.G., 1982, Transition to chaos in the Duffing oscillator, *Physical Review A*, **26**, 6, 3660-3663
21. Strogatz S.H., 1994, *Nonlinear Dynamics and Chaos*, Westview Press
22. Ueda Y., 1991, Survey of regular and chaotic phenomena in the forced Duffing oscillator, *Chaos, Solitons and Fractals*, **1**, 3, 199-231
23. Urabe M., 1965, Galerkin's procedure for nonlinear periodic systems, *Archive for Rational Mechanics and Analysis*, **20**, 120-152
24. Von Wagner U., Lentz L., 2016, On some aspects of the dynamic behavior of the softening Duffing oscillator under harmonic excitation, *Archive of Applied Mechanics*, **86**, 1383-1390
25. Von Wagner U., Lentz L., 2018, On artifact solutions of semi-analytic methods in nonlinear dynamics, *Archive of Applied Mechanics*, **88**, 1713-1724
26. Von Wagner U., Lentz L., 2019, On the detection of artifacts in Harmonic Balance solutions of nonlinear oscillators, *Applied Mathematical Modelling*, **65**, 408-414
27. Woiwode L., Balaji N.N., Kappauf J., Tubita F., Guillot L., *et al.*, 2020, Comparison of ANM and predictor-corrector method to continue solutions of harmonic balance equations, [In:] *Conference Proceedings of the Society for Experimental Mechanics Series*
28. Woiwode L., Krack M., 2023, Are Chebyshev-based stability analysis and Urabe's error bound useful features for Harmonic Balance?, *Mechanical Systems and Signal Processing*, **194**, 110265

Manuscript received February 14, 2024; accepted for print February 29, 2024

INFORMATION FOR AUTHORS

Journal of Theoretical and Applied Mechanics (***JTAM***) is devoted to all aspects of solid mechanics, fluid mechanics, thermodynamics and applied problems of structural mechanics, mechatronics, biomechanics and robotics. Both theoretical and experimental papers as well as survey papers can be proposed.

JTAM accepts full-text articles (max. 12 pages) as well as the short communications with all the requirements concerning standard publications, except a volume that is limited to 4 pages.

We accept articles in English only. The text of JTAM paper should not exceed 12 pages of standard format A4 (11-point type size, standard margins – 2.5 cm, single line spacing) including abstract, figures, tables and references.

The material for publication should be sent to the Editorial Office via electronic journal system: http://www.editorialsystem.com/jtam

Papers are accepted for publication after the review process. Blind review model is applied, which means that the reviewers' names are kept confidential to the authors. Reviewer(s) declare that there is no interpersonal relation with the author(s) that would affect the opinion and recommendation of the article for publication in JTAM. The final decision on paper acceptance belongs to the Editorial Board.

Starting from January 1, 2020, the Publisher of *Journal of Theoretical and Applied Mechanics* introduces a fee for published articles.

This applies only to papers submitted after this date and accepted by the Editorial Board for publication.

A payment of 700 EUR will be a condition for commencing the editorial procedure for upcoming articles.

After qualifying your paper for publication we will require LaTeX or TeX or Word document file and figures.

The best preferred form of figures are files obtained by making use of editorial environments employing vector graphics.

Requirements for paper preparation

Contents of the manuscripts should appear in the following order:

- Title of the paper.
- Authors' full name, affiliation and e-mail.
- Short abstract (maximum 100 words) and 3-5 key words (1 line).
- Article text (equations should be numbered separately in each section; each reference should be cited in the text by the last name(s) of the author(s) and the publication year).
- References (maximum 25) in alphabetical order.
- Titles of references originally published not in English, should be translated into English.

All the data should be reported in SI units.

Cena 40,00 zł (VAT 8%)

Contents