Investigation of the Impact of Surface Roughness, on a Ship's Drag (Hull Resistance)

Zainab Ali, Gabriella Bognár

University of Miskolc, Institute of Machine and Product Design, Egyetemváros, 3515 Miskolc, Hungary, e-mail: zainab.ali@student.uni-miskolc.hu, gabriella.v.bognar@uni-miskolc.hu

Abstract: Recently, there has been an increasing focus on maritime transport, as it offers many advantages in terms of storage and transport. As a result, shipping companies need to reduce the fuel consumption of their vessels. These companies have tried to define methods of operation and maintenance in order to reduce greenhouse gas emissions and also to reduce operating costs, thus increasing company profits. One important parameter that directly affects speed, power requirements, and fuel consumption is the hull resistance. Computational Fluid Dynamics (CFD) can be used to calculate the resistance of a rough surface using special wall functions that take into account the effect of roughness on the boundary layer near the hull. These results can be compared with those of a smooth surface. In addition to the effect of surface roughness on hull resistance to pressure, this method also allows the combination of roughness and non-linear effects such as the spatial distribution of contaminants, the movement of the ship in waves, and the effect of thrust on hull resistance. Accordingly, the aim of this research is to determine the effect of surface roughness on the ship resistance for different values of roughness height, boundary layer, and values of velocity, pressure, and kinetic energy fields for the KVLCC2 model hull by CFD using the RANS equations and the k-ω SST model. A numerical study was performed to determine how surface roughness affects the velocity field and kinetic energy.

Keywords: KVLCC2; ship hull roughness; velocity; pressure; kinetic energy; CFD; k-ωSST

1 Introduction

When the design and calculations for a ship, one of the most important things is to know the conditions under which the propeller will operate, such as the speed, pressure, kinetic energy and vortices. This is a very important parameter for predicting the thrust that the propeller can produce. Accordingly, the effect of the surface roughness on these two fields, in the working plane of the propeller, should be investigated. It is, therefore, necessary to understand and analyze the variations of the boundary layer around the ship, during motion, as this is the most important factor for studying the flow around the hull.

Prandtl [1] and Prandtl [2] defined the concept of a boundary layer as a thin zone near the surface of a body in a flowing fluid. A proper description of the physical processes taking place in the boundary layer between a fluid and a solid play an important role in fluid mechanics problems.

One of the foremost considerations in ship design and calculation, involves acquiring comprehensive knowledge about the propeller's operational conditions, encompassing factors such as velocity, pressure, kinetic energy, and vortices. The velocity field and kinetic energy levels at which the propeller functions assume the utmost significance in estimating the resultant thrust force. Consequently, it becomes imperative to investigate the influence of surface roughness on these two fields within the operational plane of the propeller. Thus, comprehending and scrutinizing the boundary layer that develops around the ship during its motion becomes pivotal, as it represents the principal determinant in studying the flow dynamics encircling the ship's hull.

Recently, several papers have been published on the analysis of the effect of surface roughness on flow parameters. Song et al. [3] conducted a numerical study on the effect of heterogeneous hull roughness on ship resistance and developed a URANS-based CFD model using the modified wall function approach. The predicted total resistance coefficients for different hull conditions were compared with experimental data from Song et al. [4] which showed a convincing agreement, where the highest error was around 6.1% for C_T of the Wigley hull with $1/4$ bow-rough and $1/4$ aft-rough conditions.

Similar observations were made by Song et al. [4] who related observations on the effects of heterogeneous hull roughness to the distribution of local wall shear stress and the roughness Reynolds number. The results showed that local differences in wall shear stress led to different roughness Reynolds numbers and hence different roughness effects depending on the location of hull roughness. The hypothesis of Song et al. [4] was confirmed in this study. Consequently, the numerical approach presented in this study can be applied to predict the effect of heterogeneous roughness on propeller propellers.

Reynolds Averaged Navier-Stokes (RANS) solvers, once developed only to evaluate the resistance of still water, have become increasingly complex, and the current generation now has unsteady-state capabilities. Shortly, the same numerical solver will be able to address problems of drag, sea-keeping, and maneuvering. CFD workshops on numerical ship hydrodynamics have been organized regularly since 1980 to assess the current state of CFD development and to set new goals [5].

In their paper, Tahara et al. [6] provided an overview of numerical methods and presented and discussed the results of traction and self-propulsion models of KRISO Container Ship (KCS), including a comparison with available experimental fluid dynamics (EFD) data. For the CFD model of a towed flat plate and a KRISO container ship (KCS) was prepared. In the wall function of the CFD model, the

roughness function of a previously created sand-grained surface was used to reproduce the roughness effect in the turbulent boundary layer. The output of the CFD simulations was then compared with the experimental results. The results showed a convincing agreement, in the case of the hull wave profile without the propeller, the maximum error observed was approximately 8%, indicating that the CFD approach accurately predicts the effect of roughness on the overall drag of the 3D hull. Finally, the effect of roughness on the different resistance components of the ship was investigated. The further evaluation took place at the CFD Workshop 2005 discussions in Tokyo, where both methods were presented.

This paper aims to analyze and interpret the variation of velocity, pressure, kinetic energy, and vorticity with surface roughness, which affect the operation of the propeller, by numerical simulation. The numerical calculations are performed by CFD with the choice of surface roughness function using the turbulent model for KRISO Very Large Crude Carrier no. 2 (KVLCC2) ship model.

2 Wall Functions

The large gradients in velocity, pressure, and kinetic energy can be handled in CFD either by direct solution or by using wall functions. Numerous experiments have been carried out to study and determine the properties of turbulent flows, particularly in the boundary layer, and these experiments have shown that the dimensionless curve of the velocity distribution (y^+, U^+) can be described by a nearly identical formula in all cases, (see Fig. 1).

In the turbulent boundary layer, the hydrodynamic effects can be expressed by the non-dimensional mean velocity profile [8], [9]

$$y^+ = f\left(U^+\right) \tag{1}$$

where U^+ is the non-dimensional velocity profile in the boundary layer, y^+ is the non-dimensional distance measured perpendicular to the surface. These parameters are defined according to the following two equations

$$U^+ = \frac{U}{U_\tau} \tag{2}$$

$$y^+ = \frac{yU_\tau}{\upsilon} \tag{3}$$

where U is the average fluid velocity, U_τ the frictional velocity is defined by the relation $\sqrt{\tau_w/\rho}$, y the wall distance, υ the kinematic viscosity, τ_w the shear stress at the wall, ρ the fluid density.

The range $0 < y^+ < 5$ is the linear range, where the effect of viscosity dominates and the equation $U^+ = y^+$ is satisfied. In the range $5 \le y^+ \le 70$, the hybrid region, a transition from a linear relationship between U^+ and y^+ to a logarithmic relationship occurs. Within the range $5 \le y^+ \le 30$, both viscous and turbulent stresses are in equilibrium, and the linear relationship between U^+ and y^+ is preserved. At $y^+ = 30$, the logarithmic range in which the dominance of turbulence is satisfied begins, and the velocity distribution from this value is given as described in [7].

$$U^+\left(y^+\right) = \frac{1}{k_{const}} \ln y^+ + B; \qquad 5 \le y^+ \le 30 \tag{4}$$

where B is a constant that considers the effect of surface roughness on the velocity distribution, and according to many experiments, the value $B = 5$ is the best value for it and $k_{const} = 0.41$ is Kármán's constant [10].

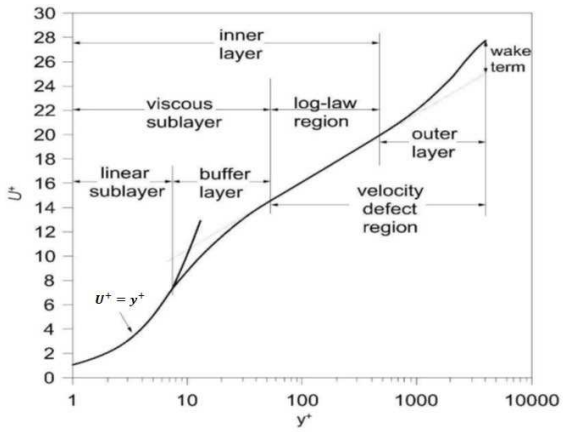


Figure 1
The non-dimensional curve of the velocity distribution in the turbulent boundary layer [7]

Surface roughness causes an increase in turbulence, which increases shear stresses at the wall and turbulence stresses, both of which reduce velocity. Roughness can be broadly divided into *k*-type and *d*-type roughness, which are the two most common types. This categorization is determined by the roughness functions used.

The primary parameter for the *k*-type roughness functions is the roughness height *k*, while the primary parameter for the *d*-type roughness functions is the pipe diameter [10]. Research on how surface roughness affects the turbulent boundary layer near the surface is of fundamental importance and has been studied since 1993 and is still ongoing, and the most important studies in this area, which we recommend the reader to review, include [10-20]. In this paper, we consider the roughness of *k*-type, since it has been shown that hull roughness is of *k*-type [21], and henceforth the term roughness is used to mean roughness of type *k*. In addition to other factors that may be used to determine roughness, the primary parameter is the roughness height *k* or the roughness height of equivalent sand k_s . The Reynolds number for roughness, which is a non-dimensional quantity, can take the place of the roughness height, and it is given by the following relationship see [21] [22]

$$k^+ = \frac{kU_\tau}{\upsilon} \tag{5}$$

The type of flow on the surface is defined according to the Reynolds number of roughness, according to this classification there are three types of flow regimes, the transiently rough regime, the fully rough regime, and the hydraulically smooth regime.

Although the same Reynolds number of roughness is recorded, it should be remembered that different types of roughness may produce different flow regimes on the surfaces [23]. For example, Schlichting [24] stated that if the surface roughness is isotropic sand grains and the value of $k^+ < 5$, then the prevailing flow regime is of the soft hydraulic type, and it turns into the coarse transitional in the range of values $5 \le y^+ \le 70$, and becomes fully coarse (the coarse and fully developed regime) when $k^+ > 70$.

Extending Nikuradse's research from 1933 [25], Schlichting used the following equations to depict the velocity profile in the turbulent boundary layer of the rough tube [24]

$$\frac{U_p U^*}{\tau_w/\rho} = \frac{1}{k}\ln\left(E\frac{U^+ y_p}{\mu}\right) - \Delta B;\ U^+ = C_\mu^{1/4} k^{1/2},$$
$$\Delta B = \frac{1}{k}\ln f_r.U_p \tag{6}$$

$$U^+ = \frac{1}{k_{const}}\ln\left(\frac{y}{k_s}\right) + B_2 \tag{7}$$

where $k_{const} = 0.41$ and the coefficients B_1, B_2 have various values for various flow regimes.

From the above it can be observed how roughness affects the flow in the velocity profile, where it causes a decrease in the velocity diagram in the logarithmic domain, this decrease is called the roughness function, and ΔU^+.

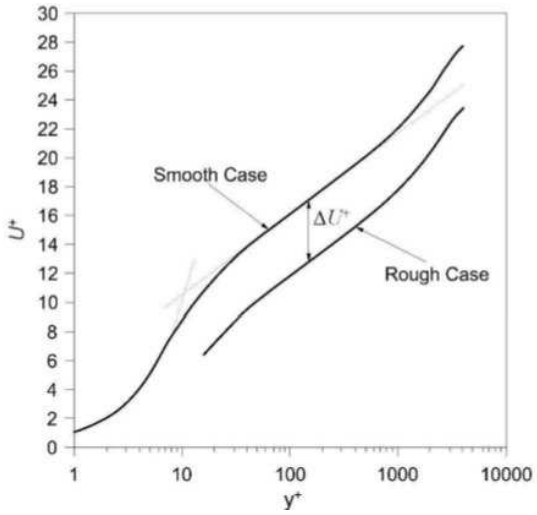


Figure 2
The effect of roughness on the velocity profile in the boundary layer region [10]

As has been previously demonstrated, the velocity profile region of the boundary layer region is where surface roughness has the greatest impact on the flow [9]. The surface roughness causes the region of complete perturbation to shift ΔU^+ (log-law region) downward in the (y^+, U^+) plot, see Fig. 2. As a result of these roughness-related variations in the velocity profile in the boundary layer region, frictional resistance increases [27] and the velocity profile, in this case, is given by the Eq. (8)

$$U^+ = \frac{1}{k}\ln\left(y^+\right) + B - \Delta U^+ \tag{8}$$

where ΔU^+ is the velocity obtained in a profile due to the roughness (velocity profile). By eliminating the expression ΔU^+ from the function provided by the equation, it is possible to represent the roughness velocity profile Eq. (1) and in the case of a smooth surface, it represents the velocity profile.

It is important to keep in mind that ΔU^+ simply disappears from Eq. (8) and this equation is transformed into Eq. (4) in the case of a smooth (without roughness) surface. Since there is no single roughness function that accounts for all types of roughness, the values of ΔU^+ are often determined empirically.

Here, it must be noted, that the wall function proposed by Demirel et al. [27] considered the effect of both coating and fouling. This wall function has been used in several reference studies, including but not limited to the study by Owen et al [28], where this function was used to investigate how performance is affected by surface roughness caused by coating and contamination. The results were very convincing. (The error in open water efficiency compared to the experimental result is 1.93%).

Given that the program that will be used in this work is the ANSYS program, which uses a roughness function that combines the characteristics of almost all roughness functions mentioned in the references so far for roughness with the modified basic wall law given by the following relation is used to incorporate the effects near the wall to simulate turbulent flow where the effect of wall roughness is most significant. Then, [29]:

$$\frac{U_p U^*}{\tau_w/\rho} = \frac{1}{k}\ln\left(E\frac{U^* y_p}{\mu}\right) - \Delta B \tag{9}$$

where $U^* = C_\mu^{1/4} k^{1/2}$ and $\Delta B = \frac{1}{k}\ln f_r U_p$ is the non-dimensional flow velocity along the wall in the boundary layer, U^* the friction velocity, μ the dynamic viscosity, f_r is the roughness coefficient that determines the amount of interference due to roughness effects, ΔB depends mainly on (type of sand, network nodes, ribs, ...) and the roughness size.

There isn't a particular roughness coefficient that corresponds to all different kinds of roughness, but ΔB is related to the non-dimensional roughness height k_s^+, where k_s^+ is the physical roughness height. It takes different forms:

- $k_s^+ \le 2.25$ hydrodynamically smooth running
- $2.25 \le k_s^+ \le 90$ transitional flow
- $k_s^+ > 90$ completely rough region

According to numerical data, we find that the effects of roughness are very small in the hydrodynamically smooth system, but they are more important in the transitional system, and greatly affect the system with a completely rough region.

The previous three roughness regimes were split in ANSYS Fluent, and the formulas proposed by Cebeci and Bradshaw [33] using Nikuradse's data [25] were used to determine ΔB for each regime as follows:

- For a hydrodynamically smooth system $k_s^+ \le 2.25$ and

  $$\Delta B = 0$$

- For a transitional system $2.25 \le k_s^+ \le 90$ and

  $$\Delta B = \frac{1}{k}\ln\left[\frac{k_s^+ - 2.25}{87.75} + C_s k_s^+\right]\sin\left[0.4258\left(\ln k_s^+ - 0.811\right)\right] \qquad (10)$$

  where C_s is the roughness constant (it is determined depending on the roughness type).

- For a system with absolute roughness $k_s^+ > 90$ and

  $$\Delta B = \frac{1}{k}\ln\left(1 + C_s k_s^+\right) \qquad (11)$$

The logarithmic velocity profile is shown by the downward slope in Fig. 3 below:

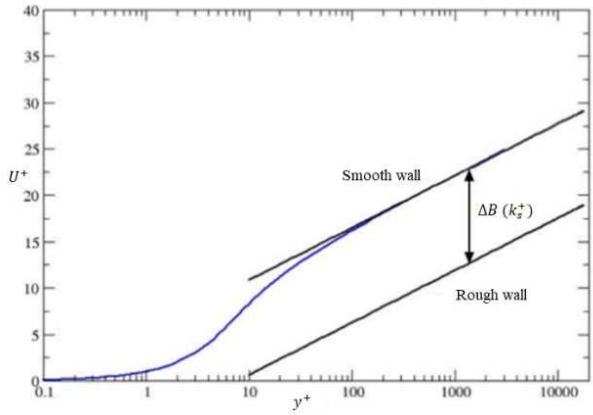


Figure 3
Downward regression of the logarithmic velocity profile [29]

Using ANSYS Fluent, multiple techniques can be employed depending on the disturbance model and near-wall processing to circumvent this issue, namely [29]:

1. Roughness height reduction as y^+ decreases. This method is to redefine the roughness height based on network optimization.

$$k_s^+ = \min\left(k_s^+, y^+\right) \tag{12}$$

This ensures that as y^+ approaches zero, k_s^+ approaches zero as well. Therefore, the grid requirement for the rough walls in this case is $y^+ > k_s^+$ in order to ensure the full effect of the roughness on the runoff.

2. Physical change of the wall. The second approach is based on the finding that the viscosity effect region is exclusively restricted to areas near smooth walls.

   The viscosity region in the rough flow is destroyed and the viscosity effects are minimal in the transitional roughness regime, when the roughness elements are somewhat thicker than the sublayer and start to impede it.

The second method (physical change of the wall) is a default treatment for rough walls in all ω-equation-based disorder models and the following ε-equation-based disorder models.

- Standard models, RNG, and the applicable k-ε model
- Reynolds stress models

When using regular wall functions and scalable wall functions, this method can be applied. More scalable wall functions can be used than regular wall functions. Other coarse wall models do not require specific calibration for fine meshes, such as the Spalart-Allmaras model. Therefore, the first method is used (decrease in roughness height as y^+ decreases).

3 Geometric Model and Boundary Conditions

The KVLCC2 ship model (the well-known KRISO Very Large Crude Carrier no. 2 model) was chosen for the calculations due to the abundance of experimental data. Since our study was performed at a low Froude number $F_r = 0.142$, space is missing the effect of free surface deformations was ignored.

Knowing that Froude number, a non-dimensional number, is a cross-sectional flow characteristic defined as the following relation:

$$F_r = \frac{v_{ship}}{gL_{wl}}$$

where v_{ship} : ship velocity ($m.s^{-1}$), g : acceleration of gravity ($m.s^{-2}$), and L_{wl} : length of the water line (m)

For assessing the mathematical equations and models pertinent to the standard case, a plate possessing analogous properties to the one employed in Schultz experiment [31] was specifically chosen.

3.1 Geometrical Dimensions and Boundary Conditions for the KVLCC2 Tanker Model

The essential geometric measurements of the KVLCC2 model are contrasted with those of the original ship in the accompanying table. The reduction ratio is $\lambda = 58$

Table 1 shows the dimensions of the KVLCC2 ship and its model and, Fig.4 shows the CAD model of KVLCC2 [30].

Table 1
The dimensions of the KVLCC2 ship and its model

Geometric dimension	Symbol	Full-scale KVLCC2	Model KVLCC2	Unit
Length between perpendiculars	L_{pp}	320	5.5172	m
Breadth (molded)	B_{molded}	58	1	m
Draft (molded)	T	20.8	0.3586	m
Blockage coefficient	C_b	0.8	0.8098	-
Wetted surface area without appendages	S	27194	8.0838	m^2
Displacement	V	312622	1.6023	m^3

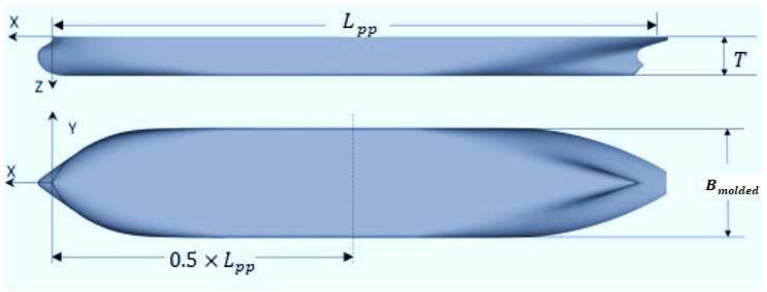


Figure 4
Carrier ship model [30]

Figure 5 shows the boundary conditions and ship location in the test channel, as follows:

$$L' \times B' \times T' = 4.615L_{pp} \times 2.885L_{pp} \times 1.5L_{pp}$$

where L', B', T' are the length, width, and height of the geometric domains for the ship.

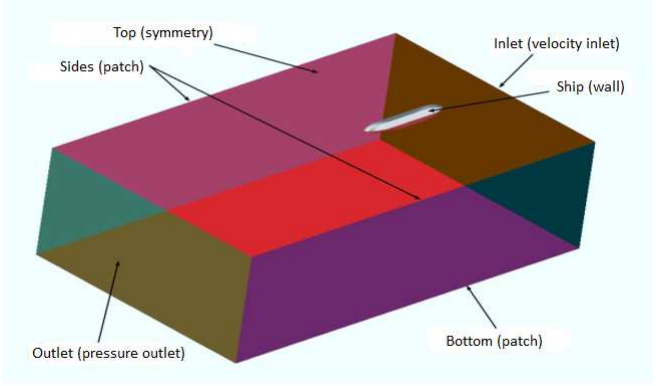


Figure 5
The boundary conditions around the ship

The ship is fixed and the fluid is moved at the same speed as the ship, achieving the Reynolds number, at which the vessel operates. The fluid inlet is the Inlet surface, and the outlet is the Outlet surface. The density is 998.2 kg/m^3, the viscosity is 0.001003 kg/m^{-4} and the ratio of specific heat is 1.4.

3.2 Geometrical Dimensions and Boundary Conditions for the Plate

Regarding the calibration plate, Table 2 shows the dimensions of Schultz plate and Fig. 6 shows it [31].

Table 2
The dimensions of the plate [31]

Geometric dimension	Symbol	Value	Unit
Length	L_{plate}	1.52	m
Breadth	B_{plate}	590	mm
Hight	T_{plate}	3.2	mm
Edge turning radius	r_{plate}	1.6	mm

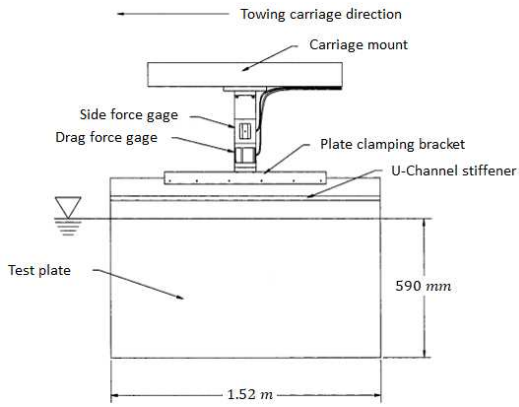


Figure 6
Validation plate [31]

Figure 7 illustrates the geometric domain used in the case of the plate has the geometric dimensions defined according to Schultz's experiment [31] as follows:

$$L"\times B"\times T" = 6L_{plate} \times 1.5L_{plate} \times 3.5L_{plate}$$

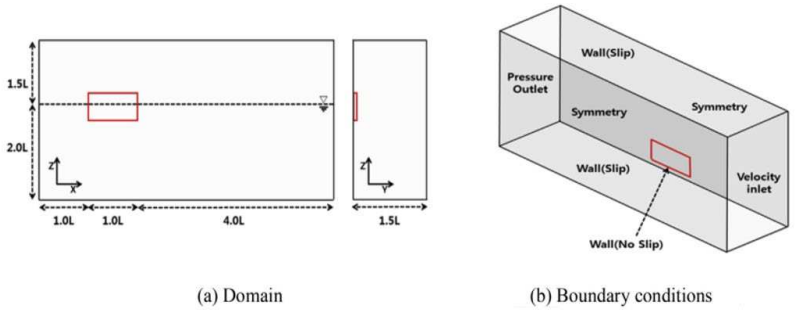


Figure 7
Computational domain and boundary conditions around the plate

4 The Fluid Flow Model

The Navier-Stokes equations and the conservation of mass equation explain the motion of incompressible Newtonian fluids. Turbulence and flow are described by four equations. Using CFD and ANSYS 15, solutions to the system of equations were obtained.

Menter [32] put forth the *k-ω* SST (SST-Shear Stress Transport) model. The *k-ω* and *k-ε* models are combined.

In this model, the benefits of the two models are integrated. While the *k-ε* model is utilized outside the boundary layer region in the free-flow area, the *k-ω* model is applied close to the wall within the boundary layer region. The following are the transfer equations for the disturbance rate ω and the disturbed kinetic energy *k* [32-36]:

- ***k* – equation**

$$\frac{\partial k}{\partial t}+u_j\frac{\partial k}{\partial x_i}=P_k-\beta^* k\omega+\frac{\partial}{\partial x_j}\left[\left(v+\sigma_k v_T\right)\frac{\partial k}{\partial x_j}\right],i,j=1,2,3,.. \quad (13)$$

- ***ω* – equation**

$$\frac{\partial \omega}{\partial t}+u_j\frac{\partial \omega}{\partial x_j}=\alpha S^2-\beta\omega^2+\frac{\partial}{\partial x_j}\left[\left(v+\sigma_\omega v_T\right)\frac{\partial \omega}{\partial x_j}\right]+2\left(1-F_1\right)\sigma_{\omega 2}\frac{1}{\omega}\frac{\partial k}{\partial x_i}\frac{\partial \omega}{\partial x_i} \quad (14)$$

- **Eddy viscosity μ_t -equation:**

$$\mu_t=\frac{\rho\alpha_1 k}{\max\left(\alpha_1\omega,\mathrm{SF}_2\right)},S\frac{\partial u}{\partial y'} \quad (15)$$

$$F_1=\tanh\left\{\left\{\min\left[\max\left(\frac{\sqrt{k}}{\beta^*\omega},\frac{500v}{y^2\omega}\right),\frac{4\sigma_{\omega 2}k}{CD_{k\omega}y^2}\right]\right\}^4\right\},$$

$$F_2=\tanh\left[\left[\max\left(\frac{\sqrt{k}}{\beta^*\omega y},\frac{500v}{y^2\omega}\right)\right]^2\right], \quad (16)$$

$$P_k=\min\left(\tau_{ij}\frac{\partial u_i}{\partial x_j},10\beta^*k\omega\right),$$

$$CD_{kw}=\max\left(2\rho\sigma_{\omega 2}\frac{1}{\omega}\frac{\partial k}{\partial x_i}\frac{\partial \omega}{\partial x_i},10^{-10}\right).$$

The blending function is used to change between the two turbulence models. The model coefficients are presented in the following Table 3.

Table 3
Coefficients of k-ω SST model [37]

β^*	α_2	β_1	σ_{k1}	σ_{k2}	$\sigma_{\omega 1}$	α_2	β_2	$\sigma_{\omega 2}$
0.09	0.555	0.075	0.85	1	0.5	0.44	0.0828	0.856

5 Mesh Generation

Figure 8 shows the structured grid employed in the numerical simulation, which was designed approximately with 482576 nodes in the case of a symmetric configuration. This grid was generated using ICEM program and was specifically oriented perpendicular to the ship's surface to ensure an accurate representation of the viscous flow field surrounding the ship. In terms of cell distribution, the bow, stern, and run sections of the ship were assigned a higher cell count compared to the central section. The minimum size of a cell ranged to be $2.1078 \times e^{-10} m^3$ to $4.0556 \times e^{-10} m^3$, depending on the inflow velocity. This varying cell size enabled finer resolution in regions of higher flow complexity and facilitated capturing the relevant flow physics with greater fidelity.

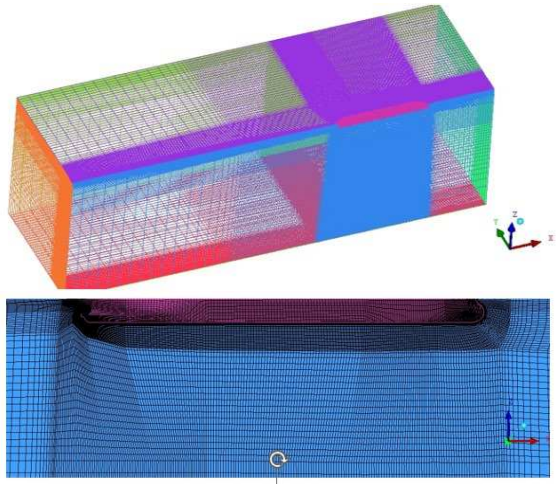


Figure 8
Structured grid around the KVLCC2 carrier (upper figure) and the refinement in the boundary layer region (lower figure)

The structured grid utilized in the numerical simulation of the calibration plate is shown in Fig. 9.

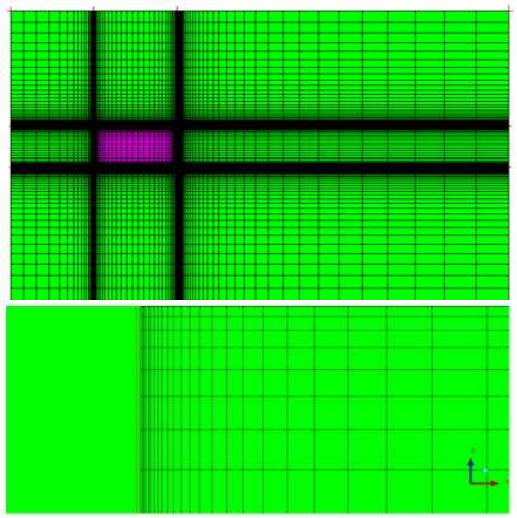


Figure 9
Structured grid around the plate (upper figure) and the refinement in the boundary layer region (lower figure)

This grid, designed for symmetric configuration, consisted of 1187640 nodes. The grid generation process was carried out using the ICEM program. To capture the flow characteristics near the plate surface accurately, a growth rate of 1.2 was chosen for grid refinement in the boundary layer region (see Fig. 9). This refinement strategy allowed for a more refined mesh resolution in the vicinity of the plate, ensuring enhanced fidelity in capturing near-wall flow phenomena. The structured nature of the grid facilitated efficient computational performance and maintained grid regularity, aiding in convergence and overall solution accuracy.

6 Solution Method

In our study, the finite element method (FEM) implemented within ANSYS serves as the primary numerical approach for accurately approximating the behaviour of our system.

To solve the continuity and momentum equations simultaneously for pressure and velocity instead of resorting to a pressure correction approach, coupled solver, with a Pressure-Based type, was implemented (which is a mix between a simple scheme and a PISO scheme). In terms of spatial discretization, least squares cell-based was chosen for the gradient and considered the pressure and the momentum as a second-order upwind while turbulent kinetic energy and specific dissipation ratio were considered as a first-order upwind.

The simulation was run under a steady-state setting, implying that the solution remains constant over time. The flow variables were also effectively and precisely initialized using the Hybrid initialization approach.

For the plate analysis, the pressure-velocity scheme was employed in conjunction with compressive volume fractions. To capture the gradient, a least squares cell-based approach was adopted, with a focus on - PRESTO! - pressure. As for the momentum, turbulent kinetic energy, and specific dissipation ratio, a second-order upwind scheme was utilized in a steady-state setting.

7 Validation Study

7.1 Assessment of Numerical Results for the Plate

The roughness function employed in ANSYS is currently undergoing study and development. Therefore, it is necessary to initially test this function on a standard case to ensure the reliability and accuracy of the results [31].

A comparison between the numerical and experimental results for plate total resistance coefficient C_T, considering various roughness height values (d), is presented in Table 4. The experimental analysis involved testing the plate at two Froude numbers (F_r). The results indicate that the error rate increases with higher Froude numbers and roughness heights.

KVLCC2 operates at a significantly lower Froude number of 0.142 [30] compared to the values listed in Table 4. Therefore, it is anticipated that the results will be even more accurate in this specific case.

Table 4
Comparison between EFD and CFD (Present) of total resistance coefficient

Velocity ($m.s^{-1}$)	F_r	$d\left(\mu m\right)$	$C_{T,EFD}$	$C_{T,CFD}$	**Error %**
2.0	0.518	0	0.003605	0.00359	0.4
		85	0.003663	0.00377	3.1
		129	0.003783	0.00397	5.1
3.8	0.984	0	0.003226	0.00320	0.8
		85	0.003423	0.00353	3.3
		129	0.003500	0.00369	5.4

It is observed that the maximum difference between the outcomes is 5.4% at a Froude number of 0.984 and a roughness height of 129 μm. This disparity is considered highly acceptable. Thus, it can be concluded that when employing the

conditions (roughness function, RANS equations, and the $k-\omega$ SST numerical model) for studying plate resistance, the results exhibit consistent agreement with the experimental data, both quantitatively and qualitatively. Consequently, these conditions were considered suitable for the vessel's circumstances and were implemented in the subsequent stage of analyzing KVLCC2.

7.2 Assessment of Numerical Results for KVLCC2

The experimental value ($C_{T,EFD}$) of the total ship resistance coefficient for the KVLCC2 ship is available at Froude number 0.142 [30]. Considering that the total resistance coefficient defined in the relationship,

$$C_T = C_d + C_w,$$

where C_d is drag coefficient (the sum friction coefficient C_f and the pressure coefficient τ_w), and C_w is wave resistance coefficient. Since the Freud number is relatively small then the resistance of waves is neglected and drag resistance is dominated.

Table 5 illustrates the comparison between the results of EFD and CFD (obtained by ANSYS) for the total resistance coefficient.

Table 5
Comparison between EFD and CFD (Present) of total resistance coefficient.

Velocity ($m.s^{-1}$)	F_r	$C_{T,EFD}$	$C_{T,CFD}$	Error %
1.047	0.142	0. 00411	0. 00394	4.1

Error percentage is calculated by the following relationship,

$$\frac{C_{T,EFD} - C_{T,CFD}}{C_{T,EFD}} \times 100.$$

The relative error 4.1%, refers to a convincing agreement between the experimental and numerical results.

8 Numerical Results and Discussion

The surface roughness of the hull has a notable impact on the propeller plane. When the hull surface is rough, it introduces disturbances and irregularities in the flow around the ship, particularly in the vicinity of the propeller.

These surface irregularities can lead to the formation of turbulent boundary layers and vortices in the flow. As the flow encounters these disturbances, it interacts with the propeller plane, causing several significant effects on velocity, pressure, kinetic energy, and vortex formation.

In this section, the surface roughness is defined by Eqs. (3) and (4), which applied in the numerical solution of Eqs. (5) and (6), in addition to the mass and momentum conservation laws.

8.1 The Effect of Surface Roughness on the Kinetic Field

Figure 10 illustrates how the velocity field changed in the area where propellers were being used. The changes of the velocity field in the propeller working area, where we clearly notice how with the increase in surface roughness the area of slow flow increases within the range (0-0.2 $m.s^{-1}$) and the flow direction perpendicular to the velocity gradient increases (the direction perpendicular to the symmetry plane). This means that the inconsistency in the flow in the propeller disc plane will increase with the increase in roughness, and this will inevitably lead to an increase in the oscillations around the propeller, which in turn will be transmitted to the stern of the ship and its hull.

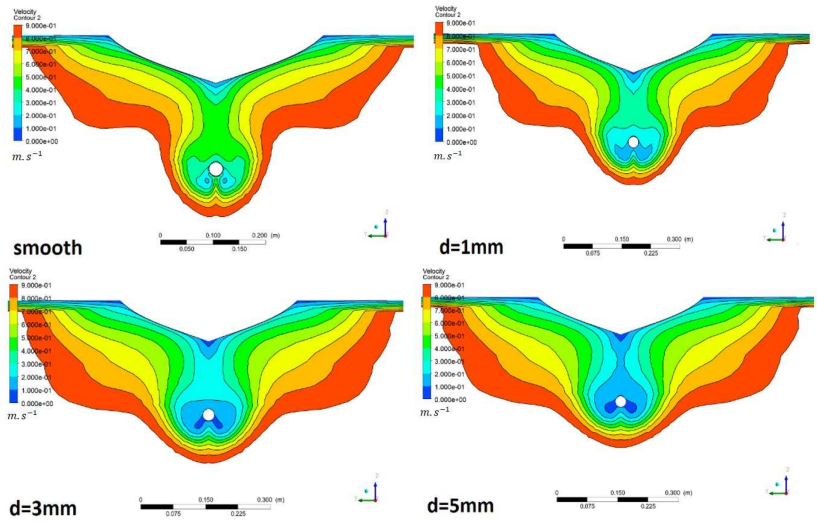


Figure 10

The changes of the velocity field according to the surface roughness in propeller plane for d = smooth, 1, 3, 5 mm

The increase in slow-flow regions within the propeller working area can affect the thrust generated by the propellers. where the presence of slow-flow areas can reduce the efficiency of the propellers, resulting in lower thrust production. This can potentially impact the overall performance and manoeuvrability of the ship.

Furthermore, the changes in flow direction perpendicular to the velocity gradient can affect the drag experienced by the ship's hull. where the increase in roughness can lead to an irregular flow pattern, resulting in higher drag forces acting on the hull. The drag force can hinder the forward movement of the ship, requiring more power to overcome the resistance and maintain desired speeds.

8.2 The Effect of Surface Roughness on the Kinetic Field

Figure 11 shows the change of the kinetic energy field according to the surface roughness in the working plane of the propeller. As we notice from this figure, with the increase in roughness, the kinetic energy gradient increases in the perpendicular direction to the plane of symmetry, and the value of kinetic energy increases in the propeller disc circle, and this indicates an increase in the intensity of the vortices entering the propeller disc, as we know in the case of bulk vessels and transport vessels. The lower part of the stern takes the form of a tube and accordingly, we have two huge vortices that enter into the plane of the propeller.

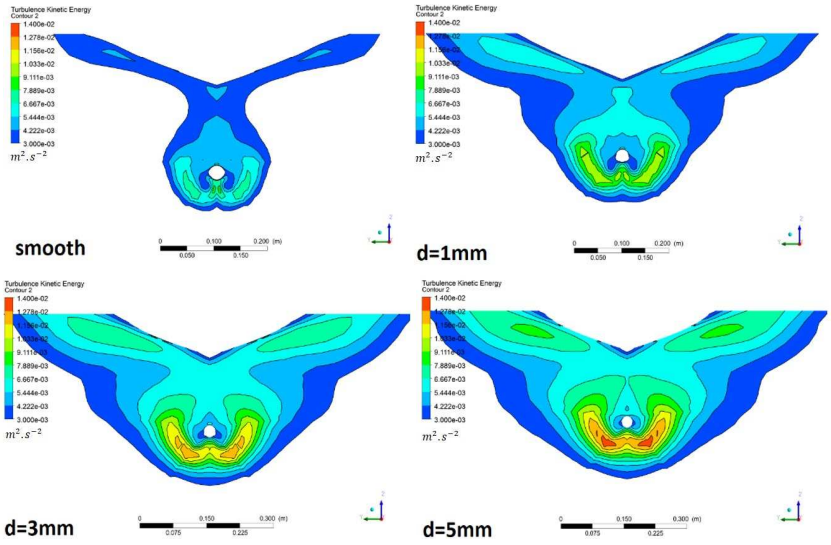


Figure 11
The changes of the kinetic energy field according to the surface roughness in propeller plane for d = smooth, 1, 3, 5 mm

8.3 The Effect of Surface Roughness on the Formation of Vortices

As shown in Fig.12, behind the hull of the studied ship, there are two large longitudinal vortices (A), in addition to two small vortices at the surface (B), and two small vortices at the propeller axis installation area (C). The eddies at the

surface are greatly influenced by the free surface and waves that form during sailing. The most important and influential vortices are the two huge vortices (A) because they enter directly into the propeller working area. These huge eddies form behind the tubular part of tankers and bulk carriers.

Accordingly, the intrusion of vortices into the propeller plane triggers a cascade of negative consequences that impact both the propeller system and overall fuel efficiency.

First, the presence of vortices induces heightened propeller vibrations, which can lead to mechanical instabilities and potential damage to the propeller structure. These vibrations not only compromise the structural integrity but also contribute to decreased thrust generation.

Second, the stresses exerted on the propeller blades increase because of the vortices interacting with the propeller. These elevated stresses can lead to premature fatigue and wear of the propeller blades, further impairing their performance and reducing the thrust force generated. The reduced thrust force necessitates higher power consumption to maintain desired speeds and propel the ship effectively.

Moreover, the disturbances caused by the vortices disrupt the smooth flow patterns and increase drag around the propeller, leading to an overall decrease in propulsion efficiency. This inefficiency translates into increased fuel consumption as more power is required to overcome the resistance and maintain the desired propulsion performance.

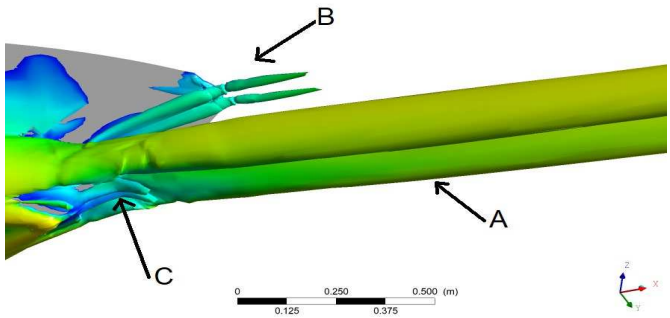


Figure 12
Vortices forming behind the hull of the KVLCC2 tanker

Conclusions

In this research, the CFD technique was used to investigate the effect of surface roughness on vehicle resistance, and the following conclusions were reached:

- Propeller performance is significantly affected by the surface roughness of the hull.

- As the surface roughness of the hull increases, the velocity field in the working area of the propeller becomes more inconsistent, and the velocity gradient increases in the direction perpendicular to the flow. This leads to an increase in vibration in the propeller, which can negatively affect propeller performance and cause increased vibration and stress on the propeller blades.
- As the surface roughness increases, the kinetic energy gradient in the direction perpendicular to the plane of symmetry also increases. This results in an increase in the intensity of vortices entering the propeller, which can further degrade the performance of the propeller.
- When designing and calculating a vessel, it is essential to take surface roughness into account as it affects the performance of the propeller.
- An increase in surface roughness leads to an increase in velocity field inconsistency, kinetic energy, and vortex formation, which in turn can negatively affect propeller performance.
- RANS equations, k-ω SST model, and the roughness function used in the Ansys program give very good results in marine applications and help save material costs and time, especially when calculating resistances.
- Roughness has a significant impact on the forces acting on the vessel and the flow characteristics around the hull.
- An increase in surface roughness leads to an increase in velocity field inconsistency, kinetic energy, and vortex formation, which in turn can negatively affect propeller performance.
- These results are useful for predicting the energy required to operate the ship in marine conditions, fuel consumption, and greenhouse gas emissions.

Nomenclature

Symbols

B	-	constant considers the effect of surface roughness on the velocity distribution
B_{molded}	m	breadth (molded)
B_{plate}	mm	breadth of the plate
B'	m	width of the geometric domain for the ship model
B''	m	width of the geometric domain for the plate
C_b	-	blockage coefficient
C_d	-	drag coefficient
C_f	-	friction coefficient

C_p	-	pressure coefficient
C_s	-	roughness constant
C_T	-	total resistance coefficient
$C_{T,EFD}$	-	total resistance coefficient according to experimental results
$C_{T,CFD}$	-	total resistance coefficient according to numerical results
C_w	-	wave resistance coefficient
d	m	roughness height value
f_r	-	roughness coefficient
F_r	-	Froude number
g	$m.s^{-2}$	acceleration of gravity
h	m	the characteristic linear dimension
k_{const}	-	Kármán's constant
k	m	roughness height
k_s	m	the equivalent sand roughness height (physical roughness height)
k_s^+	-	non-dimensional roughness height
L_{plate}	m	length of the plate
L_{pp}	m	length between perpendiculars
L'	m	length of the geometric domain for the ship model
L''	m	length of the geometric domain for the plate
L_{wl}	m	length of the water line
p	N/m^2	pressure
Re	-	Reynolds number
r_{plate}	mm	edge turning radius for the plate
S	m^2	wetted surface area without appendages
T	m	draft (molded)
T_{plate}	mm	draft of the plate
T'	m	draft of the geometric domain for the ship model
T''	m	draft of the geometric domain for the plate

U	$m.s^{-1}$	the mean velocity of the object relative to the fluid
U_e	$m.s^{-1}$	the free flow velocity
U_p	-	the non-dimensional flow velocity along the wall in the boundary layer
U^+	-	the non-dimensional velocity in the boundary layer (velocity profile)
U^*	$m.s^{-1}$	friction velocity
U_τ	$m.s^{-1}$	frictional velocity
ΔU^+	-	roughness function
V	m^3	displacement
y	m	wall distance
y^+	-	the non-dimensional distance measured perpendicularly to the surface

Greek letters

α, β	-	*k-ω* SST coefficients
δ	m	the boundary layer thickness
ε	-	the rate of dissipation of the turbulent kinetic energy
μ	kg/ms	dynamic viscosity
μ_t	kg/ms	Eddy viscosity
ν	m^2/s	kinematic viscosity
v_{ship}	$m.s^{-1}$	ship velocity
ν_T	m^2/s	turbulent viscosity
ρ	kg/m^3	the density of the fluid
σ	N/m^2	normal stress
τ_w	N/m^2	shear stress
ω	$1/S$	specific dissipation rate

Acknowledgement

The first author was supported by Dr. Nawar Abbas, Assistant Dr-Eng at Marine Engineering Department, Faculty of Mechanical and Electrical Engineering, Tishreen University, Latakia, Syria. The authors were supported by project no. 129257 implemented with the support provided to the corresponding author from the National Research, Development and Innovation Fund of Hungary, financed under the K18 funding scheme.

References

[1] H. Schlichting, K.Gersten: "Boundary layer theory". Newyork, USA: Springer- Verlog Berline Heidelberg, 2000. 1960

[2] J. D. Anderson: "Ludwig Prandtl's boundary layer," Physics Today, vol. 58, no. 12, pp. 42-48. December, 2005, doi: 10.1063/1.2169443

[3] S. Song, Y.K. Demirel, C. De Marco Muscat-Fenech, D. Sant, T.; Villa, T. Tezdogan, A. Incecik: "Investigating the Effect of Heterogeneous Hull Roughness on Ship Resistance Using CFD". J. Mar. Sci. Eng, vol. 9, no. 2, pp. 202. 2021, https://doi.org/10.3390/jmse9020202

[4] S. Song, R. Ravenna, S. Dai, C. De Marco Muscat-Fenech, G. Tani, Y.K. Demirel, M. Atlar, S. Day, A. Incecik: "Ex-perimental investigation on the effect of heterogeneous hull roughness on ship resistance". Ocean Eng, vol. 223, p. 108590. 2021

[5] S. Song, Y.K. Demirel, M. Atlar, S. Dai, S. Day, O. Turan: "Validation of the CFD approach for modelling roughness effect on ship resistance". Ocean Eng., vol. 200, p. 107029. 2020, doi: 10.1016/j.oceaneng.2020.107029

[6] Y. Tahara, J. Ando: " Comparison of CFD and EFD for KCS container ship in without/with propeller conditions ". In: Gothenburg: A Workshop on Numerical Ship Hydrodynamics. Chalmers University of Technology, Gothenburg, Sweden. 2000, vol. 192, pp. 63-70, doi: https://doi.org/10.2534/jjasnaoe1968.2002.63

[7] S. B. Pope: "Turbulent Flows". UK, Cambridge University Press. 2000

[8] H. Herwig: "Strömungsmechanik". Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3-11. 2002, doi: 10.1007/978-3-662-10107-0_1

[9] M. P. Schultz, G. W. Swain: "The influence of biofilms on skin friction drag," Biofouling, vol. 15, no. 1–3, pp. 129–139, May 2000, doi: 10.1080/08927010009386304

[10] A. E. Perry, W. H. Schofield, P. N. Joubert: "Rough wall turbulent boundary layers," J. Fluid Mech., Vol. 218, pp. 405-438. September 1990, doi: https://doi.org/10.1017/S0022112090001057

[11] J. Nikuradse: "Laws of Flow in Rough Pipes", Forschung auf dem Gebiete des Ingenieurwesens. Ausgabe B Band 4, July/August, 1933, no.1292

[12] M. A. Shockling, J. J. Allen, and A. J. Smits: "Roughness effects in turbulent pipe flow", J. Fluid Mech. 2006, doi: 10.1017/S0022112006001467

[13] M. P. Schultz, K. A. Flack: "The rough-wall turbulent boundary layer from the hydraulically smooth to the fully rough regime", J. Fluid Mech, vol. 580, pp. 381-405. 2007, doi: 10.1017/S0022112007005502

[14] J. Nikuradse: "Laws of Flow in Rough Pipes," J. Appl. Phys., 1950, doi: 10.1063/1.1715007

[15] F. Hama: “Boundary-layer characteristics for smooth and rough surfaces”, Trans. - Soc. Nav. Archit. Mar. Eng, vol. 62, pp. 333-351. 1954, URL= https://cir.nii.ac.jp/crid/1572824499659264512

[16] R. A. Antonia, R. E. Luxton: “The response of a turbulent boundary layer to a step change in surface roughness Part 1. Smooth to rough”. J. Fluid Mech, vol. 48, no. 4, pp. 721-761. 1971, doi: 10.1017/S0022112071001824

[17] R. A. Antonia, R. E. Luxton: “The response of a turbulent boundary layer to a step change in surface roughness. Part 2. Rough-to-smooth,” .J. Fluid Mech, vol. 53, no. 4, pp. 737-757. 1972, doi: 10.1017/S002211207200045X

[18] P. M. Ligrani, R. J. Moffat: “Structure of transitionally rough and fully rough turbulent boundary layers,” .J. Fluid Mech, vol. 162, pp. 69-98. 1986, doi: 10.1017/S0022112086001933

[19] P. R. Bandyopadhyay: “Rough-Wall Turbulent Boundary Layers in the Transition Regime”. J. Fluid Mech, vol. 180, pp. 231-266. 1987, doi: 10.1017/S0022112087001794

[20] P. A. Krogstad, R. A. Antonia, L. W. B. Browne: “Comparison between rough and smooth-wall turbulent boundary layers”. J. Fluid Mech, vol. 245, pp. 599-617. 1992, doi: 10.1017/S0022112092000594

[21] M. Schultz: “The effect of biofilms on turbulent boundary layers,” Ph.D. Thesis, Florida Institute of Technology. 1998

[22] M. P. Schultz, G. W. Swain: “The effect of biofilms on turbulent boundary layers,” J. Fluids Eng. Trans. ASME, 1999, doi: 10.1115/1.2822009

[23] M. P. Schultz: “Effects of coating roughness and biofouling on ship resistance and powering”, Biofouling, vol. 23, pp. 331-341. 2007, doi: 10.1080/08927010701461974

[24] H. Schlichting: “Boundary layer theory: Seventh edition.,” 1979

[25] Nikuradse, J: "Laws of flow in rough pipes [English translation of Stromungsgesetze in rauhen Rohren]," VDI-Forschungsheft, vol. 361, pp. 1-22. 1930

[26] Y. K. Demirel, M. Khorasanchi, O. Turan, A. Incecik: “A parametric study: Hull roughness effect on ship frictional resistance,” RINA, R. Inst. Nav. Archit. - Int. Conf. Mar. Coatings, 2013

[27] Y. K. Demirel, O. Turan, A. Incecik: “Predicting the effect of biofouling on ship resistance using CFD,” Appl. Ocean Res. vol.6 2, pp. 100-118. 2017, doi: 10.1016/j.apor.2016.12.003

[28] D. Owen, Y. K. Demirel, E. Oguz, T. Tezdogan, and A. Incecik: “Investigating the effect of biofouling on propeller characteristics using CFD,” Ocean Eng. vol. 159, pp. 505-516, 2018, doi: 10.1016/j.oceaneng.2018.01.087

[29] Ansys Fluent 2020 R1-Theory Guide, https://ansyshelp.ansys.com/account/secured?returnurl=/Views/Secured/corp/v201/en/flu_th/flu_th.html?q=ansys%20fluent%20theory%20guide. 2020

[30] "MOERI KVLCC2 Geometry and Conditions, SIMMAN 2008, FORCE Technology" http://www.simman2008.dk/KVLCC/KVLCC2/kvlcc2$_$geometry.html

[31] M. P. Schultz: "Frictional resistance of antifouling coating systems," J. Fluids Eng. Trans. ASME, vol. 126, no. 6, pp. 1039-1047. 2004, doi: 10.1115/1.1845552

[32] F. R. Menter: "Two-equation eddy-viscosity turbulence models for engineering applications," AIAA J., vol. 32, no. 8, pp. 1598–1605, Aug. 1994, doi: 10.2514/3.12149

[33] T. Cebeci, "Turbulence models and their application: efficient numerical methods with computer programs," Springer. vol. 24, no. 3, pp. 407. 2004, doi: 10.1016/j.euromechflu.2004.08.001

[34] Hoch, Toralf: " Development of a "numerical test bench" for turbine wheel gas meters," Diss. Duisburg, Essen, Universität Duisburg-Essen, Diss. 2011

[35] CFD Online, "*k*-ω SST model," https://www.cfd-online.com/Wiki/SST_k-omega_model, August, 2013

[36] Wilcox D. C: "Turbulence modeling for CFD," DCW industries, La Canada, vol. 98, no. 980, pp. 405. 1993, doi:10.1017/S0001924000027032

[37] Menter F. R., Kuntz M., and Langtry R: "Ten years of industrial experience with the SST turbulence model," Turbulence, heat and mass transfer, vol. 4, no. 1, pp. 625-632. 2003

Human-Machine Co-Working for Socially Sustainable Manufacturing in Industry 4.0

Martin Mareček-Kolibiský[1], Samuel Janík[1], Miroslava MÍkva[1], Peter Szabó[1], György Czifra[2]

[1] Slovak University of Technology in Bratislava, Faculty of Material Science and Technology in Trnava, Jana Bottu 2781/25, 917 24 Trnava, Slovakia martin.marecek-kolibisky@stuba.sk, samuel.janik@stuba.sk, miroslava.mlkva@stuba.sk, peter.szabo@stuba.sk

[2] Óbuda University Bánki Donát Faculty of Mechanical and Safety Engineering Népszínház u. 8, 1081 Budapest, E-mail: czifra.gyorgy@bgk.uni-obuda.hu

Abstract: Human-machine cooperation is an activity used to maximize job openness for all workers by removing barriers of languages, disability, age, gender barriers and maximizing employee well-being and motivation. The diverse technologies providing physical and cognitive assistance should facilitate attractiveness and facilitate employment, and thus social sustainability within the production section. The main goal of this paper is to analyze the current state of human-machine cooperation and identify the requirements for future human-machine cooperation for socially sustainable manufacturing in Industry 4.0.

Keywords: human-machine cooperation; Industry 4.0; employment; production processes; social sustainability

Introduction

The world has witnessed three industrial revolutions since the end of the eighteenth century, which have brought major leaps in the efficiency and productivity of industrial activities. The 4th Industrial Revolution and Digitization Society are currently taking place on a global scale. We encounter elements of digitization not only in industrial enterprises and industry as such, but they can be found in everyday life as well.

While the first and second industrial revolutions were characterized by mechanization based on the invention of the steam engine and electrification of production processes, the third industrial revolution was defined by more progressive automation of processes to production [1].

A characteristic element of the fourth industrial revolution is the digitization of all systems within the organization and their interconnection into one whole, such a revolution can be referred to as Industry 4.0. Industry 4.0 is characterized by interactions and communication between industrial equipment (machines) and cyberphysical systems for real-time operations management, the Internet of Things, artificial intelligence, robotics, cybersecurity, and other elements and technologies that contribute to technical sophistication, increased competitiveness, and production automation [2, 3].

The main idea of industrial transformation is to increase the competitiveness of enterprises, through increasing resource efficiency and productivity [4]. Quality of work, quality of processes, overall quality, and product safety are important for maintaining and improving the competitiveness of companies. The issue of quality in every industry has become a parameter of a company's survival in a turbulent competitive market. In addition to the quality of products and services, the success of companies also depends on the performance of the processes taking place in the system [5, 6].

The article is structured as follows: section 1 provides the theoretical background characterizing sustainable production, followed by human-machine collaboration in industrial practice. Section 2 describes the empirical data on the issue under study, obtained through a survey, then the research questions and hypotheses are stated. In Section 3, the research questions and hypotheses are evaluated and interpreted. Section 4 provides a discussion of the paper's topic, followed by a conclusion, including a suggestion of possible directions for future research.

1 Theoretical Background

One of the goals of the implementation of Industry 4.0 is to increase the professional knowledge and qualifications of people, and thus increase the well-being of employees under the guarantee of sustainable jobs. In Industrial Revolution 4.0, there is no competitive battle between workers and machinery. Industry 4.0 offers opportunities for more efficient use of human potential in cooperation with machines [3].

The EU's population is aging and the EU's working-age population will fall by 1/3 by 2050 [7]. In addition to this change in society, new working styles, working from home and working with robots are becoming popular, and societal and working lives are being transformed [8].

1.1 Socially Sustainable Production

Social sustainability was emphasized only after the Rio Conference in 1992. The United Nations Conference on Environment and Development (UNCED) sets out in its Agenda 21 human and social issues and their impact on sustainability. The first part of Agenda 21 emphasizes the importance of combating poverty, protecting and promoting human health, and creating an impetus for sustainable human settlements, a social and economic dimension [9]. Researchers also define social sustainability as *"a code of conduct for human survival and growth"* and *"must be achieved in a mutually accessible and prudent manner"* [10]. Social responsibility can be defined as *"the obligation of a company to use its resources in a way that is beneficial to society, through engaged participation as a member of society, consideration of society as a whole, and improving the well-being of society as a whole without regard to direct profits"* [11].

This concept of social sustainability can be extended to include the management of social resources, including people's skills and abilities, relationships, and social values. The United Nations Framework for Sustainable Development (UNDSD) classifies the dimensions of sustainable development and includes the social and economic environment. In the social dimension, the identified indicators are equality, education, health, housing, safety, and population. Social Sustainability (SU) is grouped into three categories (SU development, SU bridging, and SU maintenance) [12, 13].

It is these three categories of social sustainability that speak of social sustainability as an approach that helps humanity address social issues such as poverty, equality, education, wages, human rights, and diversity. However, social problems in industrialized economies differ from emerging economies due to their very different social standards. Social sustainability seems to be more difficult to accept and understand in many enterprises. Measuring the impact of social responsibility is a more challenging task for organizations, especially small and medium-sized enterprises. The concept of corporate social responsibility includes activities related to the social dimension of sustainability, but can have different meanings depending on the context and interpretation. In companies, we often discuss the concept of sustainable production [14, 15].

In addition to research, the concept of sustainable production has also moved to small and medium-sized enterprises, especially in industrial production. Sustainability of production is based on three areas: economic, environmental, and social. Sustainable production can be defined as the production of products in a way that minimizes environmental impacts and takes on the social responsibility of employees, the community, and consumers throughout the product life cycle and achieves positive economic results. The results of aligning organizations with the goals of socially sustainable production are clear. Decent jobs help keep employees at work, occupational safety and health care reduce illness and absence, and continuous employee training provides them with higher quality and

productivity. At present, the sustainability of production is most closely linked to the environment, for example, companies in the automotive industry. The aim is to reduce emissions in production through which it is possible to reduce the environmental impact by 45% per vehicle produced. Automotive companies such as Volkswagen, Tesla, etc. came up with a new concept for the production of electric cars. The mentioned examples of sustainable production with respect to the environment are related to human-machine cooperation. The reason for this cooperation is and will be new technologies and machines and at the same time a declining demographic curve. Man-machine cooperation is expected to contribute to reducing emissions, greenhouse gases, and industrial waste. Humanity is entering a period where the industry's intention is to affect the climate and the environment as little as possible. However, we cannot forget the man and his stable working conditions and environment at the same time, so a man in the production environment is complemented by a robot and two human-machine entities work together [15, 16, 17].

1.2 Man-Machine Cooperation

In order to prepare for labor shortages in the near future, it is necessary to take into account the fact that the working style of employees will change. Humans will work in coexistence with intelligent systems and robots. The production system in industrial enterprises will be fully automated, using various technologies and machines. The focus of job fear has shifted to automation, where people are replaced by machines. So we should discuss partnerships and man-machine cooperation in the workplace. Paradigms, between human-machine cooperation, should move from taking on a role to thinking together, learning together, and working together. The vision for the future is that machines will increasingly work and behave like humans. This means that creativity, intuition, and ethics can be common to humans as well as to machines in certain elements. It is assumed that human-machine algorithms will be developed and human-machine relationships will be managed by experts. People will have to trust the decisionmaking of autonomous machines. Relationships between humans and machines will require new industrial psychology [18].

One of the central characteristics of Industry 4.0 activities is the integration of two entities, the machine, modern technological progress and people (employees) [19]. It follows that future competitiveness should not only be ensured by superiority in productivity based on automation, but especially in the offer of added value to customers. For this reason, meaningful integration of the strengths of both the human and machine entities will increase production flexibility [20]. Successful cooperation and interaction of people with different machines (innovative technological hardware and software components) will be of great importance in various areas of industrial production (automotive industry, engineering industry, electrical industry, metallurgy) and also in the field of agricultural production.

In order to achieve a symbiosis of man and machine [21, 22]. The Industrial Revolution, in which industry, as well as society as a whole, finds itself, is transforming the design, engineering, production, operation, and service of products and production systems [23].

As stated by Krupitzer in his research [22], in which he analyzes the current state of human-machine interaction in Industry 4.0. Initial research and scientific efforts in the study focused on fully manageable systems. Over time, research has focused on adaptive mechanisms. This has led to the requirement to establish human-machine elements and to work together. In a complex the man-machine system can no longer be considered individual isolated units, but as a dynamic team working together on a common task. It is natural that even if some of the jobs of operators in production remain, some will not survive as we know them today. New profiles of workers with specific skills will be needed immediately, where manual work will be reduced in favor of cognitive and analytical skills and the way of working will be fundamentally changed. Information technology and work activities such as data analysis come to the front. According to the estimates of the US statistical office, there is talk of 1.37 million people in the US who will be retrained for the so-called "New viable" professions. Professions that do not currently exist at all, but will require skills and abilities in the field, such as (analysis of big data of users and entities, internet of things, markets with applications and web, virtual reality, creators of computer systems, cooperation with stationary robots, humanoid robots, etc.). These positions will include software developers, database administrators, computer systems engineers, and computer and information research scientists [15, 24].

Advanced modern digital and industrial technologies will help people stay in, return to, or join modern manufacturing companies and workgroups. Thanks to technological developments, such as new connectivity options and intelligent technologies between components, machines and humans, industrial production systems are increasingly evolving towards the idea of leaner, and more integrated production, real-time data monitoring, evaluation, and adaptation to production conditions [25].

The new work environment, based on the ideas of cyberphysical factories and the digital twin, will directly affect the operator, the nature of the work, and create new working connections between people and machines in the workplace, but also between the digital and physical environments. The future of companies through transformation to Smart Factories will require a new design and engineering philosophy for production systems focused on socio-technological transformation. Automation, robotics, and other modern technologies are considered elements that could further improve and expand human capabilities [21, 26]. The expansion and improvement of human capabilities in Smart Factory will be controlled by the Operator 4.0 model, where the operator will be understood as an "intelligent and skilled operator" performing its work not only with robots but also with intelligent machines using cyberphysical systems to achieve advanced human-machine

interaction and achieving a man-machine working symbiosis in automation. This understanding of Operator 4.0 is based on the assumptions of the industrial production of the future, which will require the analysis of big data of users and entities, the Internet of Things application, virtual reality cooperation with stationary robots and humanoid robots. The result will be the creation and development of new skills and knowledge of operators. In the future, the operator will be understood in a different sense than today. The operator will need to be qualified and professionally focused on data analysis, working with information systems, cloud solutions, and the Internet of Things. It is very likely that human-machine cooperation will take place using a computer [27].

A study [28] describes simultaneous localization and mapping (SLAM) technology. This technology is used in robots and robotic devices that evaluate and scan space in real-time. Using similar technology, new robots will be created in the industrial environment, which will be able to relieve people in the production process. Hancock [29] expressed the idea of human-machine interactions with respect to social sustainability. According to him, machines and automation should adapt to the cognitive and physical requirements of people in a dynamic way. In such a sense, adaptive automation aims to optimize man-machine cooperation and efficiently distribute man-machine work in a production system. The idea of adaptive automation will help increase the efficiency of the production system in a sustainable way, man and machine will achieve symbiosis in the production system and achieve production goals. The main goal of this adaptive automation paradigm is to achieve efficient production efficiency, prevent errors, and thus increase quality and eliminate forms of waste and improve the mental and physical burden on people. Everything is focused on the fact that people should never be subordinated to machines and automation, but on the contrary, machines, and automation must be helpful to people. According to Hancock and others, in order to achieve the sustainable development of human society, that is, the symbiosis between man and machine, automation is needed through the use of intelligent automation systems that will enable man's goals and plans to be met. Romero presented the involvement of "Enterprise Architecture (EA)", which represents a set of knowledge between man-machine cooperation. EA considers the socio-technological aspects of systems, combines management and engineering practices, highlights key requirements, principles, and models, includes people, business information, and technology processes, and describes the company's future position [25, 29, 30].

Innovation, based on the human-machine cooperation paradigm, benefits primarily from the advent of new technologies and ideas in the industrial environment. The advantage of open systematic innovation is primarily the use of machine intelligence on complex networks in the environment and the ability to quickly select those innovative technologies with the greatest potential. Technologies based on the ideas of a socially sustainable business environment, technologies that make work easier for people, minimize environmental pollution,

and improve the economic environment of the company. Closed innovation in the business environment will lead to a well-modeled search for business opportunities that can benefit the whole community. Patented systematic innovation will benefit from the man-machine function, joint innovation at a competitive advantage through rapid decision making, the creation of new markets, and the alignment of products and services with market dynamics. The technologies needed for machine intelligence are already available. The use of these technologies can have a positive impact on people's behavior and business development. The needs of today's market, as expressed by today's and tomorrow's consumers, call for advanced innovation processes that are fast and lead to customer-tailored products and services that are efficient. In this complex environment, intelligent machines will play an important role in the future. The impact of emerging events that have the potential to change the world of work and life will continue to evolve exponentially, resulting in the constant development of innovation, both inside and outside the business environment [31].

For Industry 4.0, costs and sustainable development are key aspects to consider when implementing new technologies. Cyber-physical systems, cybersecurity, blockchain, and additive manufacturing play an important role in the redistributed production model that promotes social sustainability. Technologies such as digital twins and Big Data will enable better data analysis in cooperation in the context of man and machine [32, 33, 34].

Lagashev [35] says in his research that cloud computing is currently one of the most widely used technical solutions for data processing and interconnection of this data within machines. And in the presented article he discusses the issue of a cloud server, which also deals with human-machine cooperation.

1.1.1 Man-Machine Cooperation in Industrial Practice

German industrial corporations in the automotive and engineering industries are among the leaders in Europe and in the world. Companies are technologically advanced, they are introducing new technologies and they have the financial means to implement new elements of the most modern technologies.

When the government of the Federal Republic of Germany came to the public in 2011 with the term Industry 4.0, all German corporations and companies began to implement elements of Industry 4.0. Industry 4.0 was created to improve the economic and industrial environment in Germany and Europe. The issue of the aging population in the EU and the labor shortage in the industry are also significant. As a result, technologies and machines have begun to be introduced into manufacturing companies and industries that can replace people or make work easier. This Industry 4.0 idea is not about removing people from production. Rather, it should motivate people to make their work easier, one of these paradigms being man-machine cooperation.

2 Materials and Methods

Empirical data on the researched issue were obtained using a scientific questionnaire. The questionnaire contained 37 closed questions, the first part was focused on finding out the identification and demographic characteristics of respondents, and in the second part we focused on the following four questions:

RQ1: Do you feel threatened by the introduction of new technologies in your organization?

RQ2: Do you currently consider your employer's social behavior (employee care) to be socially responsible?

RQ3: In what areas do you consider your employer's behavior to be socially responsible?

RQ4: Which skills do you consider most important in terms of digitization and job automation (0 don't know; 1 least important to 5 most important)?

Based on the research questions and for the purpose of the paper, the following research hypotheses were defined:

Hypothesis 1: In the fear of employment by introducing new technologies, there is a significant difference between employees of different positions.

Hypothesis 2: There is a significant difference in skills needed with regard to digitization and automation between employees in different job positions.

It was 556 respondents who filled in the questionnaire. Respondents answered the questions in the questionnaire as representatives for the company, not as individuals in terms of their employment status. Due to the thematic focus of our contribution, we focused on companies operating in the industrial sector with small (10 to 49 employees), medium (50 to 249 employees), and large (250 and more employees). We have excluded micro-enterprises (1 to 9 employees) from our research because the topic of human-machine cooperation will be implemented significantly in small, medium, and large enterprises. After filtering out the variables, we looked at a research sample of 322 respondents.

Data processing was performed using Microsoft Excel and IBM SPSS Statistics 28.0.0.0. The interpretation of the data was processed through statistical methods, such as histograms, pie charts, and chart analyses. Statistical quadratic tests and ANOVA were used to test the relationships between dependent and independent variables.

A deeper distribution of respondents operating in individual sectors of industrial production can be found in Figure 1. We focused mainly on the 4 largest industrial sectors in Slovakia.

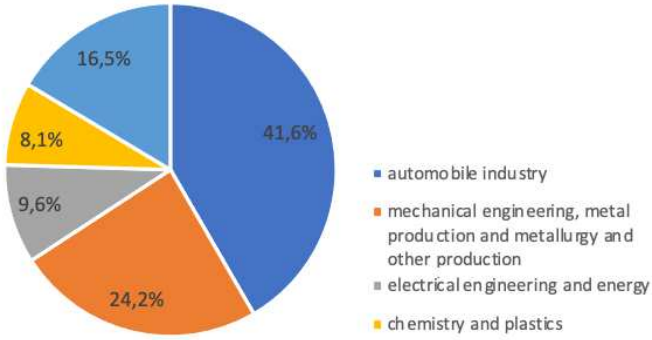


Figure 1
Distribution of respondents operating in individual sectors of industrial production

The distribution of respondents according to the size of the company according to the number of employees in which they work is shown in the pie chart in Figure 2. The graph shows that the largest part of respondents come from large companies (72.4%). It is in large companies that a massive integration of human-machine entities is expected. The rest of the respondents come from medium-sized enterprises (19.9%) and small enterprises (7.8%). In another question, we examined the representation of respondents depending on gender and job position. Based on these data, we can conclude that 64.3% of respondents were men, and 35.7% were women. Part of socially sustainable production is balancing gender equality across the organization, and based on the results of the analysis, the authors state that the current distribution is not in line with the trend of sustainable development, which may result in fewer inaccuracies in predictions for future human-machine cooperation with regard to gender diversity.

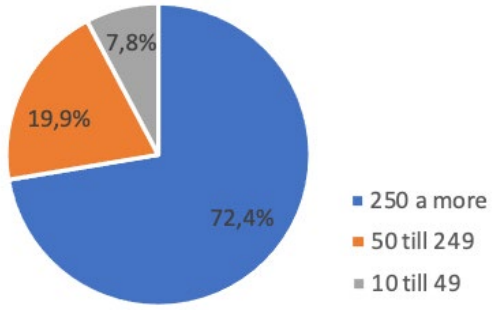


Figure 2
Representation of the relative number of respondents and the size of the company according to the number of employees

The distribution of respondents by job position is shown in Table 1. The term employee specialist is understood as an employee who may or may not be in a managerial position, but in terms of his job description is key to the management of production. Further, a 'production worker' is not a person who carries out operational activities directly in production, but a person who manages operational activities (i.e. the lowest level of management within the relevant organisational management structure, e.g. a teamleader). For this reason, we consider these views to be comparable.

Table 1
Distribution of respondents by job position and men/woman

Work position	**Overall**		**Men**		**Women**	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
employee specialist	102	31.7%	71	34.3%	31	27.0%
administrative staff member	80	24.8%	31	15.0%	49	42.6%
production worker	61	18.9%	42	20.3%	19	16.5%
management position	59	18.3%	46	22.2%	13	11.3%
other	20	6.2%	17	8.2%	3	2.6%

3 Research Results

In the following section, the individual research questions are evaluated and interpreted, as well as the research hypotheses that the authors considered. Due to the insignificant number of respondents who indicated other in the job position, these responses were excluded when examining the research questions and hypotheses as the results would not have relevant predictive power.

Research Question 1: **Do you feel threatened by the introduction of new technologies in your organization?**

We used basic descriptive statistics to evaluate the research question. In this research question, we examined how employees feel threatened by the introduction of new Industry 4.0 technologies. The significance of the threat in this research question is defined by the future expected degree of threat to their current job position. The obtained results are shown in Table 2. Based on the above results, it can be stated that 93 respondents (31%) state that their position is not in danger or employees do not feel threatened by the implementation of new technologies and 82 respondents (27%) say, "no, I assume that this will have a

significant positive effect on my work". The fact that 58% of all respondents have a positive attitude towards the introduction of new technologies in companies is a very important factor for the future of the business environment and the competitiveness of Slovak companies. In such work environments, human-machine cooperation will be implemented much better and more smoothly. From a job standpoint, it is worth mentioning the words "yes, I am worried about my job" (15%) and "yes, I am afraid that this will have a significant negative impact on my job" (10%) by production workers. From the results, we can conclude that the introduction of new technologies into companies from the perspective of employees can be perceived as a positive feature of the 4th Industrial Revolution, which focused on a high degree of automation and direct man-machine cooperation. The fourth industrial revolution, focusing on human-machine cooperation in the context of socially sustainable production, does not aim to remove employees from companies, but on the contrary to simplify their work or provide them with new jobs with an adequate retraining program.

Table 2
The feeling of endangering employees by introducing new technologies in the company

	employee specialist		**administrative staff member**		**production worker**		**management position**		**overall**	
	Abs. freq.	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**
yes, I am worried about my place	2	2%	3	4%	9	15%	1	2%	**15**	**5%**
yes, I am afraid it will have a significant negative impact on my work	5	5%	3	4%	6	10%	0	0%	**14**	**5%**
I think it will affect my work to a minimum	14	14%	18	23%	12	20%	7	12%	**51**	**17%**
I do not feel threatened	31	30%	23	29%	17	28%	22	37%	**93**	**31%**
I did not think about it	13	13%	9	11%	8	13%	2	3%	**32**	**11%**
no, I assume that this will have a significant positive effect on my work	34	33%	19	24%	5	8%	24	41%	**82**	**27%**
I am worried about my work for other reasons (e.g. economic consequences of COVID-19)	3	3%	5	6%	4	7%	3	5%	**15**	**5%**
overall	**102**	100%	**80**	100%	**61**	100%	**59**	100%	**302**	**100%**

Hypothesis 1: **In the fear of employment by introducing new technologies, there is a significant difference between employees of different positions.**

Table 3
Results of Hypothesis 1

Chi-Square Tests				
Please indicate which sector you work in:		**Value**	**df**	**Asymptotic Significance (2-sided)**
industrial production	**Pearson Chi-Square**	**48.736**	**18**	**0.000**
	Likelihood Ratio	51.539	18	0.000
	Linear-by-Linear Association	3.407	1	0.065
	N of Valid Cases	302		

The hypothesis was tested based on the job positions from which the possibility of job positions, which have been included in other, was excluded. The results are shown in the Table 3, the hypothesis was verified by Chi-Square Test and the strength of the correlation was determined using Cramer´s V value in the Table 4.

Table 4
Results of Hypothesis 1 Cramer´s V

Symmetric Measures				
Please indicate which sector you work in:			**Value**	**Approximate Significance**
industrial production	Nominal by Nominal	Phi	0.391	0.000
		Cramer's V	**0.226**	0.000
	N of Valid Cases		302	

Significance came out less than 0.05, that is, we reject H0 at 0.05 level of significance and this implies that there are significant differences among the workers in their concern about their job position. According to the value of Cramer's V is 0.226 so the result is that the dependence between the variables is moderately strong.

Research question 2: **Do you currently consider your employer's social behavior (employee care) to be socially responsible?**

In this research question, we examined the opinion of employees on the social area, with which we can connect social sustainability. The results obtained are shown in Table 5 depending on the job position. Based on the above results, it can be stated that a total of 218 respondents (72%) rate the employer's behavior as socially responsible. These results were evaluated depending on the variable job position and from the above results it is worth noting that up to 54 respondents, which is 92%, who work in the management positions evaluate the employer as socially responsible.

Table 5
Socially responsible (socially sustainable) behavior of the employer

	employee specialist		**administrative staff member**		**production worker**		**management position**		**overall**	
	Absolute frequency	**Relative frequency**	**Absolute frequency**	**Relative frequency**	**Absolute frequency**	**Relative frequency**	**Absolute frequency**	**Relative frequency**	**Absolute frequency**	**Relative frequency**
yes	69	68%	62	78%	33	54%	54	92%	**218**	**72%**
do not know	23	23%	15	19%	20	33%	4	7%	**62**	**21%**
no	10	10%	3	4%	8	13%	1	1%	**22**	**7%**
overall	**102**	100%	**80**	100%	**61**	100%	**59**	100%	**322**	**100%**

Research Question 3: **In what areas do you consider your employer's behavior to be socially responsible?**

In the third research question, we examined the opinion of employees on a specific social area in which their employer behaves socially responsibly. The results obtained are shown in Table 6 as absolute and relative numbers depending on the employee's job position. In the given question, the respondents had the opportunity to choose from several answers. Based on the above results, it can be stated that we received a total of 338 responses, of which a maximum of 193 (57%) responses were listed under the option "activities and measures to promote health". Other answers were "work-life balance" answered by 82 (24%) respondents. It was 38 (11%) respondents described “support for disadvantaged employees” and “support for vulnerable communities” was identified by 25 (7%) by respondents as the area that respondents considered least affected by the employer's actions.

Table 6

Areas of social behavior of the employer

	employee specialist		**administrative staff member**		**production worker**		**management position**		**overall**	
	Abs. freq.	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**	**Abs. freq.**	**Rel. freq.**
activities and health promotion measures	64	62%	54	62%	25	39%	50	60%	**193**	**57%**
support for disadvantaged employees	9	9%	9	10%	9	14%	11	13%	**38**	**11%**
work-life balance	22	21%	21	24%	22	34%	17	20%	**82**	**24%**
support for vulnerable communities	8	8%	3	3%	8	13%	6	7%	**25**	**7%**
overall	**103**	100%	**87**	100%	**64**	100%	**84**	100%	**360**	**100%**

Research questions 2 and 3 are directly related to the socially sustainable area and say what activities employers carry out in order for this social area to develop and be sustainable.

Research question 4: **Which skills do you consider most important in terms of digitization and job automation (0 don't know; 1 least important to 5 most important)?**

Table 7
Ability / skill

	Value	employee specialist		administrative staff member		production worker		management position		overall	
		Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.	Abs. freq.	Rel. freq.
technical (professional) skills	**0**	3	3%	1	1%	3	5%	0	0%	**7**	**2%**
	1	26	25%	16	20%	10	16%	7	12%	**59**	**20%**
	2	7	7%	9	11%	6	10%	4	7%	**26**	**9%**
	3	5	5%	12	15%	9	15%	10	17%	**36**	**12%**
	4	29	28%	17	21%	13	21%	20	34%	**79**	**26%**
	5	32	31%	25	31%	20	33%	18	31%	**95**	**31%**
digital skills	**0**	3	3%	1	1%	2	3%	0	0%	**6**	**2%**
	1	23	23%	14	18%	6	10%	5	8%	**48**	**16%**
	2	11	11%	9	11%	9	15%	9	15%	**38**	**13%**
	3	11	11%	14	18%	12	20%	12	20%	**49**	**16%**
	4	18	18%	12	15%	7	11%	14	24%	**51**	**17%**
	5	36	35%	30	38%	25	41%	19	32%	**110**	**36%**
ability to learn	**0**	3	3%	0	0%	1	2%	0	0%	**4**	**1%**
	1	26	25%	23	29%	10	16%	7	12%	**66**	**22%**
	2	7	7%	3	4%	11	18%	7	12%	**28**	**9%**
	3	9	9%	15	19%	8	13%	13	22%	**45**	**15%**
	4	20	20%	18	23%	14	23%	15	25%	**67**	**22%**
	5	37	36%	21	26%	17	28%	17	29%	**92**	**30%**
flexibility, adaptation	**0**	5	5%	1	1%	2	3%	0	0%	**8**	**3%**
	1	17	17%	15	19%	6	10%	5	8%	**43**	**14%**
	2	13	13%	10	13%	8	13%	11	19%	**42**	**14%**
	3	20	20%	23	29%	19	31%	13	22%	**75**	**25%**
	4	18	18%	17	21%	14	23%	11	19%	**60**	**20%**
	5	29	28%	14	18%	12	20%	19	32%	**75**	**25%**
social (ability to get along with other people)	**0**	5	5%	1	1%	3	5%	0	0%	**9**	**3%**
	1	16	16%	14	18%	7	11%	10	17%	**47**	**16%**
	2	28	27%	20	25%	18	30%	13	22%	**79**	**26%**
	3	27	26%	19	24%	17	28%	21	36%	**84**	**28%**
	4	11	11%	14	18%	9	15%	8	14%	**42**	**14%**
	5	15	15%	12	15%	7	11%	7	12%	**41**	**14%**
overall		**102**	100%	**80**	100%	**61**	100%	**59**	100%	**302**	**100%**

In the last research question, the authors dealt with the abilities or skills that employees consider most important with regard to digitization and job automation. The research question, by its very nature, deals with the future requirements for the ability of employees in human-machine cooperation. In the questionnaire survey, respondents commented on the question on a scale from 0 (I do not know);

1 (least important) to 5 (most important). Respondents had the opportunity to comment on the following skills: technical (professional) skills; communication skills; organization of time within work, and work tasks; ability to manage and make decisions; ability to learn; ability to work under pressure; digital skills; language (foreign languages); social (ability to get along with other people); initiative (entrepreneurship, commitment); flexibility, adaptation; creativity and creativity. As some skills and knowledge do not directly relate to our research area, human-machine cooperation in socially sustainable production, we have decided to select only the following skills/competencies that we consider essential in human-machine cooperation. The skills and competencies are evaluated in Table 7. Overall, the most numerous skills were of a technical (professional) nature and digital skills. It was 110 (36%) respondents rated digital skills as the most important and up to 95 (31%) respondents rated technical (professional) skills as the most important. We assume that the effective synergy between the mentioned digital skills and professional skills will be key in the integration of employees affected by Industry 4.0 technologies in the context of human-machine cooperation.

Hypothesis 2: **There is a significant difference in skills needed with regard to digitization and automation between employees in different job positions.**

The hypothesis was tested based on the job positions. The results are shown in the Table 8, the hypothesis was verified using analysis of variance.

Table 8
Results of Hypothesis 2 - ANOVA

ANOVA						
		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
technical (professional) skills	Between Groups	4.311	3	1.437	0.630	**0.596**
	Within Groups	663.723	291	2.281		
	Total	668.034	294			
digital skills	Between Groups	3.521	3	1.171	0.524	**0.666**
	Within Groups	654.070	292	2.240		
	Total	657.591	295			
ability to learn	Between Groups	4.202	3	1.401	0.593	**0.620**
	Within Groups	695.009	294	2.364		
	Total	699.211	297			
flexibility, adaptation	Between Groups	5.998	3	1.999	1.069	**0.362**
	Within Groups	542.234	290	1.870		
	Total	548.231	293			
social (ability to get along with other people)	Between Groups	0.240	3	0.080	0.500	**0.985**
	Within Groups	464.565	289	1.607		
	Total	464.805	292			

For each skill category, the significance came out greater than 0.05, i.e., we do not reject H0 at the 0.05 level of significance and this implies that there are no significant differences in skills among workers with respect to digitization and automation of work. For all positions tested, these skills are equally important.

4 Discussion

Due to the advantages and disadvantages of human-machine cooperation, collaboration is very important, a machine can represent a person in certain activities and a person will have more time and energy for other activities that the machine cannot handle (handling more complex parts of products). From the results we analyzed, we came to the conclusion that employees in companies are aware of the need to introduce new technologies, which we can state based on the results focused on the need for skills and abilities. Employees, respondents rated technical (professional) skills as well as digital skills as the most necessary, which is undoubtedly the basis for human-machine cooperation. We can define precise paths of activities and procedures for the machine so that they are economical and efficient for the company, and that is why one will have to develop these two areas of skills, which serve precisely to coordinate with the robot. Cooperation between man-machine opens up new possibilities for advancing not only industry but also everyday life.

Conclusion

In the last decade, technology and technological progress have gone exponentially to the forefront. Today, humanity, science, and industry know the technologies that undoubtedly make people's daily lives easier. Technological progress is still advancing and at present, we cannot even realize what awaits us in the industry in the future, but the application of the Industry 4.0 paradigm, accelerating and improving production processes, focusing on the sustainability of production and production processes in accordance with the 3 dimensions of sustainability (economic, social and environmental). The views of authors [21, 24, 31] and scientific researchers differ. Many authors state in a global sense the negative impact on human-machine cooperation, others discuss the positive impact on man and his work. Man in the production process is an irreplaceable aspect, just like a machine. There must be a definition of areas of work for the machine, for man, and their cooperation. Collaboration in terms of outlining the activities that will be performed by a man and by machine. Defining, for example, specific tasks such as feeding parts to a machine/robot for the human assembly, so that man does not interfere with the production process (feeding parts) of the machine and the machine, robot into human work activities (assembly of parts). Furthermore, to define the distances between the machine or robot and human, if a human approaches a specified distance, which can endanger a human, he must cause the machine to stop its movement, and activity. Defining ergonomic requirements when working with loads, wherein such an activity, the work of the robot is a human aid. The primary purpose of introducing technology, automation, and digitization into the industry is due to a weakening workforce and the facilitation of human activities.

Without the implementation of the basic elements and technologies of Industry 4.0, companies will not be able to constantly adapt to the new challenges that come with this new paradigm. Without basic innovations and implementations within the Industry 4.0 trend, small, medium, and large Slovak companies will not be able to apply technologies that would support human-machine interaction. Employees, especially production workers, based on the current results of a questionnaire survey, more than 65% of employees would say that they generally do not feel threatened by the introduction of new technologies into the company in which they work. More than 85% of employees do not feel threatened by administrative staff. It is the production and administrative staff that are expected to be the most vulnerable groups in terms of the introduction of digitization-related technologies and Industry 4.0.

Acknowledgment

This work was supported by project VEGA 1/0721/20 "*Identification of priorities of sustainable human resources management with regard to disadvantaged employees in the context of Industry 4.0*".

The paper is a part of project KEGA No. 018TUKE-4/2022 „*Creation of new study materials, including an interactive multimedia university textbook for computer-aided engineering activities* ".

References

[1] KAGERMANN. H. 2015. Change Through Digitization-Value Creation in the Age of Industry 4.0. In: *Management of Permanent Change. Springer Fachmedien Wiesbaden, Wiesbaden*, 23–45

[2] SHROUF, F.; ORDIERES, J.; MIRAGLIOTTA, G. 2014. Smart Factories in Industry 4.0: A Review of the Concept and of Energy Management Approached in Production Based on the Internet of Things Paradigm. In: *Proceedings of the 2014 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, 697–701. https://doi.org/10.1109/IEEM.2014.7058728

[3] FIFEKOVÁ, E.; NEMCOVÁ, E. 2016. Industry 4.0 and its implications for EU industrial policy. [12-2021]. Available on the Internet: http://www.prog.sav.sk/sites/default/files/2018-03/Priemysel.4.0.a.jeho_.implikacie.pre_.priemyselnu.politiku.pdf

[4] USTUNDAN, A.; CEVIKCAN, E. 2018. *Industry 4.0: Managing The Digital Transformation*; Springer International Publishing: Cham, Switzerland

[5] POTKÁNY, M.; GEJDOŠ, P.; LESNÍKOVÁ, P.; SCHMIDTOVÁ, J. 2020. Influence of quality management practices on the business performance of Slovak manufacturing enterprises. Acta Polytechnica Hungarica. 17, 161–180 s. [2-2022] http://acta.uni-obuda.hu/Potkany_Gejdos_Lesnikova_Schmidtova_106.pdf

[6] ŠOLC, M.; KOTUS, M.; GRAMBALOVÁ, E.; KLIMENT, J.; PALFY, P. 2019. Impact of corrosion effect on the quality and safety of refractory materials. Syst. Saf. Hum. Tech. Facil. Environ. 1, 760–767. https://content.sciendo.com/view/journals/czoto/1/1/article-p760.xml?product=sciendo

[7] https://ec.europa.eu/eurostat

[8] HAYASHIDA, N. 2018. Sensecomputing for Human-Machine Collaboration trough HUman Emotion Understanding. In: *Sci.Tech. Journal*. 54(5)

[9] UNCED. 1992. United Nations Conference on Environment and Development. http://sustainabledevelopment.un.org/index.php?page=view&nr=23&type=400

[10] SHARMA, S.; RUUD, A. 2003. On the path to sustainability: integrating social dimensions into the research and practice of environmental management, In: *Business Strategy and the Environment*, 12(4), 205–214 s.

[11] VAN DER WIELE, T.; KOK, P.; MCKENNA, R.; BROWN, A . 2001. A Corporate Social Responsibility Audit within a Quality Management Framework. In: *J. Bus. Ethics,* **31**, 285–297 s.

[12] UNDSD .2001. Indicators of Sustainable Development: Guidelines and Methodologies. [1-2022] http://www.un.org/esa/sustdev

[13] VALLANCE, S.; PERKINS, H. C.; DIXON, J. E. 2011. What is social sustainability? A clarification of concepts, In: *Geoforum*, 42 (3), 342–348s.

[14] GUGLER, P.; SHI, J. Y. 2009. Corporate social responsibility for developing country multinational corporations: lost war in pertaining global competitiveness?. In: *Journal of Business Ethics*, 87 (1), 3–24 s. https://doi.org/10.1007/s10551-008-9801-5

[15] SARTAL, A. a kol. 2020. The sustainable manufacturing concept, evolution and opportunities within Industry 4.0: A literature review. In: *Sustainable Manufacturing – Review Article*. 12(5). 1 -17s.

[16] WILSON, C. 2018. Designing the purposeful world: the sustainable development goals as a blueprint for humanity. New York: Routledge

[17] DESPEISSE, M.; MBAYE, F.; BALL PD. 2012. The emergence of sustainable manufacturing practices. Prod Plan Control, 23(5): 354–376 s. https://doi.org/10.1080/09537287.2011.555425

[18] BOTHA, A.P. 2016. Developing executive future thinking skills, International Association for Management of Technology, *IAMOT 2016 Conference Proceedings*, 951 – 972 s.

[19] NELLES, J.; KUZ, S.; MERTENS, A.; SCHLICK, Ch. M. . 2016. Human-centered design of assistance systems for production planning and control:

The role of the human in Industry 4.0. In: *2016 IEEE International Conference on Industrial Technology (ICIT). IEEE*, 2099–2104. https://doi.org/10.1109/ICIT.2016.7475093

[20] BRETTEL, M.; FRIEDERICHSEN, N.; KELLER, M.; ROSENBERG, M. . 2014. How Virtualization, Decentralization and Network Building Change the Manufacturing Landscape: An Industry 4.0 Perspective. *International Journal of Information and Communication Engineering* 8 (1), 37–44 s. doi.org/10.5281/zenodo.1336426

[21] ROMERO, D.; STAHRE, J.; WUESt, T.; NORAN, O.; BERNUS, P.; FAST-BERGLUND, L.; GORECKY, D. . 2016a. Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies. In: *CIE46 Proceedings*. 11(1)

[22] KRUPITZER, C.; LESCH, V.; ZÜFLE, M.; KOUNEV, S.; MÜLLER, S.; EDINGER, J.; BECKER, C.; LEMKEN, A.; SCHÄFER, D. 2020. A Survey on Human Machine Interaction in Industry 4.0. 1(1), 45 s. https://doi.org/10.1145/1122445.1122456

[23] BOSTON CONSULTING GROUP. 2015. Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries. [12-2021]. https://inovasyon.org/images/Haberler/bcgperspectives_Industry40_2015.pdf

[24] ZAHIDI, S.; RATCHEVA, V.; LEOPOLD, T.A. 2018. Towards a Reskilling Revolution A Future of Jobs for All. http://www3.weforum.org/docs/WEF_FOW_Reskilling_Revolution.pdf

[25] ROMERO, D.; NORAN, O.; STAHRE, J.; BERNUS, P.; FAST-BERGLUND, A. 2015. Towards a Human- Centred Reference Architecture for Next Generation Balanced Automation Systems: Human-Automation Symbiosis. In: *IFIP International Conference on Advances in Production Management Systems*. 556-566 s. https://link.springer.com/chapter/10.1007%2F978-3-319-22759-7_64

[26] ROMERO, D.; BERNUS, P.; NORAN, O.; STAHRE, J.; FAST-BERGLUND, A. 2016b. The Operator 4.0: Human cyber-physical systems & adaptive automation towards human-automation symbiosis work systems, production management initiatives for a sustainable world. In: *International Federation for Information Processing (IFIP)*, 677-686 s.

[27] ROMERO, D.; STAHRE, J.; TAISCH, M. 2020. The Operator 4.0: Towards socially sustainable factories of the future. In: *Computers and Industrial Engineering.* 139. https://doi.org/10.1016/j.cie.2019.106128

[28] MAC, T. T., LIN, CH-Y. HUAN, N. G., NHAT, L. D., HOANG, P. C., HAI, H. 2021. Hybrid SLAM-based Exploration of a Mobile Robot for 3D Scenario Reconstruction and Autonomous Navigation. *Acta Polytechnica Hungarica.* Vol. 18, No. 6

[29] HANCOCK, P.A.; JAGACINSKI, R.J.; PARASURMAN,R. WICKENS,C.D.;WILSON, G.F.; KABER, D.B. 2013. Human-automation interaction research: past, present and future. In: *Q.Hum. Factors Appl*. 21(2), 9-14 s. https://doi.org/10.1177/1064804613477099

[30] FASTH-BERGLUND, A.; STAHRE, J. 2013. Cognitive automation strategy- for reconfigurable and sustainable assembly systems. In: A*ssembly Autom.* 33(3), 294-303 s. https://doi.org/10.1108/AA-12-2013-036

[31] BOTHA, A.P. 2016. The Future of Artificial Intelligence – The Human-Machine Frontier, Foresight for Development, [1-2022] http://www.foresightfordevelopment.org/featured/artificial---intelligence---ii. http://foresightfordevelopment.org/featured/artificial-intelligence-ii

[32] LUKAČEVIĆ, F., ŠKEC, S., MARTINEC, T. 2022. Challenges of Utilizing Sensor Data Acquired by Smart Products in Product Development Activities. *Acta Polytechnica Hungarica.* Vol. 19, No. 4

[33] THOMAS, D. 2016. Costs, benefits, and adoption of additive manufacturing: A supply chain perspective. Int. J. Adv. Manuf. Technol. 85, 1857–1876. http://dx.doi.org/10.1007/s00170-015-7973-6

[34] USTUNDAN, A.; CEVIKCAN, E. 2018. Industry 4.0: Managing The Digital Transformation; Springer International Publishing: Cham, Switzerland

[35] LEGASHEV, L. V., BOLODURINA, I. P. 2020. An Effective Scheduling Method in the Cloud System of Collective Access, for Virtual Working Environments. *Acta Polytechnica Hungarica.* Vol. 17, No. 8

Verification of Articulatory Phonetics Features with Quantitative Data

László Czap

Institute of Automation and Infocommunication, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary, czap@uni-miskolc.hu

Abstract: This paper aims to refine the base data set of visemes – the visual counterparts of phonemes – with quantitative data to provide accurate input for visual speech synthesis (a talking head that supports the training of speech production of deaf and hard-of-hearing children). Measurement-based features extend the existing data and refine our previously used dynamic model of articulation. This requires the definition of two major types of data simultaneously: the shape of the mouth, which can be examined relatively simply in an ordinary camera image, and the position of the tongue, the analysis of which requires the use of medical-level imaging devices and the processing of their signals. Articulatory phonetics can be divided up into three areas to describe consonants. These are voice, place, and manner respectively. This study aims to confirm the description of the place of articulation with measurement data. Data derived from the shape and position of the tongue is suitable for determining the place of articulation of sounds. In the case of vowels, we estimated the tongue position with the centroid of the tongue while in the case of consonants, we define the place of articulation with the measured distance of the tongue from the palate. To measure these, we used MRI and US images and determined tongue contours with an automated process. The results of this analysis statically define data for articulation keyframes for visual speech synthesis. We applied our results to improve the existing Hungarian transparent talking head with a more accurate model based on the clarification of the dynamic features. We also adapted the same model to the Chinese Shaanxi Xi'an dialect.

Keywords: Quantitative tongue description; Articulatory phonetics; Place of articulation; Talking head; Viseme features

1 Introduction

Previous studies show that visual information on the physiological processes of human speech greatly contributes to understanding the complex mechanism of speech formation and, through this, to the effective development of speech synthesis methods [1]. The radiological and monitoring processes currently available, such as magnetic resonance imaging (MRI) [2], computer tomography

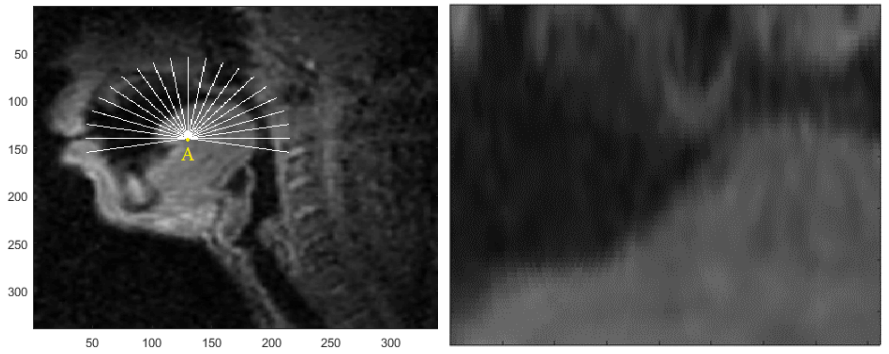
(CT) [3], ultrasound (US) [4], electropalatography (EPG) [5], or electromagnetic articulography (EMA) [6] are indispensable in getting to know the dynamic features of articulation. This is because the morphological and geometric data obtained with the help of imaging techniques can be used to map the articulation movements belonging to the given speech signal. This is essential, for example, in the parameterization of a talking head imitating the articulation. In this research quantitative data from a series of MRI and US images have been derived. Thus, we provided appropriate parameters for our animation algorithm. The main feature of this application is to show the tongue movements in a transparent-faced talking head. The basic items of this animation are the visemes. Such a system can be used well in speech therapy, in the design of non-native language learning training, or even in the construction of synthesizers to convert articulation features into silent speech [7]. Adaptation for a Chinese dialect has been examined as well.

The paper aims to provide a new quantitative method for analyzing tongue movements. Deaf and hard-of-hearing people are accustomed to lipreading, but unable to observe the invisible tongue movements. Without full acoustic perception, they rely on the visual modality of speech to be able to form their special speech signals. The quantitative data obtained helped in the better realization of a transparent talking head.

2 Methods and Material

The processing of MRI and US images was performed during the static and dynamic analysis. The programs for this were written in MATLAB environment, in the framework of which we fitted an auxiliary curve to the surface of the tongue based on dynamic programming [8].

The resolution of the raw MRI images is 320×320 pixels, as Figure 1a shows. As the first step of preprocessing, the image is resampled radially in the midsagittal MRI cross-section image by forming radial lines from a visually selected circle center (see point A in Figure 1a). (Here and in the following, scaled figures are in pixels.) This is necessary to avoid the appearance of two contour points in the same column of the image – where the tongue contour bends back – that the edge-detecting algorithm could not handle. For the sake of clarity, lines are only shown by ten degrees in Figure 2a but in reality, resampling is done by one degree. Arranging the sections thus obtained in a Cartesian column, a matrix is gained. The resampled image is represented in the Descartes coordinate system (Figure 1b). The bottom line contains the center point while the top line represents the points of the circumference of the circle.

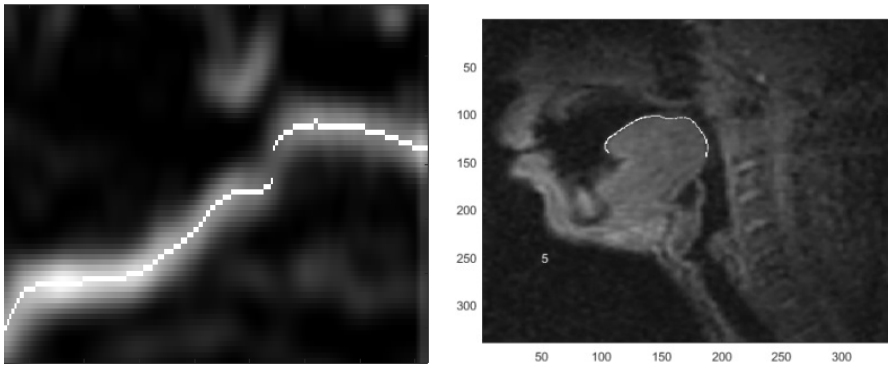


(a) (b)

Figure 1

Preprocessing, Step 1: (a) resampling the MRI image radially, (b) the Cartesian column matrix of the resampled image

In the second step, in the matrix, we find the largest cumulative luminance curve in the image obtained after edge enhancing with dynamic programming (Figure 2a). Processing is done from the left column to the right column of the image. The identified contour is indicated with white points in Figure 2a. The uneven tongue contour is smoothed by filtering before further processing. The smoothed tongue contour represents the base for the further analysis of articulation. In the image of Figure 2b, the tongue contour can be followed by projecting it back to the original image. The definition of the tongue contour offers an opportunity to perform various analyses.



(a) (b)

Figure 2

Preprocessing, Step 2: (a) The highlighted edge of the Cartesian image marks the found tongue contour with white points, (b) the tongue contour is projected onto the original image

2.1 Analysis of Tongue Position

The definition of the tongue contour makes it possible to calculate geometric features for a segmented part of it. Due to a lack of Hungarian recordings, a multilingual MRI visual database [9] was used to determine the tongue position associated with each speech sound as the male speaker produced vowels and VCV sound sequences (V: vowel, C: consonant). Through the exploration of the place of articulation in MRI images, we obtained static viseme data for each speech sound.

2.1.1 Method of Defining Tongue Position of Vowels with Quantitative Data

The idea was that by defining the centroid of the cross-section of the tongue body, we could obtain quantitative data about the horizontal and vertical positions of the tongue characteristic of the current speech sound. The centroid (C_{xy}) of the tongue is derived as the first-order momentum of the horizontal and vertical coordinates of the white points of the filled-up tongue body (1) as shown in Figure 3 [10, 11].

$$\boldsymbol{Cxy} = [\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}];\ \overline{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \boldsymbol{x} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}),\quad \overline{\boldsymbol{y}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \boldsymbol{y} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

where $f(x,y)=1$ in the white area, $f(x,y)=0$ outside the white area, and n is the number of white points.

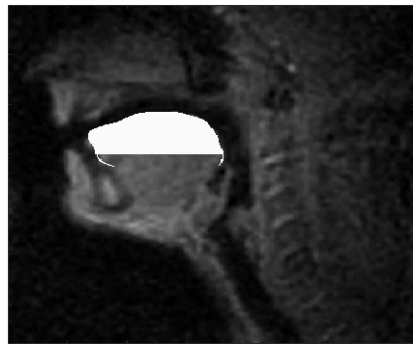


Figure 3
Filling the section of the tongue (sound /ɛ/)

We also need to determine the optimal number of pixel rows to fill the tongue body downwards from the top point of the tongue to define the centroid. Filling too few rows might lead to inaccurate measurement of the tongue position while filling too many rows might lead to oversimplification and losing the characteristic tongue position of distinct sounds.

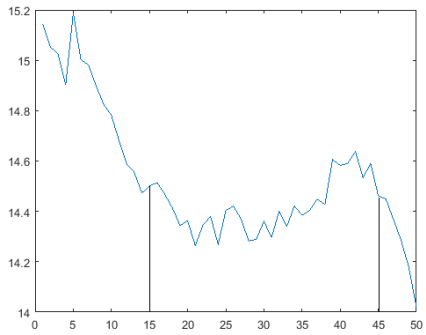


Figure 4
The variance of the centroids in pixels of the 28 vowels (vertical axis) of the database as a function of the depth of filling (horizontal axis)

To determine the filling depth, the standard deviation of the center of gravity of all vowels of the multilingual video database was examined, looking for a maximum for the highest distinction. In Figure 4, the variance (the average of deviations from the mean) decreases by filling over 45 rows of pixels as the tongue's root is less discriminative. Selecting less than 15 rows defines a cross-section representing just the top of the tongue and not the mass of it.

Based on the argumentation above, the tongue centroid for the vowels was investigated filling the depth of 42 rows of pixels for each vowel by the given resolution of the MRI image. In physical dimensions, the upper 22 millimeters of the tongue cross-section were selected.

2.1.2 Method of Defining the Place of Articulation for Consonants with Quantitative Data

The articulation of consonants is substantially different from that of vowels. This can be characterized by the place of articulation, which is determined by a gap or closure formed by the lip-tongue-jaw movement.

The place of articulation is determined by the narrowing or closure formed by the articulatory movements [12]. Thus, the place of articulation can be assigned to the place of the narrow part formed by the tongue and the unmoving part of the oral cavity.

The contour of the alveolar ridge and the palate can be defined analogously to the definition of tongue contour. The only difference is that moving away from the circle center we need to find not decreasing brightness – a falling edge – but rather increasing brightness – a rising edge. In images where the palate has no sharp edge the palate contours are defined by averaging several images.

Once we know the tongue contour and the palate contour, the distance of the two curves can be defined point by point. The distance measure derived from the definition of Nearest Neighbor Distance (NND) is suitable for determining the distance of curves consisting of a different number of points [13]. Let us take the two curves defined by their samples: U=[u_1, u_2, …, u_n] and V=[v_1, v_2, …, v_m]. Let the distance of one point of U from curve V be the distance of the point belonging to V nearest to it, and conversely, according to (2).

$$DU(i) = \min_j |u_i - v_j| \qquad DV(i) = \min_j |v_i - u_j| \tag{2}$$

Figure 5 shows the nearest neighbors of the tongue and palate.

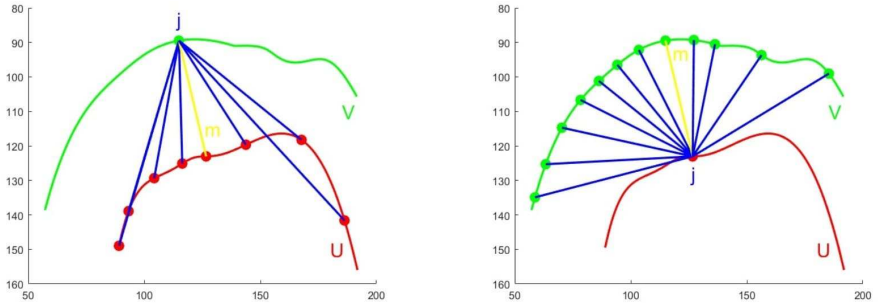
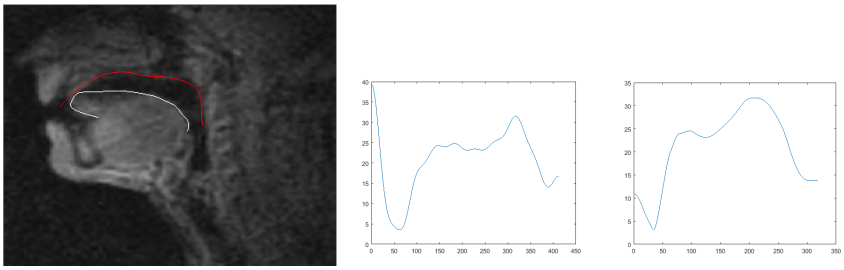


Figure 5
Graphical representation of NND

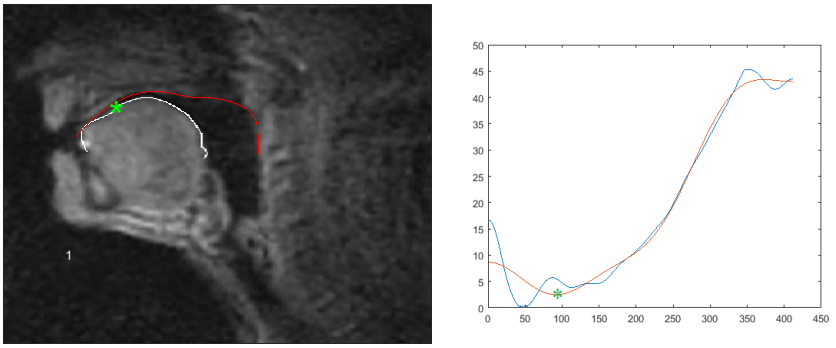
Figure 6a shows the alveolar ridge-palate contour (red line) and the tongue contour (white line). Figure 6b shows the minimum distance measured from the palate (vertical axis) to the points of the tongue (horizontal axis) while 5c shows the minimum distance measured from the tongue contour to the points of the palate. On the horizontal axis, the serial number of the tongue and palate contour points respectively, on the vertical axis the NND corresponding to the current contour point can be seen, measured in pixels. The place of articulation is considered the point of the tongue belonging to the smallest distance.



(a) (b) (c)

Figure 6
For the sound /r/: (a) The contour of the palate (red) and tongue contour (white), (b) the distance of the palate measured from the tongue, (c) distance of the tongue measured from the palate

With fricatives and approximants, a longer section of the tongue is close to the palate. With such sounds, it seems appropriate to regard the center as the middle of the whole near section. With low-pass filtering of the distance function, the place of the curve minimum can be shifted to the middle of the narrow section as is shown in Figure 7b. Discrete cosine transformation was used to filter the curve.



(a) (b)

Figure 7

(a) The tongue contour of sound /j/ and (b) the filtering of the distance function, (the blue line represents the original curve, and the red is the filtered one)

Selecting the appropriate frame of the video stream to represent the sound is a crucial point of this analysis. In the case of stop sounds and affricates, the frame before the burst is marked as the representative frame of sound; for the other consonants, the middle point of the time interval of the sound is selected.

3 Results

In our experiments, we obtained quantitative data from the determination of the tongue contour during speech. The results are shown separately for vowels and consonants, comparing them to traditional descriptive phonetic data.

3.1 Vowels

In the case of vowels, cross-section data formed along the longitudinal axis of the vocal tract are affected by jaw openness and tongue position. The narrower and wider sections of the vocal tract and the lips' shape determine the spectral properties of the excitation signal coming from the larynx [14]. Figure 8a shows the position of vowel articulation according to the phonetic parameters with the conventional representation in literature. This figure is adopted from the website of the International Phonetic Alphabet, (IPA).

On MRI recordings of vowel announcements, the tongue contour of the image taken from the center of the stationary phase of the sound was filled up to a line depth of 22 millimeters after automatic contour selection. Figure 8 shows the centroid of the tongue while pronouncing the sound /ɛ/.

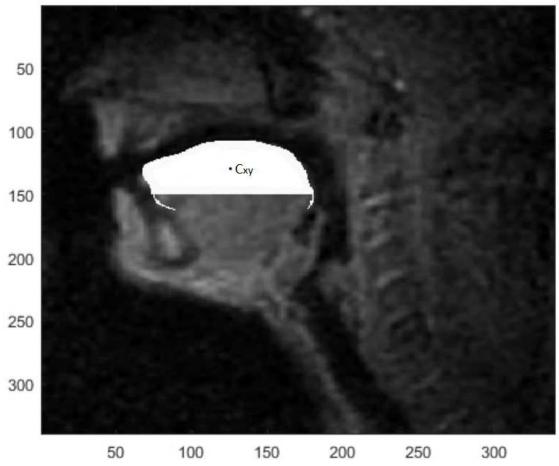


Figure 8
The centroid of sound /ɛ/

Figure 9b shows the C_{xy} centroid of the filled-up tongue in the oral cavity according to the 320×320 pixel coordinate system in, e.g. Figure 1a and Figure 2b, which visually reflects the phonetic arrangement of Figure 9a.

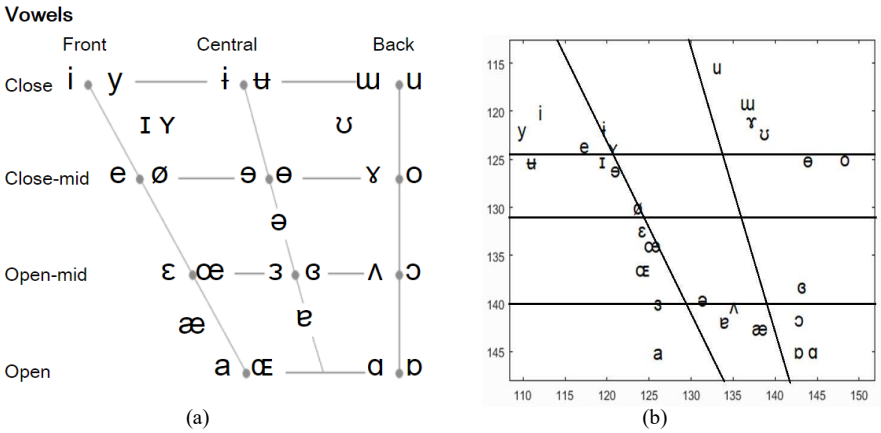


Figure 9
(a) Articulation map of vowels [15], (b) tongue centroids in the MRI images (the back part of the oral cavity is shown on the right, the front part on the left)

The traditional map of tongue position – divided vertically into four and horizontally into three sections – was compared with the measured centroid data.

In Table 1, the tongue positions located in the correct section are indicated in unshaded cells. Gray shading indicates a one-box difference either horizontally or vertically. The more white cells are in the table, the better the theoretical and the measured quantitative data match each other. No differences greater than one box were measured.

Table 1
The accuracy of the tongue positions

	horizontal	vertical		horizontal	vertical
i			o		
y			ə		1
ɨ			ɛ		
ʉ	1	1	œ		
ɯ			ɜ	1	
u			ɞ	1	
ɪ		1	ʌ	1	
ʏ			ɔ		
ʊ			æ	1	
e			ɐ		
ø			a		
ɘ			ɶ		1
ɵ	1		ɑ		
ɤ		1	ɒ		

3.2 Consonants

Table 2 shows the place of articulation of consonants. The tongue position of the bilabial and labiodental sounds was not examined. The reason for this is that in the formation of these sounds, the position of the tongue is indeterminate, that is, it adapts to the tongue position of the adjacent sounds. In these cases the place of articulation is not determined by the tongue but by the teeth and lips.

Table 1
Place of articulation of consonants [16]

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Figure 10 shows the places of articulation obtained with the method described in 2.1.2.

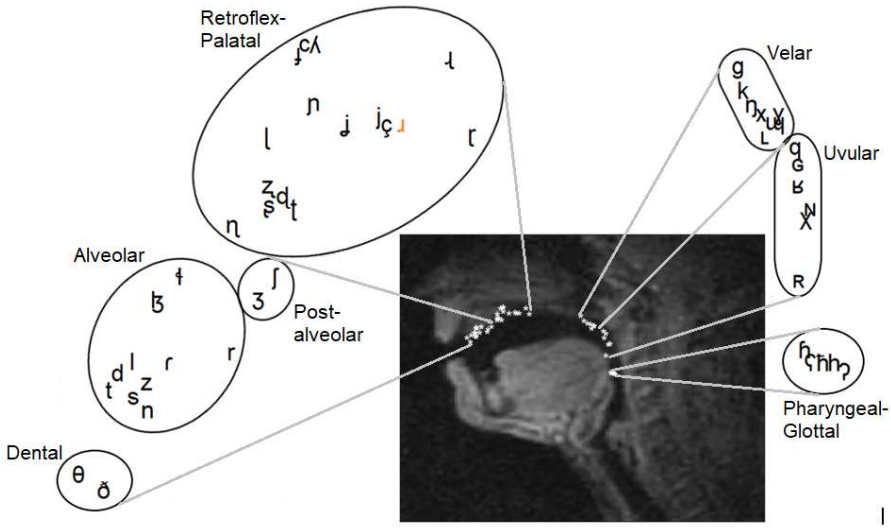


Figure 10
Calculated place of articulation of consonants marked in the MRI image

The theoretical and quantitative defined places of articulation differ only for a single sound /ɹ/ (marked with red in the figure). This means that the places of articulation defined with quantitative data match the physiological definitions well.

3.3 Application of the Results in the Speech Assistant System

The Speech Assistant system – elaborated for the Hungarian language to support the deaf in learning to speak – was further refined by incorporating the results of this work [17]. Figure 11 shows the visualized image of the reference pronunciation (bottom) and that of the sound recorded during practice (top), while on the right side, it displays the transparent talking head in two views. The talking head for the Shaanxi Xi'an dialect of Chinese and its Speech Assistant system is under development.

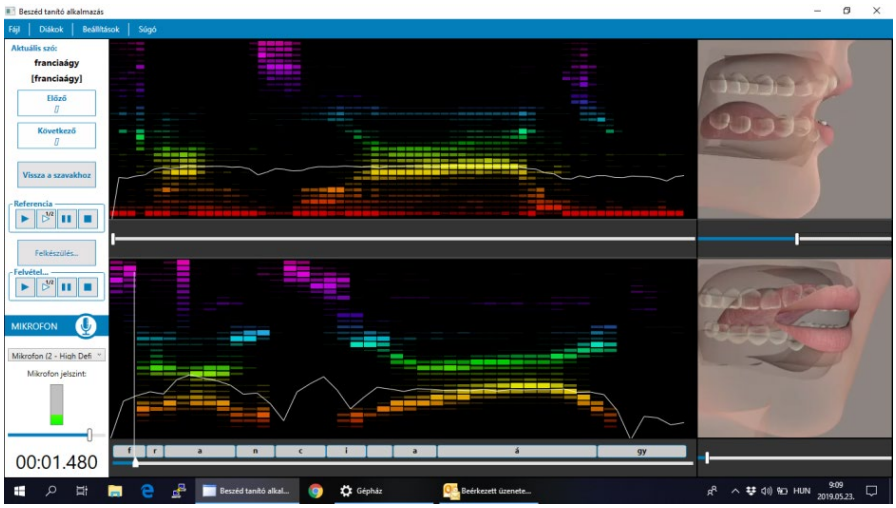


Figure 11
Screen view of the Speech Assistant during practice

Conclusions

A quantitative analysis of articulation was performed mainly to make the articulation visible in a transparent talking head. In the case of vowels, the tongue position was described with first-order momentums derived from MRI images. By consonants, the place of articulation was identified with the place of the closure or narrowing between the tongue and the alveolar ridge or the palate. The results of the approach confirm the suitability of quantitative analysis for verifying descriptive phonetic classifications. Thus, an important step towards the quantitative description of the articulation of speech production was taken. The traditional phonetic classification of speech sounds, the calculated place of articulation, and the tongue position defined by measurements are consistent with descriptions reported in the relevant literature. Supporting articulatory phonetics with quantitative data requires further, more detailed investigations. The native English speaker perfectly articulated the sounds of the International Phonetic Alphabet. The obtained results do not contradict the findings of the Hungarian

descriptive phonetic classifications. According to the testimony of Figure 9 and Table 1, the place of articulation of the vowels matches the descriptive phonetics data with at most one classification section error. The extension of the analysis to a larger number of speakers and different sound environments offers the possibility of improvement.

The results show that the analysis of the articulation of the Chinese Shaanxi Xi'an dialect speaker based on ultrasound images makes it possible to define the static data of visemes but also offers the opportunity to perform a dynamic description of the articulation. The presented analyses support visual speech synthesis with quantitative data that go beyond phonetical considerations.

References

[1] Barnaud, M. L., Schwartz, J. L., Bessière, P., and Diard, J. (2019) Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO, a perceptuo-motor Bayesian model of speech communication. PLoS ONE, 14, 1, https://doi.org/10.1371/journal.pone.0210302

[2] Miller, N. A., Gregory, J. S., Aspden, R. M., Stollery, P. J., & Gilbert, F. J. (2014) Using active shape modeling based on MRI to study morphologic and pitch-related functional changes affecting vocal structures and the airway. Journal of Voice, 28(5), 554-564, https://doi.org/10.1016/j.jvoice.2013.12.002

[3] Baum, S. R., Blumstein, S. E., Naeser, M. A., and Palumbo, C. L. (1990) Temporal dimensions of consonant and vowel production: An acoustic and CT scan analysis of aphasic speech. Brain and Language, 39(1), pp. 33-56, https://doi.org/10.1016/0093-934X(90)90003-Y

[4] Recasens, D. (1991) On the production characteristics of apicoalveolar taps and trills. Journal of Phonetics, 19(3-4), pp. 267-280, https://doi.org/10.1016/s0095-4470(19)30344-4

[5] Czap, L. (2021) Impact of Preprocessing Features on the Performance of Ultrasound Tongue Contour Tracking, via Dynamic Programming. Acta Polytechnica Hungarica, Vol. 18, No. 2

[6] Serrurier, A., Badin, P., Barney, A., Boë, L. J., & Savariaux, C. (2012) The tongue in speech and feeding: Comparative articulatory modelling. Journal of Phonetics, 40(6), 745-763, https://doi.org/10.1016/j.wocn.2012.08.001

[7] Hueber, T., Benaroya, E. L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010) Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Communication, 52(4), pp. 288-300, https://doi.org/10.1016/j.specom.2009.11.004

[8] Zhao, L., Czap L. (2019) A nyelvkontúr automatikus követése ultrahangos felvételeken. (Automatic tracking of the tongue contour on ultrasound

recordings.) *Beszédkutatás* 27(1), pp. 331-343, https://doi.org/10.15775/ Beszkut.2019.331-343

[9] sail.usc.edu/span/rtmri_ipa/je_2015.html, Accessed 18.02.2020

[10] Hu, M. K. (1962) Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, 8(2), pp. 179-187, https://doi.org/10.1109/TIT.1962.1057692

[11] Mukundan, R., and Ramakrishnan K. R. (1998) Moment functions in image analysis. Singapore: Word Scientific Press. pp. 11-24

[12] Erdogan, N., & Wei, M. (2019) Articulatory Phonetics: English Consonants. In Erdogan, N., & Wei, M. (Ed.), Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners pp. 263-284, IGI Global, http://doi:10.4018/978-1-5225-8467-4.ch011

[13] Zharkova, N., & Hewlett, N. (2009) Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /a{script}/. Journal of Phonetics, 37(2), pp. 248-256, https://doi.org/10.1016/j.wocn.2008.10.005

[14] Ivanova, S. A., & Hasko, V. (2019) Articulatory Phonetics: English Vowels. In Erdogan, N., & Wei, M. (Ed.), Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners pp. 285-301, IGI Global, http://doi:10.4018/978-1-5225-8467-4.ch012

[15] https://www.internationalphoneticalphabet.org/ipa-charts/vowels/ Accessed 22.04.2020

[16] https://www.internationalphoneticalphabet.org/ipa-charts/consonants/ Accessed 22.04.2020

[17] Czap, L., Pintér, J. M., & Baksa-Varga, E. (2019) Features and results of a speech improvement experiment on hard of hearing children. Speech Communication, 106, pp. 7-20, https://doi.org/10.1016/j.specom. 2018.11.003

The Enhancement of the Overall Group Technology Efficacy using Clustering Algorithm for Cell Formation

Lan Xuan Phung*, Trung Kien Nguyen, Son Hoanh Truong

School of Mechanical Engineering, Hanoi University of Science and Technology, Hanoi, 100000, Vietnam, lan.phungxuan@hust.edu.vn, trung.nguyenkien@hust.edu.vn, son.truonghoanh@hust.edu.vn
*Corresponding author

Abstract: Cellular manufacturing is a principal application of group technology in which machine cells and part families are generated based on their similarity in the production process to minimize overall movement cost and maximize machine utilization by using complex mathematical programming procedures or computer tools with a lot of computational effort and time to solve problems. In this study, the clustering analysis based on a similarity coefficient is developed to efficiently solve cell formation problems in both single and multiple process routings. A novel similarity coefficient is developed to integrate operation sequence especially adjacent operation, processing time, production volume, machine capacity, and multi-visits to minimize the number of actual inter-cell moves and voids in machine cells. An improved clustering algorithm is proposed for grouping machines into cells and simultaneously determining the machine sequence in cells to reduce intra-cell moves as well as selecting the best process routing for each part. The practical effectiveness of the proposed method is demonstrated through computational experiments involving eighteen test instances, varying in scale from small to large problems. When compared to other complex methods, the proposed approach not only enhances overall group technology efficacy but also significantly reduces computational time, making it a highly promising and practical solution for addressing cellular manufacturing challenges.

Keywords: cell formation; overall group technology efficacy; similarity coefficient; clustering algorithm

1 Introduction

Lean production focuses on optimizing efficiency, reducing costs, minimizing waste and improving overall quality. Cellular manufacturing (CM) is one of the key principles used to achieve these objectives and plays a crucial role in lean production. [1, 2]. CM is based on group technology principles by separating

machines into groups with similar characteristics and distributing parts into part families to achieve higher productivity and flexibility compared to traditional manufacturing [3, 4]. The principal problem in the CM system is cell formation to minimize the number of moves and voids in the cells, maximizing the utilization of machines, equipment, and labour in the production process. By grouping machines and processes in self-contained cells, CM reduces material handling and setup times. This results in a more streamlined and efficient production flow, minimizing the time required to move materials between workstations and improving overall process efficiency. For over three decades, various production-oriented methods have been introduced for cell formation problems such as mathematical programming, heuristic and (meta-) heuristic algorithm, graph partitioning, and most commonly hierarchical method. The mathematical programming methods focus on developing the model to maximize the total operations in each cell and/or to minimize the moves between cells [5, 6]. Some studies addressed heuristic algorithms based on flow-matrix to solve cell formation problems (CFP) and machine layout generation [7-9]. Another heuristic approach with two stages based on the similarity score was developed to produce the cellular facility layout [10]. Recently, (meta-) heuristic methods such as simulated annealing algorithm [11], genetic algorithm [12, 13], ant colony algorithm [14], etc. have been introduced as promising methods to obtain "acceptable" solutions with the optimization in "reasonable" computational time. Adaptive resonance theory, a class of neural networks, has also demonstrated the ability to solve the CFP [15, 16]. The graph partitioning approach represents a graph with nodes and arc weight defined as machines and similarity measure of machine pair, respectively to minimize inter-cell travels [17, 18]. In other work, the hybrid algorithm was developed to solve complex, multiple objective optimizations for the CFP [19]. The hierarchical method is the main approach to cluster analysis in which the similarity or distance function and hierarchy of clusters are determined [20, 21]. Besides time-saving and computation calculation, hierarchical clustering methods based on similarity coefficient (SC) are more flexible in integrating various production data into CFP such as operation sequence, production volume, processing time, machine capacity, etc. [22-24].

To address the problem in the hierarchical method, most researchers start from the machine-part matrix to get a transformed matrix in a more structured form as diagonal blocks by grouping machines into cells and parts into families to minimize the number of out-of-blocks referred to as exceptional elements. There are four steps in this method: production data collection, SC determination for each machine pair, clustering algorithm applying for cell formation based on SC, and part family generation. The most important production input data, machine-part matrix MP $[m_{ij}]$, is first determined. The binary matrix is commonly used as an MP matrix in simple CFP, where $m_{ij} = 1$, if part i proceeded on machine j and otherwise [14, 25]. In other work, a number presented for the operation sequence index is used instead of using the value "1" in the MP matrix [26-28]. Besides the MP matrix, processing time, and production volume, machine capability is also initial

production data for the problem. Based on production input data, the SC is determined by different formulas in the literature. McAuley developed the original SC that considers only the number of machine pairs proceeding with both operations and/or at least one operation [25]. T. Gupta and H. I. Seifoddini incorporated processing time and production volume in the SC formula [26]. The operation sequence and multi-visit problem are also considered in the SC formula [27]. This study focuses on two specific operations: the first and last operations, primarily aimed at reducing the number of inter-cell moves. However, the adjacent operation, which plays a crucial role in diminishing duplicate or repeated inter-cell moves, has not been addressed.

In the next step, the clustering algorithms based on the SC are applied to separate and group machines into cells. There are some common algorithms for group machines to form machine cells (MC) such as the single linkage algorithm (SLINK) [25] and the average linkage algorithm (ALINK) [28]. Linear cell clustering (LCC) is a method employed to generate consistent machine groups by linearly comparing the similarity scores between two machines [20]. LCC is known as a fast method in computing, a simple algorithm, and an easy solution in programming. Finally, parts can be allocated to families corresponding with assigned machines to optimize the inter-cell moves and voids in MC. In practical works, each part may have alternative process routings that make CFP more complicated. Thus, the formation of machine cells (MCs), part families, and selection of best routing need to be considered in CFP to achieve the overall objectives [29-31]. Intra-cell moves are classified into two main types: forward moves, including in-sequence and by-pass movements, and backward moves. Due to the reverse material flow, the cost of backward moves is significantly higher than that of other moves. Thus, minimizing backward moves becomes a key factor in reducing intra-cell move cost. However, most previous clustering algorithms mainly focus on reducing the inter-cell moves regardless of paying attention to minimising the backward moves and the compactness in CFP for the multi-routing problem. In the proposed work, the operation sequence, multi-visits, multi-routings, processing time, production volume and machine capacity are integrated into a similarity coefficient in a unique model to solve CFP for both single-routing and multi-routing problems.

To evaluate the group technology performance, several measuring methods are used in the literature. Three types of evaluation performance measurements were used including global efficiency, group efficiency, and group technology efficiency [32]. These measurements are quite individual and insufficient to provide an overall evaluation of the effectiveness of cell formation. To solve this problem, Nair and Narendran developed bond efficiency incorporating both inter-cell moves and compactness [27]. Lee and Anh proposed group technology efficacy (GTE) as shown in Eq. 1 for the performance measurement of cell formation considering both actual inter-cell moves and cell compactness [33]. Based on Lee's GTE, S. Raja defined GTE considering the backward moves as presented in Eq. 2 [34]. In this measurement, the effect of inter-cell moves, and backward moves are equal

although they are quite different in the actual production. Moreover, the definition of backward move in this study only covers the operations inside a cell containing the part while this move exists even for external operations of the part in a different cell.

$$Lee's\ GTE = \frac{1 - \frac{AIM}{PIM}}{\frac{NV}{NI}} \tag{1}$$

$$Raja's\ GTE = \frac{1 - \frac{AIM + BM}{PIM}}{\frac{NV}{NI}} \tag{2}$$

$$PIM = NO - NP$$

where AIM is number of actual inter-cell moves; BM is number of backward moves; PIM is number of possible internal moves; NV is the number of voids; NI is number of operation inside machine cells; NO is the number of operation outside machine cells; NP is the number of parts.

The purpose of this article is to solve CFP by considering the actual inter-cell moves and actual backward moves and the compactness through the novel SC and modified clustering algorithm. In this study, both cell formation and machine sequence are solved simultaneously in consideration of the most important factors including operation sequence, processing time, machine capacity, production volume, multi-visits and multi-routings. The results contribute to increasing the overall GTE for both single routing and multi-routing problems.

2 Methodology

2.1 Proposed Similarity Coefficient

Besides considering the important factors such as operation sequence, processing time, production volume, and multi-visits, the proposed SC integrates the key factors including adjacent operation and the compactness in the calculation. The proposed SC between machine j and machine k is determined by S_{jk} as shown in Eq. 3.

$$S_{jk} = \frac{\sum_{i=1}^{n} A_{jk}\, w_i}{\sum_{i=1}^{n} A_{jk}\, w_i + \sum_{i=1}^{n} O_{jk}\, w_i + C_{jk}/N} \tag{3}$$

where A_{jk} refers to the total actual flow through machine j proceeded by part i which uses both machine j and machine k as calculated in Eq. 4; O_{jk} refers to the actual flows to or from machine j only (excluding machine k) made by the part i;

C_{jk} is the void possibility between machine j and machine k; N is the total routes of all parts; w_i is the production volume of part i.

$$A_{jk} = \sum_{r=1}^{n_{ir}} \left(A_{irjk} + A_{irkj}\right)\left(T_{irjk} + Z_{irjk}\right) \tag{4}$$

$$A_{irjk} = \sum_{p=1}^{n_{irj}} \sum_{q=1}^{n_{irk}} a_{ij}^{pq}$$

$$\text{where } a_{ij}^{pq} = \begin{cases} 0 \; if \; b_{irj}^{p} = 0 \; or \; b_{irk}^{q} = 0 \\ 1 \; if \; b_{irj}^{p} = 1 \; or \; r_i \\ 2 \; otherwise \end{cases} \tag{5}$$

where A_{irjk} indicates the total actual flow through machine j proceeded by part i which uses both machine j and machine k in the route r as shown in Eq. 5 n_{ir} refers to the number of routes associated with part i; n_{irj} is the number of operations in which part i processes on machine j in the route r; a_{ij}^{pq} indicates the flows through machine j proceeded by part i which uses both machine j for time p and machine k for time q in the route r; b_{irj}^{p} indicates the operation index if part i moves through machine j for time p in the route r;

T_{irjk} is the proportion of minimal and maximal total processing time of part i in the route r spending on both machine j and k as shown in Eq. 6.

$$T_{irjk} = \frac{min\left(\sum_{r=1}^{n_{ir}} \sum_{p=1}^{n_{irj}} \frac{t_{irj}^{p}}{C_j}, \sum_{r=1}^{n_{ir}} \sum_{q=1}^{n_{irk}} \frac{t_{irk}^{q}}{C_k}\right)}{max\left(\sum_{r=1}^{n_{ir}} \sum_{p=1}^{n_{irj}} \frac{t_{irj}^{p}}{C_j}, \sum_{r=1}^{n_{ir}} \sum_{q=1}^{n_{irk}} \frac{t_{irk}^{q}}{C_k}\right)} \tag{6}$$

where t_{irj}^{p} is the processing time if the part i uses machine j for the time p in the route r; C_j is the machine capacity of machine j

Z_{irjk} is the proportion considering the adjacent operation between both machines j and k for part i in route r as shown in Eq. 7. It is the key factor to reduce the inter-cell moves.

$$Z_{irjk} = \frac{\sum_{p=1}^{n_{irj}} \sum_{q=1}^{n_{irk}} z_{irj}^{1\text{pq}}}{\sum_{p=1}^{n_{ij}} \sum_{q=1}^{n_{ik}} z_{irj}^{2\text{pq}}}$$

$$\text{where } z_{irj}^{1\text{pq}} = \begin{cases} 1 \; if \; (both \; b_{irj}^{p} \; and \; b_{irk}^{q} \neq 0 \; and \; \left|b_{irj}^{p} - b_{irk}^{q}\right| = 1)\,) \\ 0 \; otherwise \end{cases} \tag{7}$$

$$z_{irj}^{2\text{pq}} = \begin{cases} 1 \; if \; both \; b_{irj}^{p} \; and \; b_{irk}^{q} \neq 0 \\ 0 \; otherwise \end{cases}$$

where $z_{irj}^{1\text{pq}}$ is the number of adjacent operations between machine j in the time p and machine k in the time q in the route r by part i. $z_{irj}^{2\text{pq}}$ is the number of operations proceeded on both machine j in the time p and machine k in the time q in the route

r by part i. Only the actual flows to or from machine j (O_{jk}) is calculated by Eq. 8; where O_{irjk} is the actual flows to or from machine j (excluding machine k) made by the part i in the route r as shown in Eq. 9; o_{irj}^{pq} is the flow through machine j in the time p (excluding machine k in the time q) made by part i in the route r

$$O_{jk} = \sum_{r=1}^{n_{ir}} \left(O_{irjk} + O_{irkj}\right) \tag{8}$$

$$O_{irjk} = \sum_{p=1}^{n_{irj}} \sum_{q=1}^{n_{irk}} o_{irj}^{\mathrm{pq}}$$

$$\text{where } o_{irj}^{\mathrm{pq}} = \begin{cases} 0 \; if \; b_{irj}^{p} = 0 \; or \; b_{irk}^{q} \neq 0 \\ 1 \; if \; b_{irj}^{p} = 1 \; or \; r_i \\ 2 \; otherwise \end{cases} \tag{9}$$

C_{jk} is defined by the absolute value of the difference between the total number of parts visiting machine j and the total number of parts visiting machine k in route r described in Eq. 10. This factor takes into account the impact of voids in cells.

$$C_{jk} = \left| \sum_{i=1}^{n} \sum_{\mathrm{r}=1}^{n_{ir}} c_{irj} - \sum_{i=1}^{n} \sum_{\mathrm{r}=1}^{n_{ir}} c_{irk} \right|$$

$$\text{where } c_{irj} = \begin{cases} 1 \; if \; m_{irj} \neq 0 \\ 0 \; otherwise \end{cases} \tag{10}$$

c_{irj} is the number of operations proceeded by part i by machine j in the route r

2.2 Improved Clustering Algorithm

The clustering algorithm is modified from the LCC algorithm by incorporating the sort algorithm, which facilitates the simultaneous identification of the appropriate machine positions within the cell during MC formation, and the selection algorithm, used for determining the best routing. The flowchart of modified clustering algorithm for grouping and generating machine cell is shown in Figure 1 and the algorithm includes the following steps:

Step 1: Acquire the matrices for machine-part ($MP[ir, j]$), process time ($PT[ir, j]$), production volume ($PV[i]$), and machine capacity ($C[j]$). Set value for the similarity coefficient threshold ($sThreshold$) for group merging considerations utilized at step 3b and weight factor (q) as the ratio between backward and inter-cell move cost used at step 6.

Step 2: Calculate S_{jk} for each machine pair j and k and generate a similarity coefficient matrix. Arrange the SC in decending order.

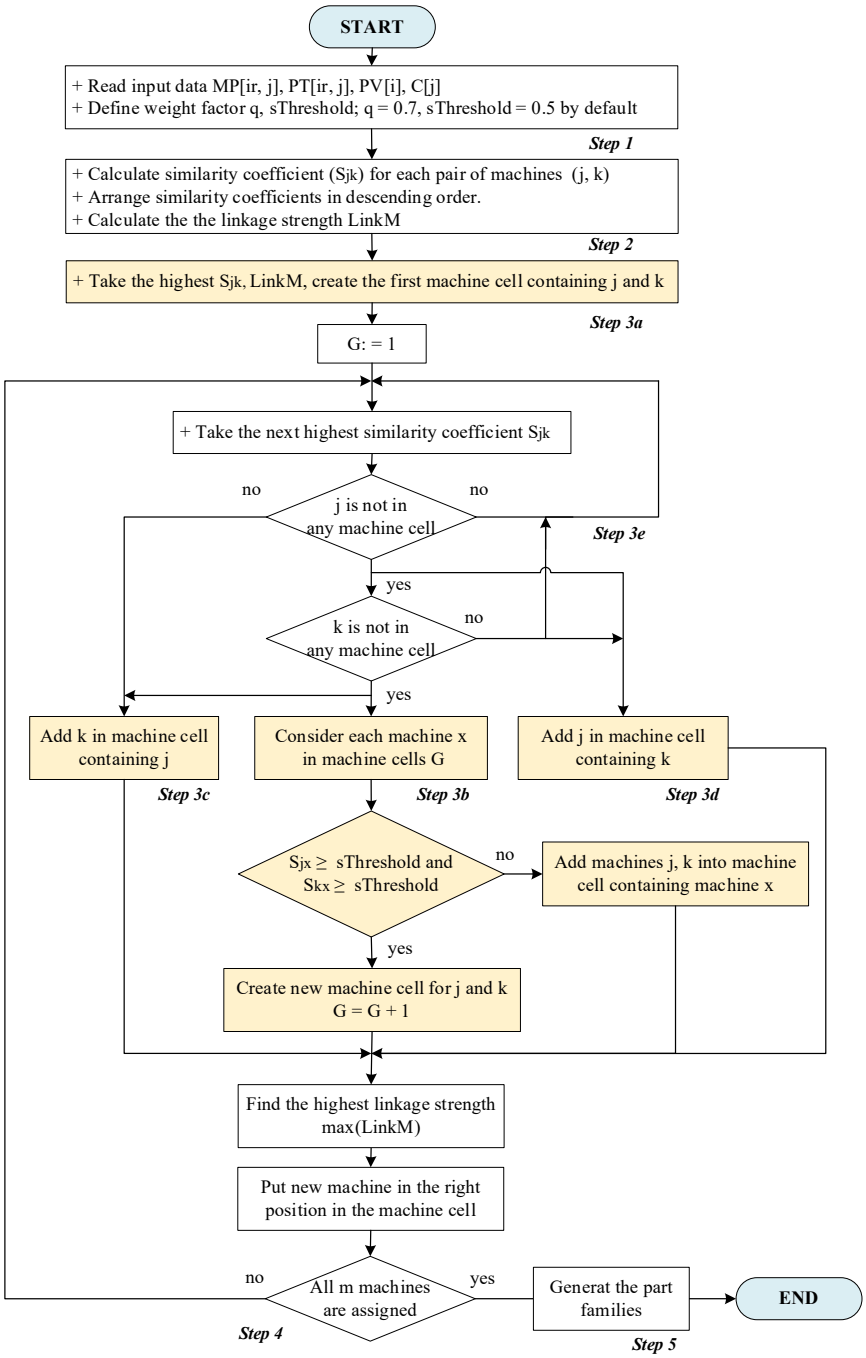


Figure 1
The flowchart of modified clustering algorithm for generating the machine cell

Step 3: Cluster machines into MC by considering the machine pair with the highest S_{jk}. During processing, the position of a machine in the assigned MC is also determined by calculating and comparing the linkage strength ($LinkM_{jk}$) between machines j and machine k. The order with the highest $LinkM_{jk}$ is given priority. With machines j and k in a cell, the linkage strength $LinkM_{jk}$ is calculated based on the total backward moves between them as shown in Eq. 11.

$$LinkM_{jk} = \sum_{i=1}^{n} \sum_{\mathrm{r}=1}^{n_{ir}} \sum_{p=1}^{n_{irj}} \sum_{q=1}^{n_{irk}} L_{jkpq}^{ir}$$

where L_{jkpq}^{ir}

$$= \begin{cases} = 1 \; if \; (b_{irk}^{q} - b_{irj}^{p} = 1 \; and \; b_{irj}^{p} \neq 0 \; and \; b_{irk}^{q} \neq 0)\,) \\ = 0.25 \; if \; (b_{irk}^{q} - b_{irj}^{p} \geq 2 \; and \; b_{irj}^{p} \neq 0 \; and \; b_{irk}^{q} \neq 0)\,) \\ = 0 \; otherwise \end{cases} \tag{11}$$

In this step, there are some cases as follows:

3a) If no MC is generated, the first MC is created containing machines j and k. The order in the first cell can be (j, k) or (k, j) depending on the comparison values of $LinkM_{jk}$ and $LinkM_{kj}$ as above-mentioned.

3b) If there is at least one generated MC and both machines j and k have not been assigned to any MC yet. Then, the merging group process is applied to determine whether both machines are assigned in a new MC or a pre-generated MC. Let x be any machine in generated MC. If both S_{jx} and S_{kx} are larger than $sThreshold$ and the number of machines in the merged cell is not larger than the maximum expected the number of machines in cells, machine j and k are merged into the assigned MC. Otherwise, a new MC is created for the machine pair.

3c) If machine j has already belonged to the cell, and machine k has not been assigned. Machine k is allocated to the cell that includes machine j. The position of machine k in MC depends on the value of $LinkM_{kx}$ and $LinkM_{xk}$ between k and each assigned machine (machine x) in the MC to find the position having the highest linkage strength value.

3d) If machine k has already belonged to the cell, and machine j has not been assigned. Machine j is assigned to the cell that includes machine k.

3e) If both machine j and machine k have already belonged to the same cell. Therefore, the machine pair can be ignored and go to the next step.

3f) If machines j and k have been allocated to two distinct cells. This information can be reserved for future processes such as merging two MCs with specific conditions.

Step 4: Repeat step 3 with the next highest S_{jk} until all m machines have been assigned to MCs.

Step 5: Assign parts corresponding to MCs to generate part families. The following sub-steps are required to assign parts into part families corresponding to the generated MCs. For special cases, further process such as merging cells is applied.

5a) For each part i with route r and each machine cell, determine NV_{ir} representing the total number of machines that are not visited by part i, and NO_{ir} which denotes the sum of operations of part i outside this cell.

5b) Part i with route r is assigned to the part families corresponding to the machine cell where the sum NVO_{ir} is minimal as calculated by Eq. 12.

$$NVO_{ir} = NV_{ir} + NO_{ir} \tag{12}$$

5c) Repeat step 5a until all parts are allocated in part cells.

For parts with multi-routings, the following processes in step 6 are proceeded to select the best routing.

Step 6: Select routings for each part

6a) For each part i in the route r, determine $SAIB_{ir}$ that is calculated by the weighted sum of the total number of actual inter-cell moves (AIM_{ir}) and the total number of actual backward moves (ABM_{ir}) as shown at Eq. 13.

$$SAIB_{ir} = AIM_{ir} + q.ABM_{ir} \tag{13}$$

6b) Among the various routings, the route with smallest $SAIB_{ir}$ is chosen. In the case of the same $SAIB_{ir}$, two additional sub-conditions are employed. The first sub-condition involves evaluating the ratio of voids for route r, denoted as $RNVI_{ir}$ and calculated using Eq. 14. The second sub-condition pertains to the processing time objective, measured as the total processing time in route r (ST_{ir}) and computed using Eq. 15. Depending on the selected objective, one of two sub-conditions is compared across various routings to determine the optimal routing, which yields the smallest value.

$$RNVI_{ir} = \frac{NV_{ir}}{NI_{ir}} \tag{14}$$

$$ST_{ir} = \sum_{j=0}^{m} \sum_{p=1}^{n_{irj}} t_{irj}^{p} \tag{15}$$

where NV_{ir} is number of voids for each part i in the route r; NI_{ir} is number of operations inside machine cells for part i in the route r

6c) Repeat Step 6b until all parts can choose the best routing.

Step 7: Stop

2.3 Modified Group Technology Efficacy (MGTE)

The modified group technology efficacy ($MGTE$) is introduced to integrate the actual backward moves (including external operations) and the weigh factor q. In practical production, the inter-cell move is the main concern and has the highest effects on the travelling cost in CFP. Thus, the inter-cell move should have a stronger effect than other factors and q should be added in $MGTE$ calculation. Depending on the pratical inter-cell and intra-cell move cost, the q value can change to meet specific production conditions but is not fixed for all cases. Using the proposed $MGTE$, the overall group technology efficacy can be evaluated as shown in Eq. 16.

$$MGTE = \frac{1 - \frac{AIM + q.ABM}{PIM}}{1 + \frac{NV}{NI}} \tag{16}$$

3 Illustrative Examples

To explain the calculation procedure of the proposed method, example 1 was generated with six parts and five machines for single routing, incorporating multi-visits, Subsequently, example 2 utilized production data featuring seven parts, ten machines, and fourteen alternative routings.

3.1 Example of Single Routing and Multi-Visits (Example 1)

Random production data is generated with a machine-part matrix of 6x5, indicating the operation sequence, and the production volume is shown in Table 1. In this example, all parts have the single routing. The SC calculation in step 2 is applied for all machine pairs, and the results are shown in Table 2. At step 3a, S_{14} is determined as the highest value (0.8703), so machines 1 and 4 should be grouped in the first machine cell. The order of machines 1 and 4 is determined by comparing the linkage strength calculation of the machine pair. Because $LinkM_{14} = 2$ and $LinkM_{41} = 1$, machine 1 should have been in front of machine to obtain the highest linkage strength value. S_{23} is the second-highest value (0.8481), so machines 2 and 3 should have been grouped in the second machine cell. The linkage strength between machines 3 and 2 is higher than that between machines 2 and 3. Hence, machine 3 should have been in front of machine 2. A similar process is executed for the other machine pairs, and two MCs are finally generated, including (1, 4) and (3, 2, 5) at the end of step 4. The parts then are assigned to the MC in the step 5. For each part, NV_{ir} and NO_{ir} for each machine cell are calculated to assign the part to the specific MC with the smallest sum of NV_{ir} and NO_{ir}. Finally, two part families are generated, including (3, 5, 6) and (1, 2, 4) according to two machine cells.

Table 1
The production data for Example 01

Part i	Machine					PV
	M1	M2	M3	M4	M5	
P1	3	1	2			160
P2		4	1, 3		2	310
P3	2, 4	1		3		280
P4		1		3	2	265
P5				2	1	80
P6	1			2		150

Table 2
Similarity coefficient matrix for Example 01

Machine	Machine				
	M1	M2	M3	M4	M5
M1	1	0.5892	0.2941	**0.8703**	
M2	0.5892	1	**0.8481**	0.3948	0.7611
M3	0.2941	**0.8481**	1		**0.7777**
M4	**0.8703**	0.3948		1	0.5406
M5		0.7611	**0.7777**	0.5406	1

3.2 Example of Multi-Routings and Selected Objective (Example 2)

The second example with a test instance size of 7x10x14 uses a production data sample from the existing literature, including 7 parts and 10 machines and 14 alternative routings [8]. The processing time for all operations is assumed to be the same. According to the calculation of the SC matrix, the machine pairs (8, 3), (6, 4), (9, 6), and (7, 10) have highest SC value. Therefore, the machine pair (8, 3) should be assigned to the first MC and the machine pair (6, 4) assigned to the second MC since the merging condition at step 3d is not satisfied.

The next highest SC is for machine pair (9, 6). Machine 6 has already been assigned to the second MC. Then machine 9 is assigned to the second MC. The machine pair (7, 10) has not been assigned to any MCs. The merging condition of this pair is satisfied because both SC values for machine pairs (7, 8) and (10, 8) are higher than the $sThreshold$ value (0.5). Hence, they are added to the first MC with the machine pair (8, 3). The process continued until all machines are assigned to MCs. At the end of step 4, two MCs corresponding with two-part families for all routings are generated, as shown in Table 3. During step 6, $SAIB_{ir}$ for all parts with alternative routings are calculated as Eq. 13. The routing with lowest $SAIB_{ir}$ is selected. For parts 3, 5 and 7, two alternative routings have the same $SAIB_{ir}$. Thus, the $RNVI_{ir}$

and ST_{ir} are calculated corresponding with two objectives: the compactness and processing time. To obtain the MC compactness for the case (7x10x14$^{(C)}$), the smallest $RNVI_{ir}$ is more important than ST_{ir}. Two-part families are determined: (P1-R1, P2-R3, P3-R2, P5-R1) and (P4-R1, P6-R1, P7-R2). To achieve the processing time objective for the case (7x10x14$^{(T)}$), the smallest ST_{ir} has a higher priority and two families are identified: (P1-R1, P2-R3, P3-R2, P7-R1) and (P6-R1, P4-R1, P5-R2). Two machine cells with machine layout for case 7x10x14$^{(C)}$ and examples of material flow for P2-R3 and P7-R2 are shown on Figure 2.

Table 3
Cell formation for all multi-routings

Part	Route	Machine										$SAIB_i$	$RNVI_i$	ST_{ir}
		M3	M5	M7	M8	M10	M1	M2	M4	M6	M9			
P1	**R1**		**3**	**4**		**5**	**1**		**2**			**1**		
P2	**R3**	**2**	**3**	**4**	**5**	**6**		**1**				**1**		
P3	R2	1	2		3	5					4	2	**0.25**	**5**
P5	R1	1		3	4	5				2		2	**0.25**	*5*
P6	**R1**			**3**		**4**		**1**	**2**			**1**		
P7	R1	1			2	3						0	*0.67*	**3**
P1	R2	2			4			1		3	5	4		
P2	R1		4			6	1	2	3	5		3		
P2	R2	2					1		3	4	5	2		
P3	R1			3	4		1		2		5	2	*0.67*	*5*
P4	**R1**							**1**	**2**	**3**	**4**	**0**		
P5	R2			3			1	2			4	**2**	*0.67*	**4**
P6	R2				2		1				3	2		
P7	R2							1	2	3	4	0	**0.25**	*4*

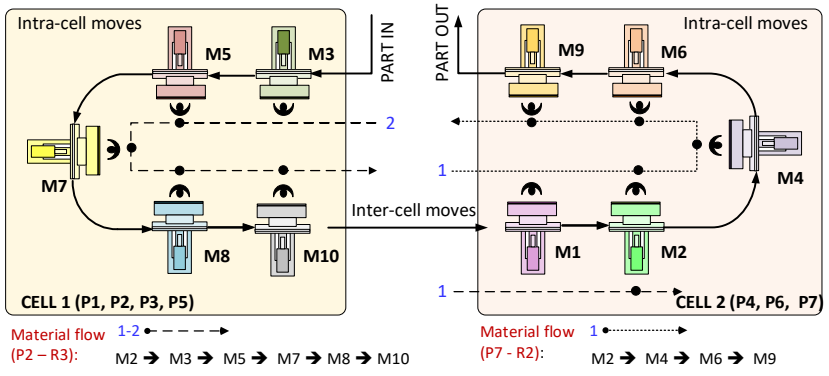


Figure 2
Cell formation and machine layout (7x10x14$^{(C)}$) and material flows for P2 - R3 and P7 - R2

4 Comparison Results and Discussion

The proposed method used seventeen common problems from the literature from small size (5x4) to large size (51x20) for the evaluation and comparison. Three kinds of GTE values including Lee's GTE [33], Raja's GTE [34] and proposed MGTE are calculated for the comparison. A developed software built by Visual C# allows to quickly calculate and display the group technology results based on the proposed method. Figure 3 shows the final solution of case 40x25 in the developed application.

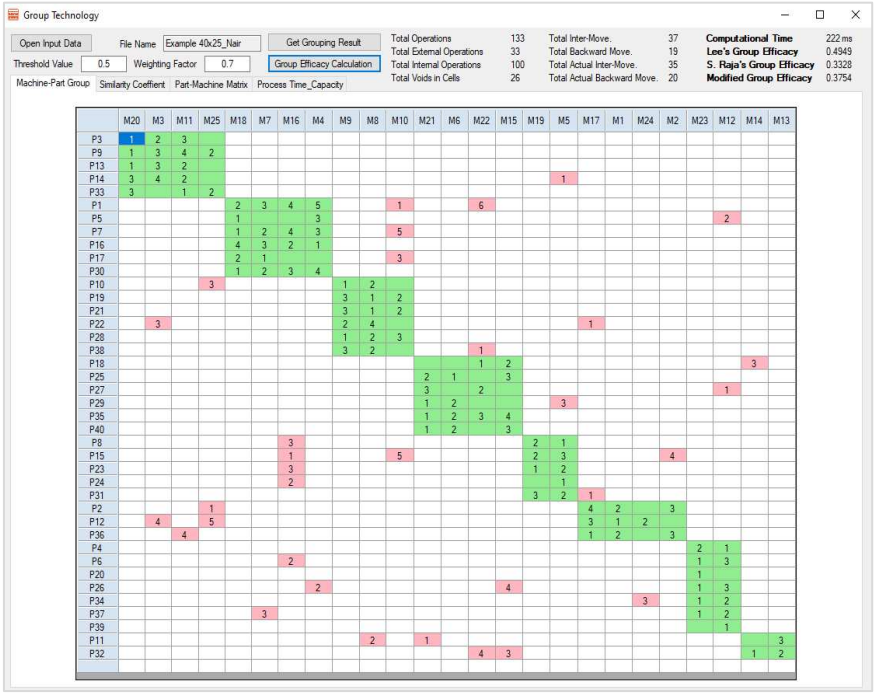


Figure 3
The result of the production data size of 40x25 in the developed application

Table 4 displays the computational results obtained by the proposed method for each test instance, along with a comparison to the best results achieved by other approaches in the literature. The proposed method generally outperforms previous studies in terms of AIM and ABM, except for the 35x18 and 43x16 examples. Although these two cases exhibit a higher number of AIM, they demonstrate superior compactness of machine cells with significantly lower NV. Furthermore, for the 40x25 example, the proposed method yields machine cells with one additional ABM, while maintaining four fewer AIM compared to the literature's methods. Consequently, the proposed method consistently achieves minimal overall

moves while ensuring higher compactness. The results demonstrate that the proposed method produces solutions with lower weighted overall moves and improved compactness across most test instances.

Table 4
Comparison of the proposed method with other methods in detail

Test No.	Size PxMxR	Best result/**Proposed Method**						Method	Best source
		NC	NO	NI	NV	AIM	ABM		
1	6x5x8	-/**2**	-/**4**	-/**14**	-/**3**	-/**4**	-/**3**	-	-
2	5x4x10	2/**2**	0/**0**	9/**9**	1/**1**	0/**0**	2/**2**	Mathematical model	[35]
3	9x9	2/**2**	4/**5**	29/**28**	20/**14**	4/**5**	1/**0**	Similarity coefficient	[24]
4	12x10	3/**3**	4/**5**	34/**33**	7/**8**	8/**5**	9/**3**	Two-mode similarity coefficient	[36]
5.1	7x10x14$^{(C)}$	-/**2**	-/**7**	-/**26**	-/**9**	-/**7**	-/**0**	-	-
5.2	7x10x14$^{(T)}$	-/**2**	-/**7**	-/**24**	-/**11**	-/**7**	-/**0**	-	-
5.3	7x10x14	2/**3**	8/**8**	22/**23**	12/**11**	11/**9**	0/**0**	Heuristic algorithm	[8]
6	18x10	3/**3**	6/**6**	50/**50**	13/**14**	7/**7**	14/**12**	Similarity coefficient	[21]
7	19x12	3/**3**	26/**20**	53/**59**	16/**30**	28/**22**	7/**6**	Simulated annealing	[11]
8.1	12x12x20$^{(C)}$	3/**3**	15/**15**	31/**32**	30/**16**	14/**18**	12/**4**	Similarity coefficient	[31]
8.2	12x12x20$^{(T)}$	3/**3**	15/**15**	31/**29**	30/**19**	14/**18**	12/**5**	Similarity coefficient	[31]
9	20x8	3/**3**	9/**9**	52/**52**	0/**0**	16/**16**	8/**8**	Flow matrix	[34]
10.1	8x9x20$^{(C)}$	2/**2**	2/**2**	26/**26**	14/**10**	2/**2**	0/**0**	Tabu search algorithm	[29]
10.2	8x9x20$^{(T)}$	2/**2**	2/**2**	26/**24**	14/**11**	2/**2**	0/**0**		
11	10x10x24	3/**3**	2/**2**	30/**30**	3/**3**	2/**2**	1/**1**	Tabu search algorithm	[37]
12	16x10x32	2/**2**	5/**5**	66/**67**	17/**17**	6/**6**	18/**11**	Similarity coefficient	[31]
13	20x20	5/**5**	14/**14**	65/**65**	18/**18**	18/**18**	21/**11**	Two-mode similarity coefficient	[36]
14	35x18	4/**5**	49/**44**	118/**123**	91/**21**	54/**60**	29/**27**	Genetic algorithm	[13]
15	40x25	8/**8**	33/**33**	100/**100**	23/**26**	39/**35**	19/**20**	Flow matrix	[34]
16	43x16	4/**5**	28/**32**	119/**115**	107/**53**	37/**44**	19/**20**	Similarity coefficient	[31]
17	45x20	4/**5**	31/**30**	129/**130**	65/**66**	41/**41**	41/**22**	Genetic algorithm	[13]
18	51x20	5/**5**	42/**30**	138/**150**	65/**84**	46/**37**	27/**29**	TOPSIS	[38]

PxMxR: *Part number* x *Machine number* x *Route number*; *(C): Compactness objective; (T): Processing time objective; NC: Number of cells*

Figure 4 presents the comparison of the proposed method's results against other approaches, specifically for the 40x25 test instance, using three GTE measures. The consideration of adjacent operation in the SC formula can significantly reduce the total number of AIM resulting in the best Lee's GTE in comparison with other methods in the literature. Raja's GTE and proposed MGTE integrated ABM in measures are also higher in the proposed method than in other previous approaches. Table 5 presents a comparative analysis of the proposed method against the best

solutions from the literature, focusing on Lee's GTE, Raja's GTE, and the proposed MGTE measures across eighteen instances. The comparison results consistently demonstrate the superiority of the proposed method over other approaches in terms of Raja's GTE and the proposed MGTE. Notably, in instances 18x10 and 45x20, the proposed method achieves smaller Lee's GTE but higher Raja's GTE than other methods. This is attributed to the proposed method yielding a final solution with one more NV while maintaining the same AIM and fewer ABM. In instances 35x18 and 43x16, the proposed method exhibits a higher number of AIM due to a higher NC. However, the substantial reduction in NV across all machine cells leads to significantly higher GTE in all measure types compared to existing methods.

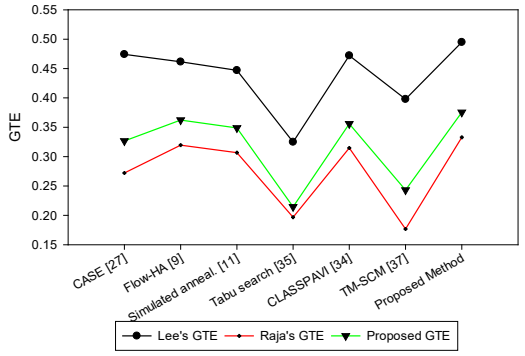


Figure 4
Comparison of the GTEs obtained by the proposed method over existing methods for problem 40x25

Furthermore, in order to assess the efficacy of the proposed method when dealing with a large-scale problem, we employ the production data 100x40 using a binary machine-part matrix, as introduced by Gonçalves [12], where 100 parts and 40 machines are involved. The original production data is also modified to change into an operation sequence-based machine-part matrix to evaluate the performance of the proposed method. The CPU time to solve these problems is less than three seconds. It indicates the merit of the proposed method to apply to big-size problems in a short computational time while still obtaining the optimal overall GTE. For all remaining evaluated instances, the computational time is significantly short in comparison with other methods in the literature due to the simple programming method [11, 13, 29, 37].

Figure 5 shows the CPU time comparison results between the proposed method and other algorithms for popular instances. The comparison results emphasize the significance of incorporating adjacent operations in the SC calculation, resulting in decreased AIM and NV in machine cells. The effectiveness of the sorting algorithm in determining machine positions during clustering leads to reduced ABM. Selecting the best routing for each part, based on both overall moves and machine cell compactness, allows for achieving optimal overall GTE in multi-routing problems within the context of CFP.

Table 5
The GTE comparison between other approaches and the proposed method

Test No.	Size PxMxR	Results from the literature			Proposed method			CPU (ms)
		Lee's GTE	Raja's GTE	Proposed MGTE	Lee's GTE	Raja's GTE	Proposed MGTE	
1	6x5	-	-	-	**0.5490**	**0.3431**	**0.4049**	3
3	5x4x10	0.9	0.45	0.585	**0.9**	**0.45**	**0.585**	1
2	9x9	0.4932	0.4685	0.4759	**0.5277**	**0.5277**	**0.5277**	8
4	12x10	0.5741	0.2870	0.3731	**0.6500**	**0.5572**	**0.5850**	12
5	7x10x14$^{(C)}$	-	-	-	**0.5428**	**0.5428**	**0.5428**	15
	7x10x14$^{(T)}$	-	-	-	**0.4857**	**0.4857**	**0.4857**	16
	7x10x14	0.3376	0.3376	0.3376	**0.4637**	**0.4637**	**0.4637**	16
6	18x10	0.6475	0.3551	0.4423	*0.6373*	**0.3906**	**0.4646**	18
7	19x12	0.4097	0.3200	0.3469	**0.4198**	**0.3646**	**0.3734**	31
8	12x12x20	0.3139	0.1200	0.1734	**0.3238**	**0.2476**	**0.2704**	20
9	20x8	0.6097	0.4146	0.4731	**0.6097**	**0.4146**	**0.4731**	10
10	8x9x20	0.6213	0.6213	0.6213	**0.6500**	**0.6500**	**0.6500**	17
11	10x10x24	0.8264	0.7851	0.7975	**0.8264**	**0.7851**	**0.7975**	22
12	16x10x32	0.7084	0.4482	0.5263	**0.7121**	**0.5554**	**0.6024**	35
13	20x20	0.5442	0.2655	0.3491	**0.5442**	**0.3982**	**0.4420**	128
14	35x18	0.3600	0.2501	0.2831	**0.4659**	**0.2912**	**0.3436**	122
15	40x25	0.4721	0.3122	0.3558	**0.4949**	**0.3328**	**0.3754**	222
16	43x16	0.3392	0.2430	0.2719	**0.4032**	**0.2667**	**0.3076**	149
17	45x20	0.4279	0.1908	0.2619	*0.4267*	**0.2999**	**0.3379**	148
18	51x20	0.4374	0.2951	0.3378	**0.4571**	**0.3131**	**0.3562**	321

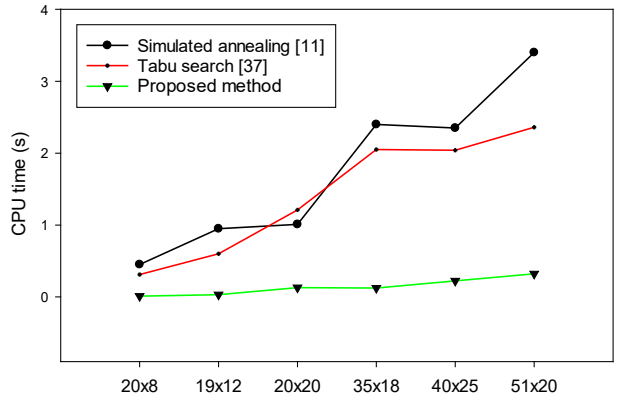


Figure 5
CPU time comparison for popular instances

Conclusions

This paper presents a novel similarity coefficient and an improved clustering algorithm to address machine cell formation and machine sequence generation simultaneously. The proposed method integrates realistic production data, such as operation sequence, production volume, processing time, machine capacity, multi-visits, and multi-routings. The modified group technology efficacy is utilized to evaluate the overall performance of the solutions for practical problems. Comparative analyses with existing methods using eighteen problems reveal the following conclusions:

- The proposed method outperforms other approaches in reducing weighted overall moves and voids in machine cells. It consistently achieves higher overall group technology efficacy for most test instances.

- The proposed method demonstrates significant time savings, with most test instances solved in less than 0.4 seconds, even big-size problem from the literature in under 3 seconds.

- The effectiveness of the proposed approach makes it a promising method for simultaneously addressing cell formation and machine sequence in complex problems within the realm of cellular manufacturing.

References

[1] Baysan, S., O. Kabadurmus, E. Cevikcan, S. I. Satoglu, and M. B. Durmusoglu, *A simulation-based methodology for the analysis of the effect of lean tools on energy efficiency: An application in power distribution industry*. Journal of Cleaner Production, 2019, **211** pp. 895-908, DOI: 10.1016/j.jclepro.2018.11.217

[2] Belhadi, A., F. E. Touriki, and S. El Fezazi, *Benefits of adopting lean production on green performance of SMEs: a case study*. Production Planning & Control, 2018, **29**(11): pp. 873-894, DOI: 10.1080/09537287.2018.1490971

[3] Pigler Cs., F.-V. Á., Abonyi J., *Scalable co-Clustering using Crossing Minimization - Application to Production Flow Analysis*. Acta Polytechnica Hungarica, 2016, **13**(2): pp. 209-228

[4] Gidwani, V. S. a. B. D., *Cellular manufacturing system practices in manufacturing industries: a pilot study*. International Journal of Indian Culture and Business Management, 2022, **25**(4): pp. 550-569, DOI: doi.org/10.1504/IJICBM.2022.122763

[5] Mahdavi, I., B. Javadi, K. Fallah-Alipour, and J. Slomp, *Designing a new mathematical model for cellular manufacturing system based on cell utilization*. Appl. Math. Comput., 2007, **190**(1): pp. 662-670 DOI: 10.1016/j.amc.2007.01.060

[6] Tóth, T., S. Radeleczki, L. Veres, and A. Körei, *A new mathematical approach to supporting group technology*. Eur. J. Ind. Eng., 2014, **8**(5): pp. 716-737 DOI: 10.1504/EJIE.2014.065734

[7] Mahdavi, I., B. Shirazi, and M. M. Paydar, *A flow matrix-based heuristic algorithm for cell formation and layout design in cellular manufacturing system*. Int J Adv Manuf Technol, 2008, **39**(9-10): pp. 943-953 DOI: 10.1007/s00170-007-1274-7

[8] Shashikumar, S., R. D. Raut, V. S. Narwane, B. B. Gardas, B. E. Narkhede, and A. Awasthi, *A novel approach to determine the cell formation using heuristics approach*. OPSEARCH, 2019, **56**(3): pp. 628-656 DOI: 10.1007/s12597-019-00381-4

[9] Mahdavi, I. and B. Mahadevan, *CLASS: An algorithm for cellular manufacturing system and layout design using sequence data*. Robot Comput Integr Manuf, 2008, **24**(3): pp. 488-497 DOI: 10.1016/j.rcim.2007.07.011

[10] Kumar, R. and S. P. Singh, *A similarity score-based two-phase heuristic approach to solve the dynamic cellular facility layout for manufacturing systems*. Eng. Optim., 2017, **49**(11): pp. 1848-1867 DOI: 10.1080/0305215X.2016.1274205

[11] Paydar, M. M., I. Mahdavi, I. Sharafuddin, and M. Solimanpur, *Applying simulated annealing for designing cellular manufacturing systems using MDmTSP*. Comput. Ind. Eng., 2010, **59**(4): pp. 929-936 DOI: 10.1016/j.cie.2010.09.003

[12] Gonçalves, J. F. and M. G. C. Resende, *An evolutionary algorithm for manufacturing cell formation*. Comput. Ind. Eng., 2004, **47**(2-3): pp. 247-273 DOI: 10.1016/j.cie.2004.07.003

[13] Boulif, M. and K. Atif, *A new branch-&-bound-enhanced genetic algorithm for the manufacturing cell formation problem*. Comput. Oper. Res., 2006, **33**(8): pp. 2219-2245 DOI: 10.1016/j.cor.2005.02.005

[14] Agrawal, A. K., P. Bhardwaj, and V. Srivastava, *Ant colony optimization for group technology applications*. Int J Adv Manuf Technol, 2011, **55**(5-8): pp. 783-795 DOI: 10.1007/s00170-010-3097-1

[15] Sudhakara Pandian, R. and S. S. Mahapatra, *Manufacturing cell formation with production data using neural networks*. Comput. Ind. Eng., 2009, **56**(4): pp. 1340-1347 DOI: 10.1016/j.cie.2008.08.003

[16] Park, S. and N. C. Suresh, *Performance of Fuzzy ART neural network and hierarchical clustering for part-machine grouping based on operation sequences*. Int. J. Prod. Res., 2003, **41**(14): pp. 3185-3216 DOI: 10.1080/0020754031000110277

[17] Selim, H. M., *Manufacturing cell formation problem: A graph partitioning approach*. Ind. Manag. Data Syst., 2002, **102**(5-6): pp. 341-352 DOI: 10.1108/02635570210432046

[18] Saral, J., S. Arumugam, I. Venkat, and A. Somasundaram, *Cellular manufacturing problem - A graph theoretic approach*. J. Adv. Mech. Des. Syst. Manuf., 2019, **13**(3): pp. JAMDSM0061-JAMDSM0061 DOI: 10.1299/jamdsm.2019jamdsm0061

[19] Danilovic, M. and O. Ilic, *A novel hybrid algorithm for manufacturing cell formation problem*. Expert Syst. Appl., 2019, **135** pp. 327-350 DOI: 10.1016/j.eswa.2019.06.019

[20] Seifoddini, H. and C. P. Hsu, *Comparative study of similarity coefficients and clustering algorithms in cellular manufacturing*. J. Manuf. Syst., 1994, **13**(2): pp. 119-127 DOI: 10.1016/0278-6125(94)90027-2

[21] Prabhakaran, G., T. N. Janakiraman, and M. Sachithanandam, *Manufacturing data-based combined dissimilarity coefficient for machine cell formation*. Int J Adv Manuf Technol, 2002, **19**(12): pp. 889-897 DOI: 10.1007/s001700200101

[22] Wu, L. and S. Suzuki, *Cell formation design with improved similarity coefficient method and decomposed mathematical model*. Int J Adv Manuf Technol, 2015, **79**(5-8): pp. 1335-1352 DOI: 10.1007/s00170-015-6931-7

[23] Goyal, K. K., P. K. Jain, and M. Jain, *A comprehensive approach to operation sequence similarity based part family formation in the reconfigurable manufacturing system*. Int. J. Prod. Res., 2013, **51**(6): pp. 1762-1776 DOI: 10.1080/00207543.2012.701771

[24] Wu, L., L. Li, L. Tan, B. Niu, R. Wang, and Y. Feng, *Improved similarity coefficient and clustering algorithm for cell formation in cellular manufacturing systems*. Eng. Optim., 2020, **52**(11): pp. 1923-1939 DOI: 10.1080/0305215X.2019.1692204

[25] McAuley, J., *Machine grouping for efficient production*. Production Engineer, 1972, **51**(2): pp. 53-53 DOI: 10.1049/tpe.1972.0006

[26] Gupta, T. and H. I. Seifoddini, *Production data based similarity coefficient for machine-component grouping decisions in the design of a cellular manufacturing system*. Int. J. Prod. Res., 1990, **28**(7): pp. 1247-1269 DOI: 10.1080/00207549008942791

[27] Nair, G. J. and T. T. Narendran, *CASE: A clustering algorithm for cell formation with sequence data*. Int. J. Prod. Res., 1998, **36**(1): pp. 157-180 DOI: 10.1080/002075498193985

[28] Seifoddini, H. and P. M. Wolfe, *Application of the similarity coefficient method in group technology*. IEE Trans (Ins. Ind. Engr.), 1986, **18**(3): pp. 271-277 DOI: 10.1080/07408178608974704

[29] Chung, S.-H., T.-H. Wu, and C.-C. Chang, *An efficient tabu search algorithm to the cell formation problem with alternative routings and machine reliability considerations*. Computers & Industrial Engineering, 2011, **60**(1): pp. DOI: 10.1016/j.cie.2010.08.016

[30] Alhourani, F., *Cellular manufacturing system design considering machines reliability and parts alternative process routings*. Int. J. Prod. Res., 2016, **54**(3): pp. 846-863 DOI: 10.1080/00207543.2015.1083626

[31] Alhourani, F., *Clustering algorithm for solving group technology problem with multiple process routings*. Comput. Ind. Eng., 2013, **66**(4): pp. 781-790 DOI: 10.1016/j.cie.2013.09.002

[32] Harhalakis, G., R. Nagi, and J. M. Proth, *An efficient heuristic in manufacturing cell formation for group technology applications*. Int. J. Prod. Res., 1990, **28**(1): pp. 185-198 DOI: 10.1080/00207549008942692

[33] Lee, K. and K. I. Ahn, *GT efficacy: A performance measure for cell formation with sequence data*. Int. J. Prod. Res., 2013, **51**(20): pp. 6070-6081 DOI: 10.1080/00207543.2013.794317

[34] Raja, S. and V. Anbumalar, *An effective methodology for cell formation and intra-cell machine layout design in cellular manufacturing system using parts visit data and operation sequence data*. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2016, **38**(3): pp. 869-882 DOI: 10.1007/s40430-014-0280-6

[35] Yin, Y., K. Yasuda, and L. Hu, *Formation of manufacturing cells based on material flows*. Int J Adv Manuf Technol, 2005, **27**(1-2): pp. 159-165 DOI: 10.1007/s00170-004-2143-2

[36] Kong, T., K. Seong, K. Song, and K. Lee, *Two-mode modularity clustering of parts and activities for cell formation problems*. Comput. Oper. Res., 2018, **100**pp. 77-88 DOI: 10.1016/j.cor.2018.06.018

[37] Chang, C. C., T. H. Wu, and C. W. Wu, *An efficient approach to determine cell formation, cell layout and intracellular machine sequence in cellular manufacturing systems*. Comput. Ind. Eng., 2013, **66**(2): pp. 438-450 DOI: 10.1016/j.cie.2013.07.009

[38] Ahi, A., M. B. Aryanezhad, B. Ashtiani, and A. Makui, *A novel approach to determine cell formation, intracellular machine layout and cell layout in the CMS problem based on TOPSIS method*. Comput. Oper. Res., 2009, **36**(5): pp. 1478-1496 DOI: 10.1016/j.cor.2008.02.012

Metaheuristic Algorithms for Related Parallel Machines Scheduling Problem with Availability and Periodical Unavailability Constraints

Mihály Gencsi

Department of Computational Optimization, University of Szeged, Árpád tér 2, 6720 Szeged, Hungary, gencsi@inf.u-szeged.hu

Abstract: The Related Parallel Machine Scheduling Problem (R-PMSP) is a type of optimal job scheduling problem. The problem is to assign different types of jobs to different parallel machines. Every machine has a speed rate that can execute a job faster or slower than other machines. This paper focuses on an R-PMSP, with availability and periodical unavailability constraints. Some jobs can also have machine preferences. The problem with these constraints is NP-hard. This study describes three metaheuristic algorithms for solving the problem. Namely, the algorithms are Genetic Algorithm (GA), Simulated Annealing (SA), and Discrete Grey Wolf Optimizer (DGWO). This article focuses on examining the performance of the algorithms, determined by the required time, to find a suboptimal threshold. Simulated Annealing proved to be the best in terms of efficiency and time required to find the suboptimal threshold. In addition, the study describes a benchmark generator method for this problem, which guarantees to create a problem with given properties and with a given optimum.

Keywords: Parallel Machines Scheduling; Availability and Periodical Unavailability Constraint; Genetic Algorithm; Simulated Annealing; Grey Wolf Optimizer

1 Introduction

Task scheduling problems can be seen in any part of our lives, for example, scheduling production lines, scheduling task execution, and assigning clients to service queues. That is why scheduling is one of the most dominant problems to be solved today. There are many approaches to solving these problems in the literature, but the diversity of the problem means many open areas. Time is one of the essential resources. Valuable time can be saved by properly allocating tasks to the machines. The motivation for this research comes from two problems that need to be solved. The first is the scheduling of patients to testing laboratories for individual tests. This can include CT, MRI, blood tests, cancer prevention tests, etc. The duration of tests varies and patients have the option of choosing laboratories. For instance, the patient has an agreement with the laboratory. The testing laboratories have different

characteristics: maximum availability (opening times) and periodical unavailability time (break times, maintenance time). Some laboratories perform their tasks faster than others (diversity of MRI machines, etc.). The second example is the scheduling of tasks to various computing devices. With so many different types available, such as PCs, supercomputers, and microcontrollers, the processing speed varies, and tasks must be assigned accordingly to optimize their running time. Utilizing a fast machine can reduce the running time of all tasks. Additionally, certain tasks may be delegated to predetermined machines for reasons such as data protection or contractual obligations. These machines possess varying characteristics, including reboot time, maintenance schedules, and cleaning intervals. The machine scheduling problem can be precisely matched with the patient scheduling problem. The characteristics of two examples are perfectly reconcilable. Patients are the tasks, and laboratories are the machines. In both cases, the goal is to minimize the makespan, i.e. minimizing the time difference between the start and the end of the task sequence. This problem is called the Related Parallel Machine Scheduling Problem (R-PMSP) with constraints.

In the next section, in chronological order, the state of the art on the problem in more detail is presented. Then, the types of scheduling problem and present current methods are described. In Section 3, a mathematical model of our problem is defined. Section 4 presents a benchmark generation method that yields the optimum. The used metaheuristic and their problem-specific modifications are described in Section 5. The computational results are discussed in Section 6. Finally, in the last section, Section 7, the results are concluded, summarized and the possibilities for further improvements are discussed.

2 Related Work

Researchers have recently defined categories and types of optimal scheduling problems. Two broad categories can be distinguished:

- **Single-stage job scheduling** Where each job consists of only one execution phase
- **Multi-stage job scheduling** Where each job consists of several execution phases that must be executed in parallel or a predefined order according to different rules

There are several types of single-stage job scheduling: single-machine scheduling, identical-machine scheduling, related-machine scheduling, and unrelated machine scheduling. Three types of multi-stage scheduling problems are known: open-shop scheduling, flow-shop scheduling, and the job-shop scheduling problem. These fundamental problems can be extended with different machine or job constraints. Machine constraints can be as follows: a machine can work for a particular time, each machine has its own time to process the information needed to complete the task, and machines can stop periodically. Task constraints can also be of many kinds: tasks must run within a given time, a task can be moved to another machine, tasks can only be available after a particular time, and tasks can be split up or not. There can be different objective functions to minimize [12]: makespan, maximum lateness, the total completion time, number of late jobs, or the total tardiness. The interested reader is referred to [5] [16] and the references therein.

2.1 The Single-Machine Scheduling Problem with Constraints

The simplest version of the problem is the single-machine scheduling problem, to which periodic or random breaks can be assigned. In [7], the single-machine scheduling problem with availability constraints is discussed. Two types of availability constraints are introduced. A machine must stop maintenance after a specific time, or a tool must be replaced after a particular processed job. In this case, the goal is to minimize the makespan. It has been shown that a single-machine problem with two maintenance constraints is NP-hard. Six types of heuristic algorithms are proposed to solve the problem, and it is shown that the best performing among them is the decreasing order with first fit algorithm (DFF). In [11], the single-machine problem is addressed under tasks due dates and machine unavailability constraints. The goal is to minimize the sum of maximum earliness and tardiness. A mathematical formulation was developed to exactly solve small problems. The Variable Neighborhood Search (VNS) was used to solve real-life problems. The VNS was extended with two knowledge module-based local searches, which improves the weaknesses of the random search of VNS. Experimental results have shown that the modified VNS can achieve optimal or near-optimal solutions in a reasonable time. In addition to these works, numerous other studies in the literature formulate the problem as a MILP model and solve the problem using various proven methods [2] [18].

2.2 The Two-Machine Scheduling Problem with Constraints

Many papers in the literature focus on the two-machine R-PMSP with various constraints. In [9], the problem was studied under the periodic availability constraint of a machine. Their goal is to minimize the makespan. They showed that the Longest Processing Time first algorithm (LPT) has a worst-case ratio of 3/2 if the problem is offline. In [8], the two-machine probability was graded under the constraint that one machine is unavailable at a given time. Their goal is to minimize

the total weighted completion time. They developed a fully polynomial-time approximation scheme (FPTAS) for this problem. They also generalized this scheme to m parallel machines. In [1], the two-machine scheduling problem with unavailability of a single machine for a specific time was addressed. Their goal was to minimize the makespan. They separately chose five cases for this problem and developed a separate solution method for each case. It was shown that these methods are efficient even for a large number of items.

2.3 Multiple Machine Scheduling Problem with Constraints

There are also papers in the literature related to multi-machine task scheduling. For multi-machine task scheduling, several types of constraints can be found in the literature, for which different algorithms have been developed. The work [3] deals with a pseudo-analysis of the classical scheduling problem, for which unavailability times for machines and release dates and delivery deadlines for tasks are introduced. A branching strategy and a new lower and upper bound for the tasks are developed based on a representation taking all the permutations of tasks. It is shown that embedding a semi-preemptive lower bound based on max-flow computations in a branch-and-bound algorithm yields very promising performance. Using this method, they were able to solve 700 tasks with 20 machines within a reasonable CPU time. The authors of [10] focus on the multi-machine task scheduling problem without extra constraints. Their goal is to minimize the maximum delay. To solve this problem, the Largest the sum of Processing time and Delivery Time first Simulated Annealing algorithm called LPDT-SA is developed. The initial solution was generated using a heuristic LPDT method. In addition to these, they used an effective solution for the representation that efficiently implements swapping and insertion into the neighborhood and avoids worse solutions. The resulting algorithm is able to solve problems with 350 tasks in 90 seconds, and the average error between the lower bounds for all 2400 random instances is 0.339%. In the study [13], the parallel machine scheduling problem is studied with multiple scheduled unavailability periods. In their presented case, they allow the tasks to be restarted. Their goal is to minimize the makespan. They first formulate the problem as a MILP for small, medium, and moderately large instances. They proposed an enumeration algorithm using lexicographic sequencing. They compared this method with the MILP model. It is shown that the proposed algorithm obtains the optimal solution and is faster. In [4], the scheduling problem of static m identical parallel machines with shared server and sequence-dependent setup times is addressed. Their goal is to minimize the makespan. They describe a MILP model for the problem and, in addition, implement a Simulated Annealing and Genetic Algorithm for large-scale problems. After comparing the efficiency of the three methods, they concluded that the GA algorithm provides better quality solutions in a reasonable computation time.

3 Model Description

In this section, we describe our problem. The problem studied is single-stage offline job scheduling. In other words, we want to assign a predefined set of jobs to machines (offline). All tasks are unrelated; this means that they do not form jobs, and we cannot break the jobs into smaller tasks or pause them (single-stage). Thus, we will use task and job as synonyms. Data are fixed and deterministic. We assign machine-specific constraints to the problem: maximum availability constraint and periodical unavailability constraint. Machines can have different speeds, which means that a machine can execute all tasks faster (or slower). Thus, the lengths of tasks are given in number of machine instructions (NMI), where one machine instruction is done in one-time unit on a unit-speed machine. We can consider one unit of time as one minute for an easier discussion. Suppose that we have an examination with ten machine instructions, a laboratory with 0.9 speed, and a faster laboratory with 0.5 speed. For example, the second laboratory has faster CT equipment and better-qualified staff than the first one. The first laboratory can complete this task in nine minutes, while the second laboratory in five minutes. The machine availability constraint means that a machine is unavailable after a particular time. For example, if we have a laboratory with a four-hour daily work schedule, we cannot assign more than four hours of examination. The periodic unavailability constraint consists of two components: maintenance time and uptime. After the end of the uptime, the machines must be periodically suspended for maintenance time. Suppose that we have a laboratory whose internal rules require that workers need to take a ten-minute break every 120 minutes. In this case, the uptime is 120 minutes, and the maintenance time is 10 minutes. The machine parameters are different for each machine. If the same constraint were imposed on all machines, we would obtain a particular case of the problem. Similarly, as we can introduce features for a machine, we can also introduce features for tasks. The machine preference for a task is set to a predefined machine and specifies that the task must be executed on that machine. It is important to point out that splitting a task into smaller sub-tasks in our problem is impossible. The tasks have no setup time, no appearance date, and no execution date. All tasks are known and available at the starting time. Our goal is to minimize the makespan.

Table 1
Notations

$J = \{j_i\}$	Jobs, $i = 1, \ldots, n$
$P = \{p_i\}$	Number of machine-instructions of each job
$JP = \{jp_i\}$	Machine preference of each job
$M = \{m_j\}$	Machines, $j = 1, \ldots, m$
$S = \{s_j\}$	Speed of each machine
$UT = \{ut_j\}$	Periodic uptime of each machine
$MT = \{mt_j\}$	Maintenance time of each machine
$A = \{a_j\}$	Available time of each machine

To formalize the problem, we need to introduce notations, summarized in Table 1. J represents the index array of the tasks. Each task is given by its NMI, p_i, from the array P. In addition, each task is assigned a jp_i value, which indicates the machine on which a task needs to be executed. If we do not have a machine preference, this value is null. M represents the index array of the machines. Each machine is assigned its speed s_j from the interval (0, 1.0], where a speed 0.5 means a two-times faster machine than speed 1 in relation to processing time. For each machine, ut_j and mt_j specifies the intervals at which the machine is active or paused. In addition, the maximum availability time of each machine a_j is also given. From the last parameters, one can compute the maximum number of segments for each machine, namely, $ns_j = \left\lfloor \frac{a\_j+mt\_j}{ut_j+ mt_j} \right\rfloor$.

Based on the above, we can formulate the following decision variables.

$$x_{kj}^{i} = \begin{cases} 1, & \text{if the } i^{th} \text{ job is run on the } k^{th} \text{ segment of the } j^{th} \text{ machine;} \\ 0, & \text{otherwise.} \end{cases}$$

With these parameters and variables, one can formulate the mathematical model of the problem. Then, such a model can be used to solve the problem by any method for ILP. Given the complexity of our problem, aiming for a concrete mathematical model is not worthy because ILP solvers cannot solve reasonable problems efficiently. Therefore, the goal was to describe the problem at hand, so that the reader could use and recreate the problem as presented.

4 Benchmark Generating Method

There is no benchmark for this problem in the literature to provide an optimal solution. In the papers mentioned in Section 2, randomly generated test instances were used. The benchmarks in the literature are not designed for such a problem: they do not contain breaks, and the constraints defined are not included. Adapting these existing benchmarks for the task is very time-consuming, and one would be unable to provide the optimum. Therefore, a benchmark generating method is presented, to ensure the optimal solution or to define a value close to the optimal one, generate gaps while keeping the optimal solution, and handle all constraints. Several parameters of our generating method are controlled, which allows for the specification of the generated problem. For example, one can set limits on the machine instruction, the maximum availability constraint, the periodical uptime, the speed of the machines, and the probability of machine preference. In addition, the number of machines and the number of gaps for each machine can be specified.

4.1 Description of the Generating Method

In the first step, the aim is to create a problem for which we know the optimal solution. It can be achieved by setting the total time of each machine to the given optimum. That is, take random uniform integer values for each machine, from a given interval, which will give the total NMI it can process (these have to be larger than the optimum and do not take maintenance times into account at this point). Set the speed of each machine, such that the execution time of their generated NMIs takes exactly the optimal time. After that, assign maintenance times and uptimes to each machine between the limits, such that the last segment is always shorter than the others, but strictly larger than 0. This will ensure that the optimum will be the given value. When generating tasks, make sure that the limit of the number of instructions is respected and that there are no gaps (idle times) on the machines. Each part of a segment is linked to a job. This means that tasks on the machine must follow each other, and only maintenance time can be added between them. Assign the maximum availability constraints to the machines randomly between the maximum availability limits. Here, we make sure that the generated value is not less than the optimum.

By careful generation, we know that all limits have been met and the optimal solution is known, since there are no idle-time slices except for pause times.

4.2 The Rules for Generating the Gaps for m Related Machines

A gap in a solution is the idle time when the machine is active but not working. In the previous step, we generated test cases with no gaps in the optimal solution. However, in most practical cases, this does not occur. Therefore, our goal is to artificially introduce gaps into a test case such that its optimal solution, Opt, does not change.

One can observe that for m related-machines, one can reduce the NMI of $m-1$ tasks by one, without affecting the optimal solution.

To see the above statement, suppose that we have $m > 1$ machines and $n \geq m$ tasks, where the NMI of all tasks are integer, that is, the smallest NMI is one. Let us randomly reduce the NMI of $m-1$ tasks by one. Consider the case where we reduce the NMI of the last tasks on the last segment of the first $m-1$ machines. Still, the optimum cannot change because on the last machine there were no reductions, and the last task finishes at the same time as before. As there are no tasks with smaller units, there is no better distribution of the tasks.

We also observe that when we reduce more than $m-1$ machine instructions in total, the optimum might decrease. Still, we can obtain a lower bound on the optimum by calculating the maximum possible decrease of the optimum. Let $r \geq m$ be smaller than the smallest NMI of the last segments times m. If we reduce the

total NMI of all tasks by r, then the optimum can improve at most by $\frac{r}{\sum_{i=1,..,m} s_i}$ unit time.

To understand this formula, suppose that we have $m > 1$ machines and $n \geq m$ tasks. We randomly reduce the NMI of some tasks, in total, by r. To analyze the worst case, we reduce the last tasks of the j[th] machine by $\frac{r}{\sum_{i=1,..,m} s_i} \cdot s_j$. Thus, the length of each machine is reduced by the same amount, $\frac{r}{\sum_{i=1,..,m} s_i}$. It modifies the optimal solution to $Opt - \frac{r}{\sum_{i=1,..,m} s_i}$.

In the deletion mechanism, we can choose whether maintaining the optimum is necessary or whether a good lower and upper bound on the optimum is sufficient. We aim to reduce the NMI of any task in the non-last segment such that it cannot be changed by any task in the last segment. Note that by generation, the last segment is always shorter than the other segments on all machines. As a consequence, the gaps cannot be moved to the last segment. This also implies that there is a maximum reduction which can be made.

Namely, we can reduce the time of tasks from a non-last segment until the combination of all possible tasks does not fill the segment better. Generating all possible solutions (combinations) is too expensive because the segments can be placed in any order within a machine having the same optimal value. Therefore, the algorithm controls this with a parameter to display all solutions or just the first possible one. The second step results in a test case that adheres to the task specification and contains gaps. If only $m - 1$ NMI is reduced in total, optimality is guaranteed, while if reduction is higher, we can bound the optimum.

The result of the generating method is shown in Figure 1, visualized in Figure 2. The figure shows a test case with three machines. It is observed that the optimal solution is 15 units, the number of tasks is 13, and there are gaps (marked with a red striped box) in the solution. The generated case obeys all constraints of the control parameters.

5 Algorithms

The R-PMSP problem with the introduced constraints is NP-hard, so we cannot give an exact algorithm that can solve our problem for real-life test cases in a reasonable time. There are exact algorithms for problems with two machines or without constraints, but if we increase the number of machines, tasks, and/or add constraints, we can only solve the problem using heuristics. Simulated Annealing (SA) and Genetic Algorithm (GA) are commonly used methods to solve these problems [4] [10]. We have used these algorithms and introduced our modifications to obtain results close to the optimal solution. In addition, we applied a discrete version of the Grey Wolf Optimizer (DGWO) to solve the problem.

5.1 Common Representation and Operators of the Algorithms

The three algorithms use the same representation for the solutions. In addition, we use the same fitness function, crossover, mutation, and initial solution generation procedure.

5.1.1 Representation of Individuals or Solutions

We can achieve faster convergence with a well-chosen representation of an individual or a solution. A good representation allows us to better explore our search space. A one-dimensional array represents individuals or solutions. The i^{th} element of the array correspond to the task j_i, and encode which segment of which machine it is assigned (see the Solution line in Figure 1). To do this, we first enumerate each segment of each machine (within the available time of the machine) sequentially from the first machine to the last. We store which segments belong to which machines in an index array. In addition, to speed up the calculations, we store the gaps (idle times) of each solution in an array, that is, how much empty space is available on each machine segment. In this way, during crossover and mutation, we do not need to calculate which segment has enough space for a task. The required storage space increases by storing the gaps in the segments but reduces the computational time considerably.

Two solutions are equivalent, where only the permutation of non-last segments of any machine is different. These individuals differ because the segments are in a different order, but the tasks assigned to a given segment are mutually equivalent. The above representation considers these solutions different. Thus, we order the segments, except the last, within the machine in descending order based on their idle time. With this sorting, the loaded segments are moved to the front (see the segments gaps in Figure 1). In the case where two segments have the same idle time, the choice is based on the number of tasks in the segments.

For example, we can see the test case from Figure 1 on the machines in Figure 2. We have three machines, 13 tasks (jobs), and the optimum is 15. In the example, we see one gap (marked with a red striped box) on the first machine and one on the first segment of the second machine. The characteristics of the machines are shown in Table 2. We see the solution representation of the same in Figure 1. The first array contains the NMI of the tasks, the second describes the segment index of the machines, the next encodes the solution, and the last array presents the gaps on the segments. We can see in the solution array the segment indexes. For example, the first task, named Job 1 in Figure 2, will be performed on the segment with index one, which is the segment of the first machine. The third task, called Job 3 in Figure 2, will be executed on the segment with index five, which is the first segment of the third machine.

Table 2

Machine characteristics

Constraints	m_1	m_2	m_3
Maximum availability	22	22	23
Speed	0.75	1.0	0.5
Uptime	30	7	16
Maintenance time	0	2	4

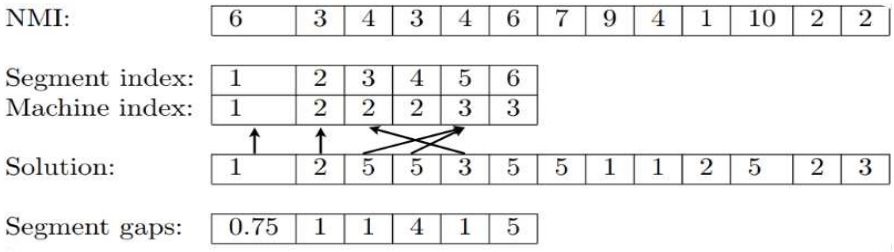


Figure 1

Solution representation

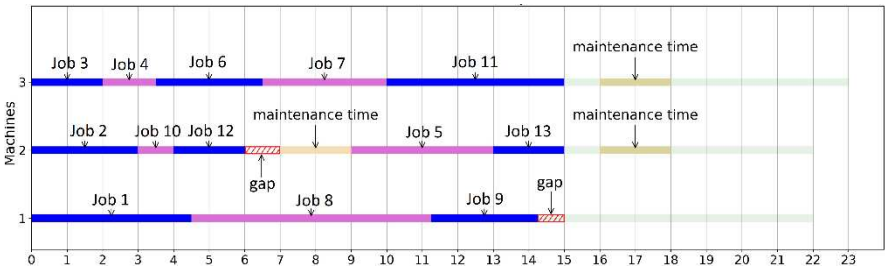


Figure 2

Generated test case with 3 machines, 13 jobs with the optimum 15

5.1.2 Generation of Initial Solutions or Individuals

The initial population or solution is generated at random. We first assign tasks with machine preference to their corresponding machines, but the segment within a machine is chosen randomly. Second, the remaining tasks are assigned to randomly selected machine segments from those where they fit in. Finally, we regenerate the individual or solution if the task cannot be assigned to a machine even after multiple attempts.

5.1.3 Fitness Function

The fitness function measures how close a given solution is to the optimal solution of the problem. It is a numeric value assigned to an individual or a solution. We could use only the makespan as a fitness function, but instead we use a more complex fitness function: we add the number of occupied segments and the total

idle time on the occupied segments to the makespan. We can formalize the fitness function as $Fitness(x) = occupied\_seg(x) + total\_idle(x) + makespan(x)$.

5.1.4 Crossover

A crossover is a genetic operator in which genetic information from two selected individuals (parents) is combined to produce new individuals (children). We have implemented several types of crossover operators that can be used with the one-dimensional array representation. Different crossovers per algorithm were shown to be better. We have implemented k-point, uniform, and three-parents crossover [14]. We swapped randomly selected segments of the individuals. For each substitution, we checked the feasibility of the solution, and dropped the children which were not feasible (the new segment does not fit into its new place). We observed that the crossovers used in this way were almost useless, and the methods converge to a wrong solution quite quickly.

Consequently, we created modified crossovers that perform segment replacement by randomly selecting the segment from those that fit on the replacement machine. In cases where it was not possible to swap due to machine preference or none of the segments fit, we left the element unchanged. In this way, the crossover operators generate more correct solutions and better traverse the search space.

5.1.5 Mutation

A mutation is a genetic operator responsible for diversifying a population. It is a slight modification of an individual or a solution. The discrete version of the Grey Wolf Optimizer requires several mutations to maintain population, namely, pack of wolves, diversification. We have implemented bit flip, swap, inverse, and reverse mutation [15]. In these implementations, segments are modified to other segments. We consider a mutation to have been used if the resulting individual satisfies the constraints of the problems. Otherwise, a new individual is generated from scratch. In addition, we implemented a modified version of each mutation named above. In these implementations, we swapped the machines, but we chose the segments randomly.

5.1.6 Selection

Selection is a rule that determines the individuals in the following population or the individuals participating in the crossover and mutation. For crossover and mutation, we randomly select individuals from the current population. We use the tournament selection rule to create a new population for our problem [6].

5.2 The Algorithms and their Adaptations

We modify the SA, the GA, and the DGWO with changes that achieve faster convergence and improve the quality of the solutions.

5.1.7 Simulated Annealing

The Simulated Annealing algorithm originates from metallurgy, which aims to change the physical properties of a material by heating and controlled cooling. The method employs an iterative motion according to the varying temperature parameters based on the annealing operation of metals.

Algorithm 1 describes the Simulated Annealing algorithm we use, where p_m denotes the mutation, p_{co} denotes the cooling scheme parameters, and $data$ encodes the problem to be solved.

The algorithm requires an initial temperature, a reduction scheme, a number of iterations, and an initial solution, which will be called the current solution. There are several temperature reductions schemes: linear, logarithmic, exponential, and quadratic. The linear scheme is the best for the problem. The temperature is reduced from the initial temperature according to the reduction mechanism. In each iteration, we generate a neighboring solution using the current solution. The current solution is compared with the neighboring solution. It is swapped if its fitness value is better than the current fitness value. Otherwise, it is swapped with a certain probability. The role of the acceptance probability is to be able to move out of the local minimum points and move towards better solutions. As the temperature decreases, the value of this acceptance probability decreases. Several stopping conditions can be introduced: after a given number of steps r, if the best solution has not improved, or the maximum number of iterations has been reached. Due to the heuristic nature of the algorithm, the optimal solution can be reached after multiple runs.

In our version, we use the mutation operators to generate a neighboring solution. In contrast to the basic algorithm, we do not generate an adjacent solution per iteration. Instead, we generate n and compare the best one with the current solution.

Algorithm 1 SimulatedAnnealing($data$, T, $iter_{max}$, p_m, p_{co}, $neighboor_{size}$)

$x_{cur} \leftarrow$ GeneneratePopulation($data$, 1);
$i \leftarrow 0$; $x_{best} \leftarrow x_{cur}$; $run \leftarrow true$;
while ($i < iter_{max}$ && run) **do**
  $temp_{cur} \leftarrow$ CalcTemp(i, T, p_{co});
  $X \leftarrow$ GenerateNeighboors(x_{cur} , p_m, $neighboor_{size}$);
  $x_{tmp} \leftarrow$ SelectBestNeighboor(X);
  if Fitness(x_{tmp}) ≤ Fitness(x_{cur}) **then**
    $x_{cur} \leftarrow x_{tmp}$;
    if Fitness(x_{tmp}) ≤ Fitness(x_{best}) **then**
      $x_{best} \leftarrow x_{tmp}$;
    end if
  else
    if Random(0, 1) < exp((Fitness(x_{cur})−Fitness(x_{tmp}))/ $temp_{cur}$) **then**
      $x_{cur} \leftarrow x_{tmp}$;
    end if

16: **end if**
17: $run \leftarrow$ ExamineStopCriterion();
18: **end while**
19: **return** x_{best}, Fitness(x_{best})

Algorithm 1
Simulated Annealing

5.1.8 Genetic Algorithm

The Genetic Algorithm is a population-based metaheuristic algorithm inspired by natural selection. Instead of one solution, we work with a set of solutions, namely, population. The elements of the population are called individuals, and the number of iterations is called generations.

Algorithm 2 presents the pseudocode of the Genetic Algorithm, where p_c is the crossover, p_m is the mutation, and p_s is the selection parameter.

Algorithm 2 GeneticAlgorithm($data$, pop_{size}, gen_{size}, p_c, p_m, p_s)

1: $pop_{cur} \leftarrow$ GeneratePopulation($data$, pop_{size});
2: $x_{best} \leftarrow$ SelectBestIndividual(pop_{cur});
3: **for** $i = 1, \ldots, gen_{size}$ **do**
4: $crossover \leftarrow \{\}$; $mutation \leftarrow \{\}$;
5: **for** $j = 1, \ldots, \frac{pop_{size}}{2}$ **do**
6: $r_1, r_2, r_3 \leftarrow$ RandInt(0, pop_{size});
7: $crossover \leftarrow crossover \cup$ Crossover(p_c, pop_{cur}, r_1, r_2);
8: $mutation \leftarrow mutation \cup$ Mutation(p_m, pop_{cur}, r_3);
9: **end for**
10: $pop_{cur} \leftarrow pop_{cur} \cup mutation \cup crossover$;
11: $x_{cur} \leftarrow$ SelectBestIndividual(pop_{cur});
12: **if** Fitness(x_{best}) > Fitness(x_{cur}) **then**
13: $x_{best} \leftarrow x_{cur}$;
14: **end if**
15: $pop_{cur} \leftarrow$ Selection(p_s, pop_{cur});
16: **end for**
17: **return** x_{best}, Fitness(x_{best})

Algorithm 2
Genetic Algorithm

The algorithm requires a population size, a generation size, and an initial population. Individuals from the current population are selected according to some selection rule. We use our modified crossover and mutation operators. Then, the population of the new generation is selected from the current population, including mutation and crossover results. This selection can be based on age or fitness. The algorithm can be stopped when a generation number is reached or if no improvement is observed after r steps.

In our implementation, we work with a fixed population. The initial population is generated randomly according to the method described in Subsection 5.1.2. In each iteration or generation, we generate new individuals equal to half the population size by crossover and mutation operators. The individuals involved here are selected randomly. We use tournament selection for the selection of new populations.

5.1.9 Discrete Version of the Grey Wolf Optimizer

The Gray Wolf Optimizer is a population-based metaheuristic algorithm inspired by nature. The method attempts to mimic the hierarchy and hunting mechanism of grey wolves. Grey wolves are organized into a hierarchy: alpha, beta, delta, and omega wolves. Each level of the hierarchy has its role, and no one can be absent from the hierarchy. The alpha wolf is at the top of the hierarchy. His role is to make decisions about the pack and to lead the pack. The beta wolf is the second level of the hierarchy. His role is to assist the alpha wolf in decision-making, to relay the decisions of the alpha wolf to the pack, but he also has a leadership role. The lowest level of the hierarchy is the omega wolf, which represents the weakest wolf in the pack. He does not have an important role in the pack, but his absence can create internal conflict. All wolves that do not belong to the previous three are delta wolves. Their role is to scout, protect, and hunt. The hunting mechanism of grey wolves also plays an important role in the algorithm. The algorithm mimics hunting, prey search, prey enclosure, and prey attack in its methods.

Algorithm 3 presents the pseudocode for the used Grey Wolf Optimizer, where p_c is the crossover, p_m is the mutation, pb_{max} and pb_{min} are the local and global stage measures, respectively.

Algorithm 3 DiscreteGWO($data$, $pack_{size}$, gen_{size}, p_c, p_m, pb_{max}, pb_{min})

```
pack ← GeneratePopulation(data, pack_size);
x_α, x_β , x_δ ← SelectBestWolfs(pack);
x_best ← x_α;
for i = 1, . . . , gen_size do
  x_α, x_β , x_δ ← LocalSearchUpdate(x_α, x_β, x_δ, p_m);
  for j = 1, . . . , pack_size do
    if Random(0, 1) ≤ pb_max − (pb_max − pb_min) · i/gen_size then
      pack_i ← SeekingMode(pack_i, p_c);
    else
      pack_i ← TracingMode(pack_i, p_c);
    end if
  end for
  x_α, x_β, x_δ ← SelectBestWolfs(pack);
  if Fitness(x_best) > Fitness(x_α) then
    x_best ← x_α;
  end if
end for
```

18: **return** x_{best}, Fitness(x_{best})

Algorithm 3
Discrete version of the Grey Wolf Optimizer

The algorithm requires a population or pack size and a number of iterations. In the algorithm, the tails α, β and δ indicate the first three solutions with the lowest fitness values. The ω denotes all other solutions. The first step of the algorithm is to generate an initial pack, from which we select the first three wolves. Then, in each iteration, a local search method is applied to the wolves α, β and δ and their values are updated. Then, for each wolf in the pack, we use the seeking mode or the tracing mode in addition to the selection probability. The stopping conditions can be the maximum number of iterations or no change in our best solution after r steps.

In detail, the local search and update algorithm (line 5 in Algorithm 3) calls the local search for the wolves α, β and δ and sorts these three wolves by fitness value. In the local search, we improve the fitness value using three types of mutation. We use two search modes in the algorithm, corresponding to the local and global search stages. We apply the search selection mentioned in line 7 in Algorithm 3, which explores the global space in the first stages of the search, clustering around local optima. In later stages, it converges to the global optimum. In the seeking mode, we aim to preserve population diversity and avoid premature convergence. This can be achieved by using the crossover between a randomly selected individual and the current individual. The tracking mode is used for local searches. The method selects the α, β and δ with fitness-based selection, and execute a crossover on the selected with the current solution [17].

6 Experimental Results

In this section, we present our computational results. First, we show the test cases, which were generated using the method described in Section 4. Then, we discuss how we determined the parameters of the algorithms. Finally, we also discuss the methods for comparing the algorithms.

6.1 Generating Test Cases

Test cases were generated using the method described in Section 4. We wanted to see the results compared to a known optimal solution for each test case, so we generated only as many gaps that did not change the optimal solution. It is impossible to directly control the number of segments and the number of tasks in the generating method. Therefore, some parameters of the generation method are controlled and others are left with a higher degree of freedom. We strictly set the parameter controlling the machine preference to 0.1 (so 10% of the tasks have preference), the number of machines to $m = 5, 10, 20$, and the machine speed to

the interval [0.8, 1.0]. We set maximum availability at [30, 120], periodical uptime at [10, 40] and the maintenance time at [0, 10]. The number of segments per machines are either 0, or chosen from [1,10] by the generator, denoted by $\overline{ns} = 0, 10$. We set the number of tasks to be chosen from three intervals: $n \in$ [25-50], [51-100] and [101-200]. For the easier, we will denote these by $\bar{n} = 50, 100, 200$. For each task, NMI is set to [1,18]. With these parameters, we generated 25 test cases. From each set of test cases, we randomly selected 5 test cases.

In total, we have the following classes for both $\overline{ns} = 0, 10$ (See in Table 3).

Table 3
Benchmark generating test case classes

m	5			10			20		
$\bar{n}$	50	100	200	50	100	200	50	100	200

6.2 Parameter Tuning of the Algorithms

We fine-tune all parameters of the three metaheuristic algorithms by controlled grid search. The test cases for fine-tuning are randomly generated. We considered each combination mentioned in Section 6.1. Each test case was run ten times. We try to define the parameters and operators based on improving the mean, minimum, and variance per iteration. We focus on keeping the variance large initially and decreasing slightly per iteration so that our search space is well-traversed. We also ensured that the average converges to the minimum by the end of the iterations. This method allowed us to filter out most of the parameters. To determine the additional parameters, we considered how many times out of ten runs reached 98% of the optimal value. We consider the winning parameter to be the one that reaches the limit more times and is closer to the optimum on average.

6.3 Experiences and Results

All runs were performed on a 2.00 GHz Intel Xeon Processor (E5-2660 v4 35M cache) with 64 GB RAM. We examine the average performance of the algorithms from 12 runs on the test cases presented in the Subsection above. Additionally, we compare the time required for the algorithms to reach a suboptimal threshold.

6.1.1 The Performance of the Algorithms

We study the performance of the algorithms on the generated test cases. First, we divide the test cases into two groups, one that does not contain periodical availability constraints, and the other that contains periodical availability constraints. Then, we run the algorithms on each test case 12 times and examine the optimum. If the algorithm does not stop after 3600 seconds, it is stopped, and the best solution is assigned to the run. We normalize the result of each run by dividing the absolute error of each result of the run with the optimal value. This value

represents the relative deviation error from the optimum. Next, we calculate the maximum, minimum, and mean of this relative for each test class. These values are grouped in increasing order by algorithm, number of machines, and tasks. For example, the group $m = 5$, $n = [51, 100]$, SA means $m = 5$, $n = [51, 100]$ using the SA. Finally, we calculate the mean, the mean minimum, and maximum error percentage of the group. In the figures, the colored column gives the relative average error, while the thin black line shows the range by the minimum and maximum of the relative error, all in percentages.

Figure 3 shows the results for the test cases without periodical availability constraints. The three algorithms obtain the optimum with a small error for the test cases with $m = 5$. As the number of machines increases, the percentage error increases linearly. The algorithms obtained the minimum out of 12 runs for all test cases. For the 20 machine runs, the maximum average error increased, but the average error did not exceed 3%. As the number of tasks increased, the performance of the algorithms increased. There may be several possible solutions for many tasks close to the optimum. In conclusion, as the number of tasks increases, the average error decreases. As the number of machines increases, the error increases.

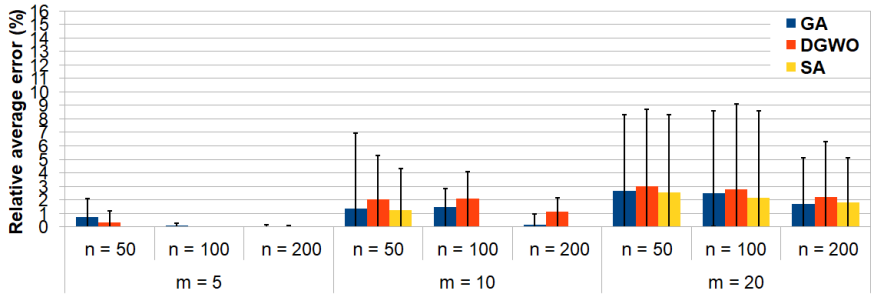


Figure 3

Test cases without periodical unavailability constraints. Relative average error from the optimum, with the minimum and maximum errors grouped by class

We can see in Figure 4 the relative average error grouped by machine, task, and algorithm for test cases with periodical availability constraints. The relative average error increases linearly as a function of the number of machines. The algorithms obtained the minimum out of 12 runs for all test cases. The relative average error of the algorithms does not exceed 5%, even for problems with 20 machines. In a few cases, it is observed that the maximum error decreases when increasing the number of tasks. The best performing algorithm here is also the SA. In this case, the maximum average does not exceed 15.5%.

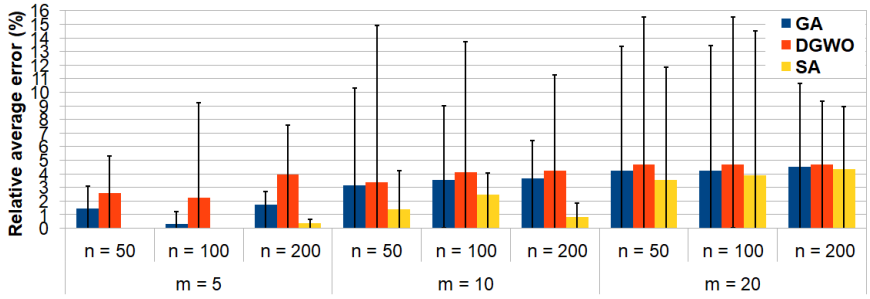
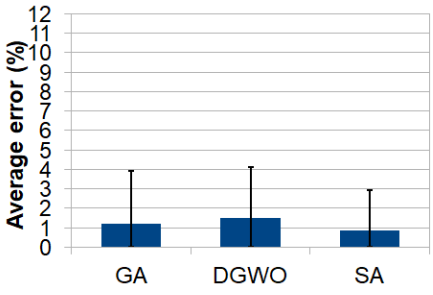


Figure 4

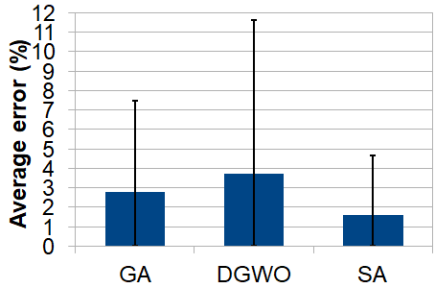
Test cases with periodical unavailability constraints. Relative average error from the optimum, with the minimum and error grouped by class

We can see in Figure 5, the overall average error calculated for each algorithm taking into account all test cases without periodical availability constraints. In general, SA has the lowest average failure rate of all the test cases. The relative average error of the algorithms does not exceed 1.5%.

Figure 6 shows the overall average error calculated for each algorithm for the all test cases without periodical availability constraints. The average error percentage for all algorithms for all test cases is below 5%, and the average maximum error percentage is below 12%. Furthermore, we can see that the average minimum error percentage is equal to 0%, which means that all algorithms find the optimal solution at least once in all test cases.



Test cases without periodical unavailability constraints. Average error from the optimum, with the minimum and maximum errors grouped by algorithm.



Test cases with periodical unavailability constraints. Average error from the optimum, with the minimum and maximum errors grouped by algorithm.

6.1.2 The Time Required to Find a Suboptimal Threshold

We study the average time required to reach some suboptimal threshold. We introduce four types of the suboptimal threshold: 30%, 20%, 10% and 5% the optimal solution plus the optimal solution. We divide the test cases into two groups,

one that does not contain periodical availability constraints and the other that contains periodical availability constraints. In the first case, we set the maximum runtime to 600 seconds because it does not require more time to find the suboptimal threshold and in the second case, it to 3600 seconds. We run the algorithms on each test case 12 times and examine the required time. If the algorithm did not reach the suboptimal threshold, we set the time to maximum runtime. We calculated the average of 12 runs. Then, we calculate the average of the average using some grouping rule. For example, we can group by the number of machines, the range of tasks, and the algorithms. In the figures, the points represent the average time required to reach the suboptimal threshold.

Figure 7 shows the average time required to reach some suboptimal threshold for test cases with periodical availability constraints. The test cases are grouped by the number of machines, tasks, and algorithms. If we increase the thresholds, the time required increases linearly. We observe that DGWO reaches the first two thresholds faster than GA. After that, DGWO slows down. SA is the fastest in all three categories.

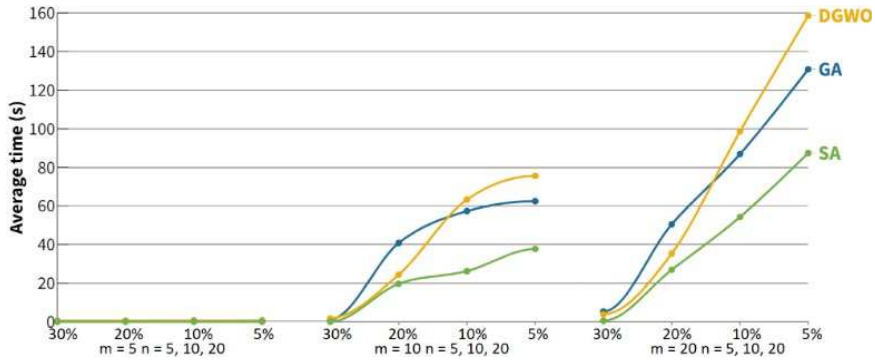


Figure 7

Test cases without periodical unavailability constraints. Relative average error from the optimum, with the minimum and maximum errors grouped by class.

Figure 8 shows the average time to reach some suboptimal threshold for test cases with periodic availability constraints. The test cases are grouped as above. If we increase the number of machines, the average running times increase exponentially.

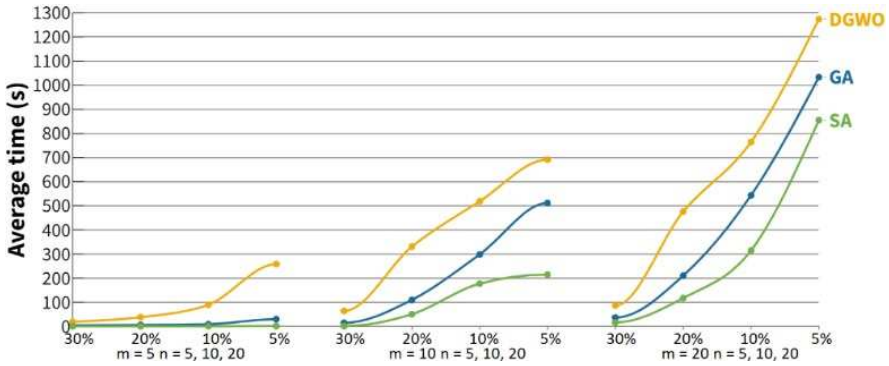


Figure 8

Test cases without periodical unavailability constraints. Relative average error from the optimum, with the minimum and maximum errors grouped by class.

Summary and Conclusions

This work focused on the related parallel machine problem (R-PMSP) with availability and periodical unavailability constraints. A robust benchmark generating method was proposed for the problem, which can generate a large variety of test cases, by controlling the parameters of the methods and knowing the optimum with certainty. In addition, a deletion mechanism that generates gaps in the solution, was offered. Gaps can be created, keeping the optimum or giving a lower and upper bound on the optimum. An implementation of three metaheuristic algorithms to solve the problem are given: Genetic Algorithm, Simulated Annealing and the Discrete version of Grey Wolf Optimizer. We introduced an efficient representation of the solution and methods to improve the algorithms in terms of solving the problem. Simulated annealing was also complemented with multiple neighborhood methods. Common operators and functions were used in the algorithms.

The performance of the algorithms was examined and compared to the average time required to find a suboptimal threshold. In the case without periodical unavailability constraints, the average relative error and the average of maximum relative errors of the algorithms are below 1.5% and 4.5%, respectively. Furthermore, with unavailability constraints, the average errors of the algorithms were below 4% and 11.5%, respectively. The running times of the algorithms increase exponentially by decreasing the threshold for many machines. However, for relatively few machines, this indicator is linear. The Simulated Annealing algorithm performs the best on average. The DGWO algorithm is not efficient for this problem, as it tends to get stuck in local optima and is time-consuming. GA takes longer on average than SA. However, it generates several near-optimal solutions, which can be beneficial in various cases where multiple solutions are needed, or we want to examine critical points. Therefore, we recommend utilizing this approach when sufficient computing time is available and several near-optimal or optimal solutions are required.

Future goals are to improve the SA and GA methods, with local search and procedures to reduce the searching space. The goal would be to generalize the SA method for semi-online R-PMSP, with constraints using local methods, but also to develop a Matheuristic, that handles the semi-online case.

References

[1] A. A. Masmoudi, M. Benbrahim. New heuristics to minimize makespan for two identical parallel machines with one constraint of unavailability on each machine. In 2015 International Conference on Industrial Engineering and Systems Management (IESM), pp. 476-480, 2015

[2] A. B. Keha, K. Khowala, J. W. Fowler. Mixed integer programming formulations for single machine scheduling problems. Computers & Industrial Engineering, 56(1):357-367, 2009

[3] A. Gharbi, M. Haouari. Optimal parallel machines scheduling with availability constraints. Discrete Applied Mathematics, 148(1):63-87, 2005

[4] A. Hamzadayi, G. Yildiz. Modeling and solving static m identical parallel machines scheduling problem with a common server and sequence dependent setup times. Computers & Industrial Engineering, 106:287-298, 2017

[5] A. Jain, S. Meeran. Deterministic job-shop scheduling: Past, present and future. European Journal of Operational Research, 113(2):390-434, 1999

[6] A. Shukla, H. M. Pandey, D. Mehrotra. Comparative review of selection techniques in genetic algorithm. In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 515-519, 2015

[7] C. Low, M. Ji, C.-J. Hsu, C.-T. Su. Minimizing the makespan in a single machine scheduling problems with flexible and periodic maintenance. Applied Mathematical Modelling, 34(2):334-342, 2010

[8] C. Zhao, M. Ji, H. Tang. Parallel-machine scheduling with an availability constraint. Computers & Industrial Engineering, 61(3):778-781, 2011

[9] D. Xu, Z. Cheng, Y. Yin, H. Li. Makespan minimization for two parallel machines scheduling with a periodic availability constraint. Computers & Operations Research, 36(6):1809-1812, 2009

[10] K. Li, S.-L. Yang, H.-W. Ma. A simulated annealing approach to minimize the maximum lateness on uniform parallel machines. Mathematical and Computer Modelling, 53(5):854-860, 2011

[11] M. Yazdani, S. M. Khalili, M. Babagolzadeh, F. Jolai. A single-machine scheduling problem with multiple unavailability constraints: A mathematical model and an enhanced variable neighborhood search approach. Journal of Computational Design and Engineering, 4(1):46-59, 10 2016

[12] N. G. Hall, C. N. Potts, C. Sriskandarajah. Parallel machine scheduling with a common server. Discrete Applied Mathematics, 102(3):223-243, 2000

[13] N. Hashemian, C. Diallo, B. Vizvári. Makespan minimization for parallel machines scheduling with multiple availability constraints. Annals of Operations Research, 213(1):173-186, Feb 2014

[14] P. Kora, P. Yadlapalli: Crossover operators in genetic algorithms. A review. International Journal of Computer Applications, 162:34-36, 03 2017

[15] S. M. Lim, A. B. M. Sultan, M. N. Sulaiman, A. Mustapha, K. Y. Leong. Crossover and mutation operators of genetic algorithms. International journal of machine learning and computing, 7(1):9-12, 2017

[16] T. Cheng, C. Sin. A state-of-the-art review of parallel-machine scheduling research. European Journal of Operational Research, 47(3):271-292,1990

[17] T. Jiang, C. Zhang, H. Zhu, G. Deng. Energy-efficient scheduling for a job shop using grey wolf optimization algorithm with double-searching mode. Mathematical Problems in Engineering, 2018:8574892, Oct 2018

[18] V. Nguyen, N. Huynh Tuong, H. Tran, N. Thoai. An MILP-based makespan minimization model for single-machine scheduling problem with splitable jobs and availability constraints. pp. 397-400, 01 2013

ImpKmeans: An Improved Version of the K-Means Algorithm, by Determining Optimum Initial Centroids, based on Multivariate Kernel Density Estimation and Kd-Tree

Ali Şenol

Department of Computer Engineering, Faculty of Engineering, Tarsus University, Engineering Faculty, 33400 Tarsus, Mersin, Turkey
alisenol@tarsus.edu.tr

Abstract: K-means is the best known clustering algorithm, because of its usage simplicity, fast speed and efficiency. However, resultant clusters are influenced by the randomly selected initial centroids. Therefore, many techniques have been implemented to solve the mentioned issue. In this paper, a new version of the k-means clustering algorithm named as ImpKmeans shortly (An Improved Version of K-Means Algorithm by Determining Optimum Initial Centroids Based on Multivariate Kernel Density Estimation and Kd-tree) that uses kernel density estimation, to find the optimum initial centroids, is proposed. Kernel density estimation is used, because it is a nonparametric distribution estimation method, that can identify density regions. To understand the efficiency of the ImpKmeans, we compared it with some state-of-the-art algorithms. According to the experimental studies, the proposed algorithm was better than the compared versions of k-means. While ImpKmeans was the most successful algorithm in 46 tests of 60, the second-best algorithm, was the best on 34 tests. Moreover, experimental results indicated that the ImpKmeans is fast, compared to the selected k-means versions.

Keywords: k-means; clustering; kernel density estimation; centroid initialization; kd-tree

1 Introduction

Clustering algorithms are unsupervised approaches, that separate data into groups, that are called clusters, according to included similarities and dissimilarities [1] [2]. Clustering approaches aim to maximize as much as possible both the similarities among the data in the same group and also the dissimilarity among the data in the different groups. In general terms, clustering algorithms are divided into five parts which are partitioning-based methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods; and DBSCAN [3], OPTICS [4], k-means [5], Affinity Propagation [6], Agglomerative Clustering [7],

HDBSCAN [8], and MCMSTClustering [9] are some examples of clustering algorithms. Some of the application areas of clustering are pattern recognition [10, 11], machine learning [12] [13], bioinformatics [14] [15], data mining [16] [17], web mining [1] [18], stream mining [12] [19], etc.

Basic k-means clustering was proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation to define linearly separable clusters. It is one of the partitioning-based clustering algorithms that divides the dataset into k clusters over randomly selected initial centroids. Although k-means is efficient and easy to use, it encounters problems if the dataset is not linearly separable. The main problems related to k-means are as follows:

- The final clusters are dependent on randomly selected centroids. As shown in Figure 1, if the randomly selected centroids are not located optimal, it fails while defining clusters.
- K-means clustering assumes that the shape of the clusters to be found is spherical. However, a minority of the datasets are spherical, and majorities are arbitrary in real-life.
- K-means cannot handle outliers because it partitions the data into k clusters without searching for outliers.
- It encounters some problems if the sizes of clusters are different.
- If the clusters are not linearly separable or overlapped, k-means encounters some issues.

Since the basic k-means clustering algorithm was proposed, many variants of it have been proposed to deal with mentioned issues that are given above [20]. Kernel k-means has been proposed to overcome the problem of identifying clusters that cannot be linearly separated [21] [22]. By using kernel methods, kernel k-means can define non-spherical clusters. However, kernel k-means run-time complexity is high, and its time complexity is high. On the other hand, to meet the need for selecting optimal initial centroids, many advanced versions of k-means were proposed, like k-means++ [22], and algorithms like Fuzzy C-Means, to automatically determine the number of clusters [23] [24]. In k-means++, cluster centers are chosen more innovatively to avoid complete randomness. Centroids are chosen step by step according to the centroids selected before to minimize the cost. This approach makes k-means++ better than basic k-means. However, this approach is not easy to perform. Fritzke proposed K-means-u*, an improved version of k-means++, to improve the limits of k-means++ [25]. But, used operations increase the complexity of the algorithm significantly. Another version of k-means to overcome the issue of selection of initial centroids was proposed by Zhang et al. [26]. In their study, although the accuracy of clustering results improved, it is unsuitable for big datasets because the algorithm's time complexity is high. Zhang et al. [27] proposed an advanced version of k-means based on density canopy to feed the k-means with the best initial centers. They used the canopy algorithm to

find the best values for k and initial centroids for k-means clustering. Although it effectively improves clustering quality, their method increases the algorithm's time complexity. As understood from explained versions of k-means clustering algorithm, we need a new k-means-based clustering algorithm that can cluster the dataset more accurately and quickly.

This study proposes a novel approach that uses multivariate kernel density estimation to find optimum initial centroids for k-means. Since kernel density estimation is a nonparametric probability density function (KDE), we use it to find denser regions to select them as initial centroids. It can find denser regions and the degree of density in any data distribution. The main contributions that this article has and state-of-the-art algorithms do not have are summarized as follows:

- The accuracy of k-means clustering algorithms increases thanks to using kernel density estimation to detect centroids of clusters.
- Because the detected centroids are also the final centroids, our approach does not need an iterative procedure to reach final clusters. Final clusters are formed in the first iteration. This method makes our approach very fast, when compared with existing methods.

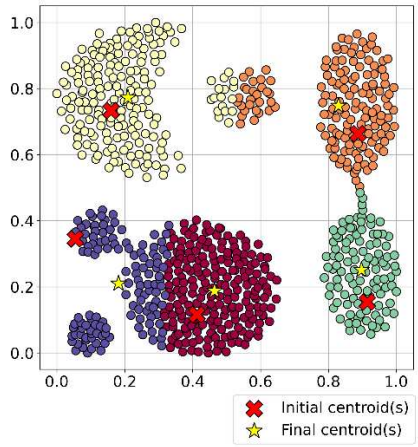


Figure 1

Final clusters of basic k-means according to randomly selected initial centers

The rest of the paper is organized as follows. In the 2nd section, related methods are explained, while in the 3rd section, the problem is stated. Then in the 4th section, details about the proposed algorithm are provided. Then, details of the experimental study are shared in the 5th section, while the work is concluded in the 6th section.

2 Preliminaries

2.1 Basic K-means

K-means clustering is based on the partitioning approach and is the basic clustering algorithm of clustering techniques. Its procedure is simple and effective if the shapes of clusters are spherical and there are no outliers. It uses an iterative approach over randomly selected initial centroids to reach the final cluster. But, just as an example is illustrated in Figure 2, initial centroids affect the final clusters directly. The main objective of iterations is to minimize the standard deviation of the dataset. The objective function of k-means is given in Equation (1).

$$J = min \sum_{j=1}^{k} \sum_{x_i \in C_i} \left\| x_i - \mu_j \right\|^2 \quad (1)$$

where k is the number of clusters, μ_j is the centroid of the j^{th} cluster, x is a data point, $\left\| x_i - \mu_j \right\|^2$ is the distance from the data point x_i to the cluster center, which is μ_j of the j^{th} cluster. Let X be the data points that construct the dataset; the pseudo-code of k-means is given in Algorithm 1.

Algorithm 1: Standard k-means

Input: Data points $X = x_1, x_2, x_3, \ldots, x_n \subseteq \mathbb{R}^d$;
k;
Output: A set of k centroids: $C = c_1, c_2, c_3, \ldots, c_k \subseteq \mathbb{R}^d$;
Initialize $C \leftarrow c_1, c_2, c_3, \ldots, c_k \subseteq \mathbb{R}^d$ at random
while C *has not converged* **do**
  $S_i \leftarrow \emptyset, \forall \in [k]$;
  foreach $x_i \in X$ **do**
    $j^* \leftarrow argmin_j \|x_i - c_j\|$;
    $S_{j*} \leftarrow S_{j*} \cup \{x_i\}$
  end
  $c_j \leftarrow \frac{1}{|S_j|} \sum_{x \in S_j} x, \forall j \in [k]$;
end

2.2 Kernel Density Estimation (KDE)

In the literature, two types of density estimation methods are commonly used. These methods are parametric and nonparametric approaches. Parametric methods like the Gaussian method assume all the data distribution is uniform and most data is gathered around the center in the circle with a standard deviation radius. In contrast, nonparametric methods assume there may be more than one denser area among the data. Namely, according to the parametric methods, there is only one peak on the curve; nonparametric methods assume there may be more than one peak. The probability density function of the univariate normal distribution with mean μ and variance σ^2 is given in Equation (2).

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{x-\mu}{\sigma}\right]^2/2} \quad (2)$$

where the value of x is in $-\infty < x < \infty$ interval. On the other hand, in addition to assuming that there may be more than one peak on the curve, nonparametric distribution estimation methods may not be uniform. Let $X = [X_1, \ldots, X_n]^T$ be an n-dimensional vector of multivariate Gaussian distribution of n-dimensional mean vector $\mu \epsilon \boldsymbol{R}^n$ and $\sum$ the covariance matrix of $n \times n$ dimensions. Therefore, the multivariate kernel distribution equation will be Equation (3) [28].

$$p(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\pi - \mu)^T \Sigma^{-1} (x - \mu)\right) \quad (3)$$

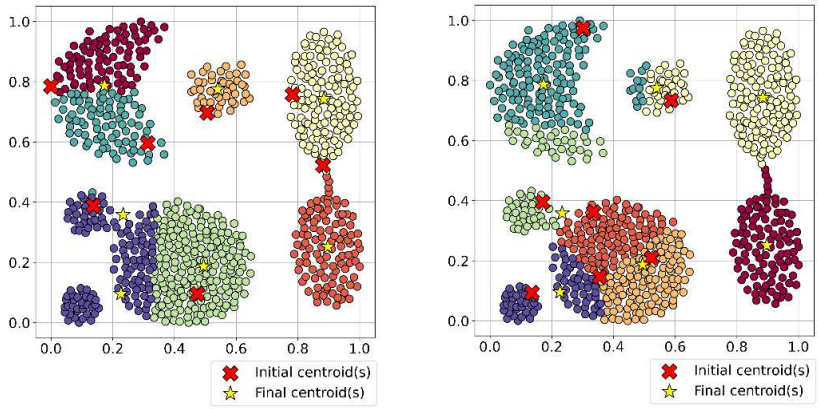


Figure 2

Two examples of the effect of randomly selected centroids on final clusters in standard k-means

As a nonparametric method, kernel density estimation tries to estimate where any new incoming data to locate according to existing data. Owing to this ability, KDE is used in many areas like machine learning, healthcare systems, stock markets, etc. [2]. As we mentioned earlier and as in the example in Figure 3, there can be multiple density peaks on the curve, and there are many types of KDE functions, known as smoothing functions, as given in Figure 4. Then, KDE is calculated as given in Equation (4), where K(.) is one of the functions in Figure 4.

$$\widehat{P_n}(n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \quad (4)$$

3 Problem Statement

The most important problem related to k-means is centroid initialization. Since the initial centroids are selected randomly in standard k-menas, both final clusters and the accuracy might be affected directly in a negative way. Although many advanced versions of k-means have been proposed, the time complexities of these algorithms

are very high. There is still a need for new k-means versions that can determine the best initial centroids and have low time complexity. In this study, we propose a new version of k-means to overcome the mentioned issues.

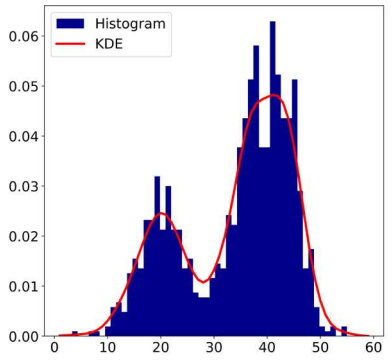


Figure 3

The relationship between histograms and peaks in KDE

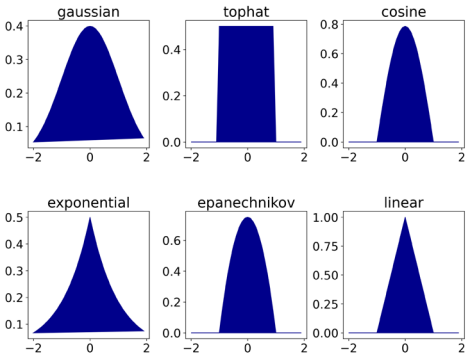


Figure 4

Types of kernel density estimation curves

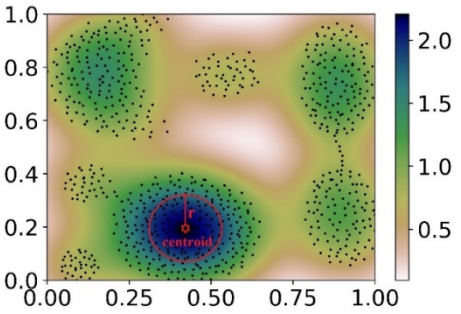


Figure 5

Example of the initial centroid and the radius used to determine the ignorance area on the Aggregation dataset

4 The Proposed Algorithm

This section describes the details of the proposed algorithm. In this study, we propose a new version of k-means clustering to overcome the issue of initial centroids determination. We try to detect peak points in the dataset to reach the goal. To find peak points, we used multivariate kernel density estimation. The purpose of finding peak points is to select determined peak points as initial centroids for k-means. Namely, the points with high KDE are candidates for initial centroids. Therefore, we select *k* points as initial centroids. However, as shown in Figure 5, selected *k* points with the highest values may be too close to each other. If these points were chosen as the initial centroids, the accuracy of the final clusters would be reduced. Therefore, in ImpKmeans, we use one more predefined parameter: ignorance radius. When searching for a new initial centroid using kd-tree and range search, we ignore data around previously selected centroids that are inside the radius of ignorance. Now, let's give more details and define the parameters used in ImpKmeans.

4.1 Definitions

Definition 1 (ignorance radius - ignorance_r): The ignorance radius determines the ignorance area around each selected centroid. This approach makes it possible for our algorithm to overcome local maxima. As given in Figure 5, if we didn't use this approach, all initial centroids would be selected from the same denser regions. This approach makes it possible to select initial centroids from different denser regions.

Definition 2 (the number of clusters - k): The number of clusters is the predefined number of clusters the user enters. However, this does not necessarily mean that there will always be k final clusters. In ImpKmeans, the formed clusters may be less than the selected k value.

Definition 3 (MultiKDE): As processed data is multidimensional, in ImpKmeans, we calculate the multivariate kernel density estimation value for each data. In addition to applying KDE to univariate data, we can apply it to multivariate datasets. To adapt the KDE to process multivariate datasets, we should use a kernel constructed by a product kernel or a radial basis function to process multidimensional datasets. Let's handle a 2-dimensional dataset. Let $X = (X_1, X_2, X_3, \dots, X_d)'$ be a sample of multivariate random variables with the density of *f(x)* defined on R^d and $\{x_1, x_2, x_3 \dots, x_n\}$ be an independent sample taken from *f(x)*. Then the multivariate kernel density estimation is calculated by Equation (5), where *K(.)* is a multivariate kernel function, and *h* is a positive bandwidth matrix.

$$\widehat{fh}(x) = \frac{1}{n|h|^{-\frac{1}{2}}} \sum_{i=1}^{n} K(h^{-\frac{1}{2}}(x - X_i)) \tag{5}$$

Definition 4 (kd-tree-based rangesearch): Figure 6 shows that kd-tree is a tree data type that can process multidimensional datasets. While placing the data into the tree,

it evaluates one dimension in each step. On the other hand, a rangesearch operation is an operation that is performed on any kd-tree to find the data inside a circle of radius of *r*. The reason to use this approach in our algorithms is that the kd-tree and the rangesearch have low computational complexity

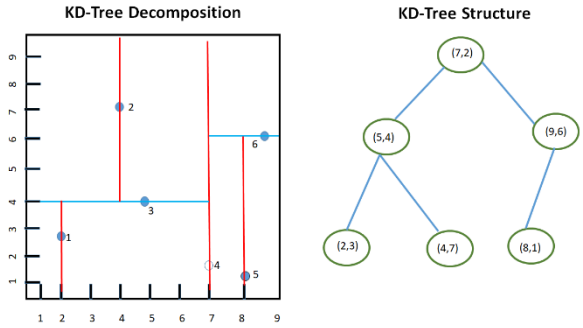


Figure 6
Decomposition of the kd-tree dataset

Algorithm 2: ImpKmeans

Input: Data points $X = x_1, x_2, x_3, \ldots, x_n \subseteq \mathbb{R}^d$;
k;
r;
Output: A set of k centroids: $C = c_1, c_2, c_3, \ldots, c_k \subseteq \mathbb{R}^d$;
$IntitialCentroids \leftarrow findInititalCentroids(X, k, r)$;
foreach $i \in k$ **do**
  $InitCentroids[i] \leftarrow argmax(MultiKDE)$;
  $MultiKDE \leftarrow deleteMultiKDE(r, InitCentroids[i])$
end
$S_i \leftarrow \emptyset, \forall \in [k]$;
foreach $x_i \in X$ **do**
  $j^* \leftarrow argmin_j \|x_i - c_j\|$;
  $S_{j*} \leftarrow S_{j*} \cup \{x_i\}$
end
$c_j \leftarrow \frac{1}{|S_j|} \sum_{x \in S_j} x, \forall j \in [k]$;

Algorithm 3: findInitialCentroids

Input: Data points $X = x_1, x_2, x_3, \ldots, x_n \subseteq \mathbb{R}^d$;
k;
r;
Output: A set of k centroids: $C = c_1, c_2, c_3, \ldots, c_k \subseteq \mathbb{R}^d$;
foreach $i \in k$ **do**
  $MultiKDE \leftarrow \frac{1}{n|h|^{-\frac{1}{2}}} \sum_{i=1}^{n} K(h^{-\frac{1}{2}}(x - X_i))$; ▷*calculate MultiKDE*
  $InitCentroids[i], j \leftarrow argmax(MultiKDE)$; ▷*find peak point*
  $kdtree \leftarrow KDTree(X)$; ▷*place data into kd-tree*
  $ind \leftarrow rangesearch(kdtree, r, x_j)$; ▷*find the data inside the ignorance area*
  $X \leftarrow delete(X, ind)$; ▷*delete the data in ignorance data*
end
return $InitCentroids$

4.2 Algorithm

As we explained above, the main contributions of ImpKmeans are that it can detect the best initial centroids for k-means and does not need an iterative search method to reach final clusters. As an example, illustrated in Figure 5, KDE makes it possible to find the denser areas, and ignorance radius makes it possible to find the initial centroids in different regions. ImpKmeans algorithm is divided into two parts:

- In the initial stage, kernel density estimation finds peak points in the dataset. k points that minimize the cost are selected as initial centroids.
- In the second stage, initial centroids proposed by kernel density estimation are given to basic k-means as initial centroids, and the dataset is clustered according to these centroids in only one iteration.

According to the abovementioned equations and explanations about our method, the pseudo-code of ImpKmeans is given Algorithms 2 and 3.

4.3 Time Complexity

Let *n* be the number of vectors of *d*-dimensions, *k* be the number of clusters in the dataset, and *i* be the number of iterations needed to be converged; the comparison of the time complexity of the proposed algorithm with the state-of-the-art algorithm is given in Table 1. Because we use kd-tree construction and range search on it, the time complexity of our algorithm is the summation of *O(dn logn)* for constructing the kd-tree and O($dn^{1-\frac{1}{d}+k}$) for reangesearch operation. In addition, *O(nkd)* is the complexity of assigning the data to selected initial centroids. Therefore, the allover time complexity of our algorithm is O($dn\ logn + dn^{1-\frac{1}{d}+k} + nkd$). This complexity could be simplified as O($dn^{1-\frac{1}{d}+k}$). As our algorithm does not use an iterative approach, it is expected to be faster, compared with the other algorithms.

Table 1
Time complexity comparison of algorithms

Algorithm	Complexity
k-means	*O(nkdi)*
k-mediods	*O(n^2kdi)*
k-means++	*O(n^2k^2di)*
FCM	*O(nkdi)*
ImpKmeans	*O(nkd)*

5 Experimental Study

5.1 Development Environment

In this study, to measure the efficiency of our algorithm, we tested it on synthetic and real datasets in the Anaconda environment by using Python programming language with the needed libraries. To measure its clustering accuracy and speed, we compared it with some state-of-the-art algorithms like k-means, kmeans++, k-medoids, and Fuzzy C-Means. All experimental studies were performed on a computer with 16 GB RAM, an Intel i7 processor, and Windows 11 operating system installed.

5.2 Experimental Setup

To be sure that each data is in the same range and to select parameters easily, the data were normalized with the min-max normalization. The equation of min-max normalization is given in Equation (6).

$$z_{ij} = \frac{x_{ij} - min x_j}{\max x_j - \min x_j} \tag{6}$$

Additionally, we used ARI (Adjusted Rand Index), Purity, and Silhouette Index to evaluate and compare the clustering quality of the algorithms. Equations about these indices are given in Equations (7), (8), and (9), respectively, where n_{ij}, a_i, b_j, be values obtained from the contingency table, k the number of clusters, and c and t are the maximum count of data in the related clusters.

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_i}{2}\right] / \binom{n}{2}} \tag{7}$$

$$Purity = \frac{1}{N} \sum_{i=1}^{k} max_j \left| c_i \cap t_j \right| \tag{8}$$

$$SI = \frac{1}{n} \sum_{i=1}^{k} \sum_{x \in C_i} \frac{b(x) - a(x)}{max\ (a(x), b(x))} \tag{9}$$

where *a(x)* is the average distance to all the data of the cluster that *x* is in, and *b(x)* is the average distance to all the data of the closest cluster that x is not in.

5.3 Used Datasets

Synthetic and real datasets were used in the experimental study to compare the success of our algorithm with the state-of-the-art algorithms. Since the main purpose of our approach is to improve the accuracy of k-means and reduce the time complexity, the selected datasets are spherical in general. On the other hand, to measure the efficiency of our algorithm on the imbalanced dataset, we select some imbalanced datasets like Outliers, Aggregation, and Thyroid. Details of the datasets used in the experimental study are given in Table 2.

Table 2
Used datasets

Dataset	Type	# of Features	# of data	# of class	Reference
Outliers	Synthetic	2	700	4	[29]
Corners	Synthetic	2	2000	4	[29]
Iris	Real	4	150	3	[30]
Breast Cancer	Real	8	699	2	[30]
Aggregation	Synthetic	2	788	7	[31]
Thyroid	Real	4	215	2	[30]
Xclara	Synthetic	2	3000	3	[32]
Twenty	Synthetic	2	1000	20	[33]
2d-10c	Synthetic	2	2990	10	[33]
2d-20c	Synthetic	2	1517	20	[33]
2d-3c	Synthetic	2	625	3	[33]
2d-4c	Synthetic	2	1260	4	[33]
D31	Synthetic	2	3100	31	[34]
R15	Synthetic	2	600	15	[34]
Diamond9	Synthetic	2	3000	9	[33]
Sizes1	Synthetic	2	1000	4	[33]
DS-850	Synthetic	2	850	5	[33]
Fourty	Synthetic	2	1000	40	[33]
S-set1	Synthetic	2	5000	16	[33]
St900	Synthetic	2	900	9	[33]

5.4 Experimental Procedure and Parameter Setting

In the experimental study, we used a random search method with randomly selected parameters to reach the best results for each algorithm. We run each algorithm on each dataset 100 times with randomly selected parameters of each algorithm for each index (ARI, Purity, and SI). The highest obtained value of each index on each dataset was the best value for the selected algorithm. Similarly, the parameters enabling us to reach this value were the best. On the other hand, we also compared the speed of algorithms on selected datasets.

5.5 Results on Both Synthetic and Real Datasets

We used the procedure explained in Section 5.4 to find the best parameters for each algorithm and clustering results. Obtained results are shown in Tables 3, 4 and 5. The ARI values of ImpKmeans, k-means, k-medoids, FCM, and k-medoids are shared in Table 3. Additionally, visual results are provided in Figure 7. According to the results, it is obvious that our algorithm is more successful on 16 datasets over 20 datasets. On the other hand, k-means, k-means++, k-medoids, and FCM were

the best on 9, 12, 6 and 8 datasets over 20 datasets, respectively. Iris, D31, Sizes1, and DS-850 were the datasets in that our algorithm was not the most successful. But the ARI values that our algorithm achieved were very close to the best values. Therefore, we can say that our algorithm is the best in datasets by the aspect of ARI.

Regarding Purity, our algorithm was the most successful on 15 datasets over 20, while k-means, k-means++, k-medoids, and FCM were the best on 11, 7, 5 and 12 datasets over 20 datasets, respectively. When we examined our algorithms' results on the datasets in which it was not the best; its Purity values were very close to the best. So, in Purity, our algorithm is very competitive compared to the other algorithms. As for SI, it was seen that our algorithm was the most successful on 15 datasets over 20 with k-means++, while k-means, k-medoids, and FCM were the most successful on 9, 6 and 8 datasets. As real datasets, we tested the algorithms on Iris, Breast Cancer, and Thyroid. Our algorithm was more successful in Breast Cancer and Thyroid. As for Iris, our algorithm is very close to the best values. Consequently, as presented in Table 6, our algorithm appears to be more successful when compared with the other algorithms.

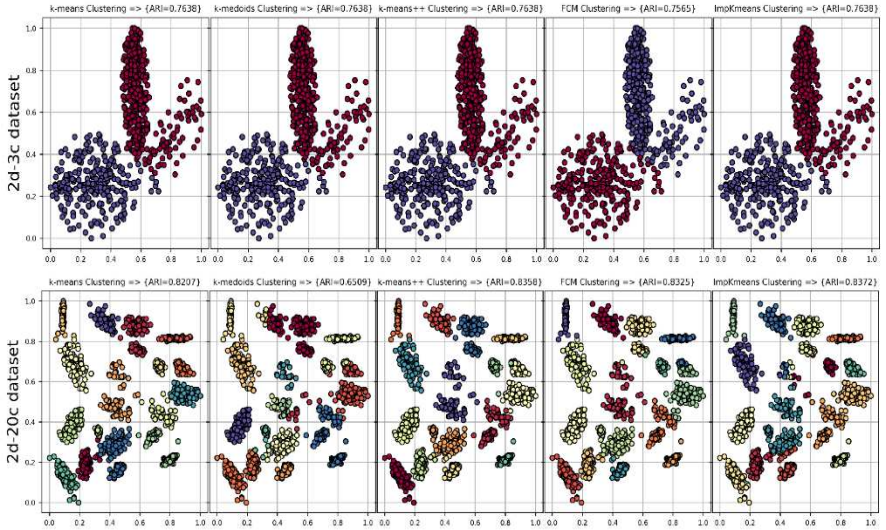


Figure 7
Cont.

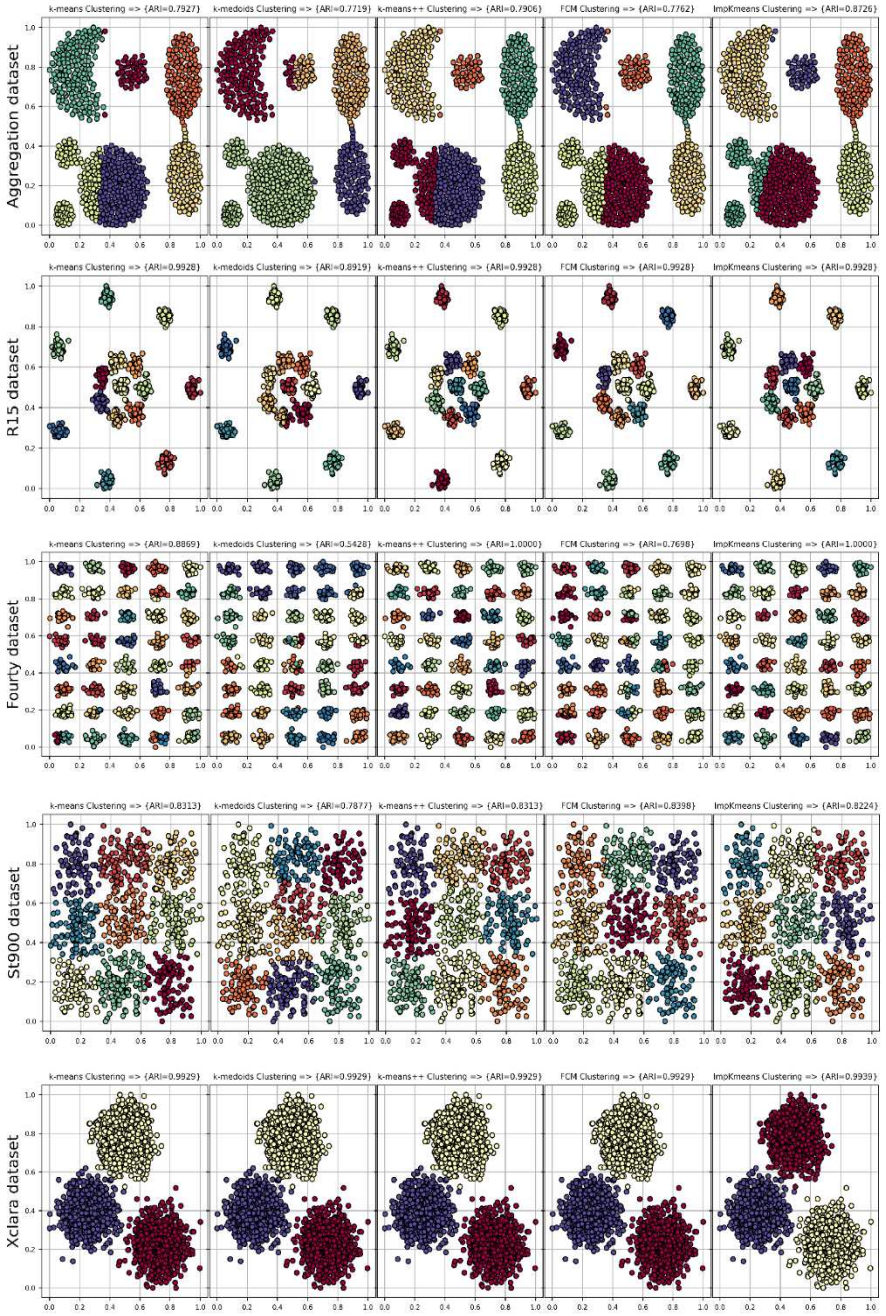


Figure 7
Cont.

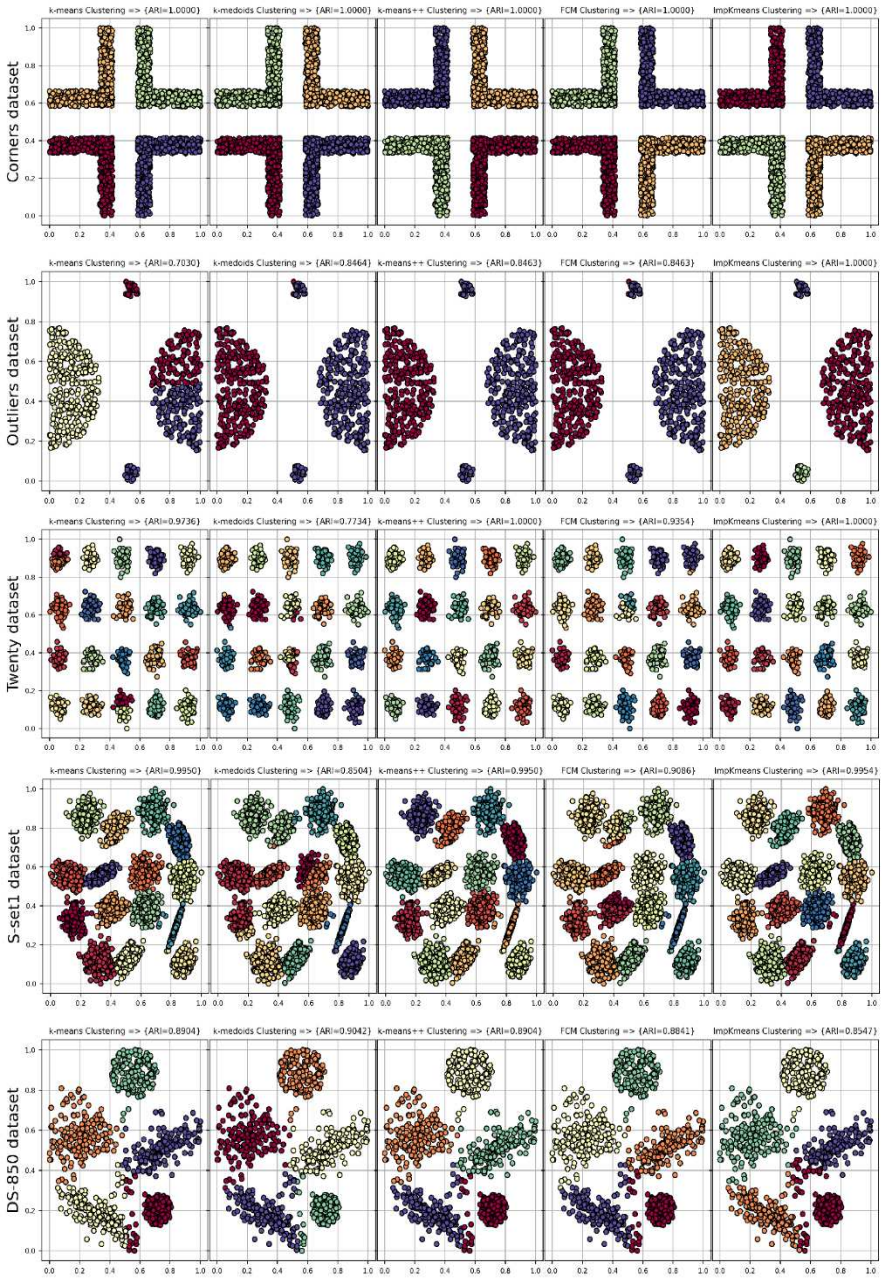


Figure 7
Cont.

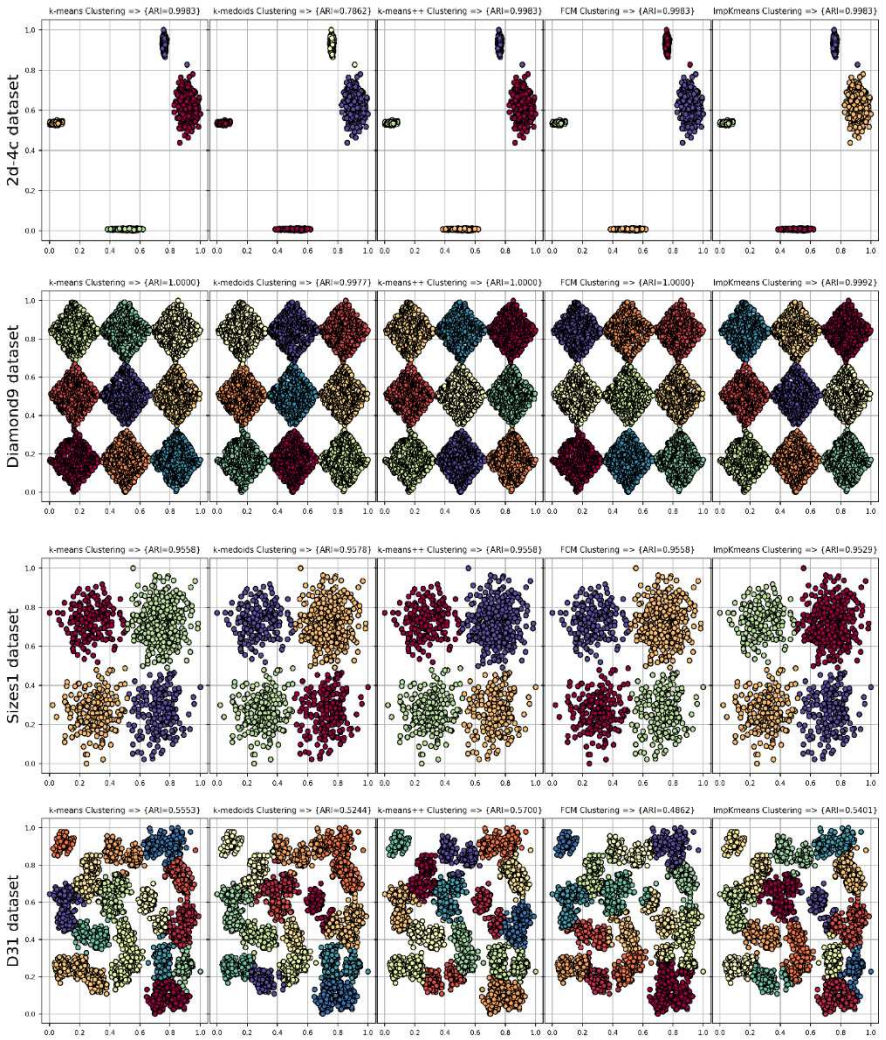


Figure 7
Visual clustering results (ARI values) of algorithms on the synthetical datasets

Table 3
ARI evaluation

	FCM	Lloyd's k-means	k-means++	k-medoids	ImpKmeans
Outliers	0.8463	0.703	0.8463	0.8464	**1**
Corners	**1**	**1**	**1**	**1**	**1**
Iris	0.7287	0.7163	0.7163	**0.743**	0.7163
Breast Cancer	0.8178	**0.8391**	**0.8391**	0.8284	**0.8391**

Aggregation	0.7762	0.7927	0.7906	0.7719	**0.8239**
Thyroid	0.6927	0.6283	0.6283	0.2055	**0.8434**
Xclara	**0.9929**	**0.9929**	**0.9929**	**0.9929**	**0.9929**
Twenty	**1**	0.9736	**1**	0.7734	**1**
2d-10c	**0.9967**	**0.9967**	**0.9967**	0.804	**0.9967**
2d-20c	0.8325	0.8207	0.8358	0.6509	**0.8377**
2d-3c	0.7565	**0.7638**	**0.7638**	**0.7638**	**0.7638**
2d-4c	**0.9983**	**0.9983**	**0.9983**	0.7862	**0.9983**
D31	0.5254	0.5553	**0.57**	0.5244	0.5489
R15	**0.9928**	**0.9928**	**0.9928**	0.8919	**0.9928**
Diamond9	**1**	**1**	**1**	0.9977	**1**
Sizes1	0.9558	0.9558	0.9558	**0.9578**	0.9558
DS-850	0.8841	0.8904	0.8904	**0.9042**	0.8904
Fourty	0.9006	0.8869	**1**	0.5428	**1**
S-set1	0.995	0.995	0.995	0.8504	**0.9954**
St900	**0.8398**	**0.8313**	**0.8313**	0.7877	**0.8313**

Table 4
Purity evaluation

	FCM	**Lloyd's k-means**	**k-means++**	**k-medoids**	**ImpKmeans**
Outliers	**1**	**1**	**1**	**1**	**1**
Corners	**1**	**1**	**1**	**1**	**1**
Iris	**0.98**	0.96	0.9667	0.9667	0.9733
Breast Cancer	0.9671	0.97	0.97	**0.9742**	**0.9742**
Aggregation	0.9949	**0.9962**	0.9949	**0.9962**	0.9949
Thyroid	0.9488	**0.9628**	0.9535	0.9023	**0.9628**
Xclara	**0.9997**	0.9987	0.999	0.9977	**0.9997**
Twenty	**1**	**1**	**1**	0.8	**1**
2d-10c	**0.9997**	0.9993	0.9993	0.8013	**0.9997**
2d-20c	**0.8471**	0.8154	0.8451	0.6711	0.8457
2d-3c	0.9902	**0.9944**	0.993	0.993	**0.9944**
2d-4c	**1**	**1**	**1**	**1**	**1**
D31	**0.4835**	**0.4835**	0.4823	0.4777	**0.4835**
R15	**0.9967**	**0.9967**	**0.9967**	0.9317	**0.9967**
Diamond9	**1**	**1**	**1**	0.999	**1**
xxSizes1	0.984	**0.986**	0.983	0.983	0.985
DS-850	0.9953	0.9988	0.9988	0.9918	**1**
Fourty	0.95	0.925	**1**	0.475	**1**
S-set1	0.9976	0.9976	0.9976	0.869	**0.9978**
St900	**0.9256**	0.9211	0.9211	0.9033	0.9211

Table 5
SI evaluation

	FCM	Lloyd's k-means	k-means++	k-medoids	ImpKmeans
Outliers	0.6128	0.5173	**0.6136**	0.6119	**0.6136**
Corners	0.5697	0.5534	0.5698	0.4693	**0.5699**
Iris	0.618	**0.6295**	**0.6295**	**0.6295**	**0.6295**
Breast Cancer	0.597	0.5966	0.5966	**0.5968**	0.5966
Aggregation	0.5279	0.5365	0.5365	**0.5385**	0.5365
Thyroid	0.5382	0.5755	**0.5852**	0.1909	**0.5852**
Xclara	**0.6945**	**0.6945**	**0.6945**	**0.6945**	**0.6945**
Twenty	**0.738**	0.6993	**0.738**	0.5472	**0.738**
2d-10c	**0.8368**	**0.8368**	**0.8368**	0.724	**0.8368**
2d-20c	0.6133	0.5967	**0.6172**	0.5374	0.6171
2d-3c	0.5517	0.5557	0.5557	0.5557	**0.5562**
2d-4c	**0.8738**	**0.8738**	**0.8738**	0.7184	**0.8738**
D31	0.4606	0.4821	**0.4832**	0.4483	0.4791
R15	**0.7528**	**0.7528**	**0.7528**	0.6843	**0.7528**
Diamond9	**0.5487**	**0.5487**	**0.5487**	0.5486	**0.5487**
xxSizes1	**0.5934**	**0.5934**	**0.5934**	**0.5934**	**0.5934**
DS-850	0.5635	0.5646	0.5646	**0.5652**	0.5646
Fourty	0.6082	0.6206	**0.6852**	0.4324	**0.6852**
S-set1	**0.7116**	**0.7116**	**0.7116**	0.6116	**0.7116**
St900	0.4417	**0.4436**	**0.4436**	0.4201	**0.4436**

Table 6
Overall cluster quality comparisons of the algorithms

Algorithms	ARI	Purity	SI	Total
k-means	9	11	9	29
m-medoids	6	5	6	17
k-means++	12	7	**15**	34
FCM	8	12	8	28
ImpKmeans	**16**	**15**	**15**	**46**

5.6 Speed Analysis

As we explained in Table 1, our algorithm is expected to be fair regarding run-time complexity. Because our algorithm does not use any iterative approach. In our approach, the time-consuming stage is the initial stage, in which the kernel density estimation based on initial centroids is determined. Experimental studies also support our idea, as seen in Figure 8. On the other hand, in some datasets, like Twenty, in which the number of clusters is high compared to the others, the

consumed time in the ImpKmeans is slightly higher. In other words, we can say that the higher the number of clusters, the higher the run-time complexity for ImpKmeans.

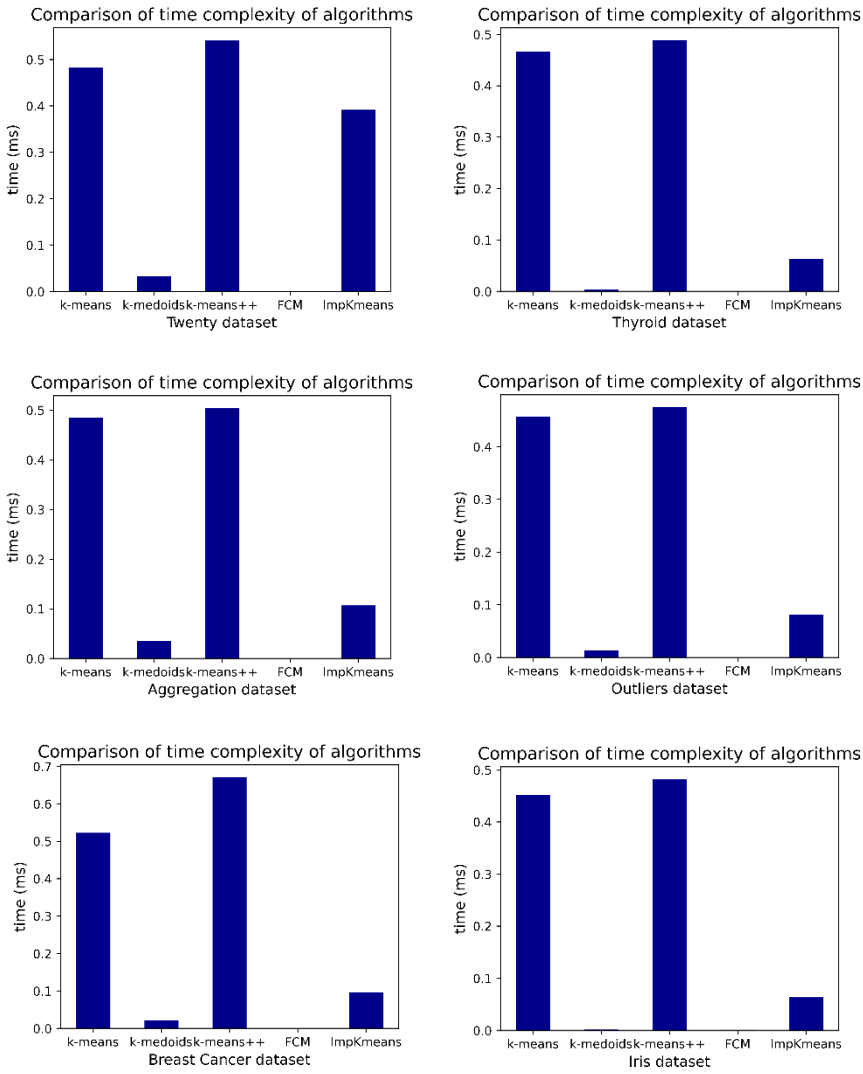


Figure 8

Time comparison of the algorithms on some of the used datasets

Conclusion and Future Work

In this study, we proposed a new advanced version of k-means, named "ImpKmeans", shortly to overcome the issues of initial centroid determination and the time complexity, that the other versions of k-means face. Our approach gets its

power from using multivariate kernel density estimation, to find the denser regions among the data. Eligible k number of peaks, are selected as initial centroids, according to density. This approach makes our algorithm superior to the compared algorithms, in terms of clustering quality. Moreover, since the selected initial centroids are also the final cluster centroids, our algorithm produces the final clusters, in only one iteration. This approach makes our algorithm both, more effective and faster.

A significant experimental result was observed while testing algorithms on Outliers and Thyroid datasets. Our algorithm reached 100% and 84.34%, while the second-best algorithm reached 84.64% and 69.27% clustering quality, respectively. As the experimental studies also support, our algorithm has both successful clustering quality and low time complexity.

In the future, plans to examine various studies, addressing datasets with arbitrary-shaped clusters, will be conducted.

Code availability

Python implementation of the proposed clustering algorithm is shared on GitHub (https://github.com/senolali/ImpKmeans).

References

[1] Aggarwal, C. C. and C.K. Reddy, Data Clustering: Algorithms and Applications. 2014: CRC Press Taylor and Francis Group

[2] Węglarczyk, S. Kernel density estimation and its application. in ITM Web of Conferences. 2018, EDP Sciences

[3] Ester, M., H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996, AAAI Press: Portland, Oregon. pp. 226-231

[4] Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander, OPTICS: ordering points to identify the clustering structure. SIGMOD Rec., 1999. 28(2): pp. 49-60

[5] Lloyd, S. P., Least squares quantization in PCM. IEEE Trans. Inf. Theory, 1982. 28: pp. 129-136

[6] Frey, B. J. and D. Dueck, Clustering by Passing Messages Between Data Points. 2007, 315(5814): pp. 972-976

[7] Lance, G. N. and W. T. Williams, A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. The Computer Journal, 1967, 9(4): pp. 373-380

[8] Campello, R. J. G. B., D. Moulavi, and J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2013

[9] Şenol, A., MCMSTClustering: defining non-spherical clusters by using minimum spanning tree over KD-tree-based micro-clusters. Neural Computing and Applications, 2023. 35(18): pp. 13239-13259

[10] Sathya, B. and R. Manavalan, Image Segmentation by Clustering Methods: Performance Analysis. International Journal of Computer Applications, 2011, 29: pp. 27-32

[11] Li, C., F. Kulwa, J. Zhang, Z. Li, H. Xu, and X. Zhao, A Review of Clustering Methods in Microorganism Image Analysis, in Information Technology in Biomedicine, E. Pietka, et al., Editors. 2021, Springer International Publishing: Cham. pp. 13-25

[12] Şenol, A. and H. Karacan, A Survey on Data Stream Clustering Techniques. European Journal of Science and Technology, 2018(13): pp. 17-30

[13] Kumar, V., M. S. Chauhan, and S. Khan, Application of Machine Learning Techniques for Clustering of Rainfall Time Series Over Ganges River Basin, in The Ganga River Basin: A Hydrometeorological Approach. 2021, Springer, pp. 211-218

[14] Yu, Z., H.-S. Wong, and H. Wang, Graph-based consensus clustering for class discovery from gene expression data. Bioinformatics, 2007, 23(21): pp. 2888-2896

[15] Zou, Q., G. Lin, X. Jiang, X. Liu, and X. Zeng, Sequence clustering in bioinformatics: an empirical study. Briefings in bioinformatics, 2020, 21(1): pp. 1-10

[16] Han, J., M. Kamber, and J. Pei, Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 2011. 5(4): pp. 83-124

[17] Sabor, K., D. Jougnot, R. Guerin, B. Steck, J.-M. Henault, L. Apffel, and D. Vautrin, A data mining approach for improved interpretation of ERT inverted sections using the DBSCAN clustering algorithm. Geophysical Journal International, 2021

[18] Rambabu, M., S. Gupta, and R. S. Singh, Data Mining in Cloud Computing: Survey, in Innovations in Computational Intelligence and Computer Vision. 2021, Springer. pp. 48-56

[19] Şenol, A. and H. Karacan, Kd-tree and adaptive radius (KD-AR Stream) based real-time data stream clustering. Journal of the Faculty of EngineeringArchitecture of Gazi University, 2020. 35(1): pp. 337-354

[20] Şenol, A., Standard Deviation-based Centroid Initialization for K-means, in 3rd International Anatolian Congress on Scientific Research. 2022: Kayseri. pp. 523-530

[21] Schölkopf, B., A. Smola, and K. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, 1998, 10(5): pp. 1299-

1319

[22] Arthur, D. and S. Vassilvitskii, k-means++: The advantages of careful seeding. 2006, Stanford

[23] Bellman, R., R. Kalaba, and L. Zadeh, Abstraction and pattern classification. Journal of Mathematical Analysis and Applications, 1966. 13(1): pp. 1-7

[24] Ruspini, E. H., A new approach to clustering. Information and Control, 1969. 15(1): pp. 22-32

[25] Fritzke, B., The k-means-u* algorithm: non-local jumps and greedy retries improve k-means++ clustering. CoRR, 2017. abs/1706.09059

[26] Ze-bao, Z. J. C. S., Algorithm for Initialization of K-Means Clustering Center Based on Optimized-Division. 2009

[27] Zhang, G., C. Zhang, and H. Zhang, Improved K-means algorithm based on density Canopy. Knowledge-based systems, 2018. 145: pp. 289-297

[28] Şenol, A., VIASCKDE Index: A Novel Internal Cluster Validity Index for Arbitrary-Shaped Clusters Based on the Kernel Density Estimation. Computational Intelligence and Neuroscience, 2022. 2022: p. 4059302

[29] Kools, J. 6 functions for generating artificial datasets. 2023 July 10, 2023]; Available from: https://www.mathworks.com/matlabcentral/fileexchange/41459-6-functions-for-generating-artificial-datasets

[30] Dua, D. and C. Graff. UCI Machine Learning Repository. 2021; Available from: http://archive.ics.uci.edu/ml

[31] Gionis, A., H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Trans. Knowl. Discov. Data, 2007. 1(1): p. 4-es

[32] Zelnik-Manor, L. and P. Perona, Self-tuning spectral clustering, in Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004, MIT Press: Vancouver, British Columbia, Canada. pp. 1601-1608

[33] Parmar, M. Clustering Datasets. 2022 8.8.2022]; Available from: https://github.com/milaan9/Clustering-Datasets

[34] Veenman, C. J., M. J. T. Reinders, and E. Backer, A Maximum Variance Cluster Algorithm. IEEE Trans. Pattern Anal. Mach. Intell., 2002. 24: p. 1273-1280

Bolt Preload Variations During Repeated Tightenings

Talal Alsardia

Department of Railway Vehicles and Vehicle System Analysis, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, 1111 Budapest, Hungary
e-mail: alsardia@edu.bme.hu

Abstract: Bolt preload prediction, through torque value, is challenged by friction variations within the bolted joint. Accurately estimating the initially achieved preload, is a persistent problem. This study aims to examine the repeatability of the bolt tightening process, under constant torque values, using experimental data. The experiments were conducted on bolts and nuts with a black finish surface, and the preload and nut factor variations were examined under four lubrication scenarios.

Keywords: Bolt preload variation; Torque-preload relationship; Lubrication; Nut factor

1 Introduction

Fixtures are necessary to join two or more elements together and a wide variety of fixtures are available in the industry. Choosing the right fixture and providing a reliable design requires a thorough system analysis [1] [2]. Among these fixtures, bolted joints assumes an easy-to-release link between parts. Bolted joints have a wide range of applications, such as those seen in [3-5]. In bolted joints, the power is transmitted from one part to another through friction. The task of the bolted link is to assume the normal force allowing this friction. The normal force called preload is a result of the bolt tightening process. The highest allowed bolt preload is usually estimated as a percentage of the bolt material yield strength [6]. Both insufficient or excessive bolt preload can lead to joint failure. Therefore, several methods are used to control the bolt preload, such as torque and angle control, bolt elongation control, and torquing control [7]. During bolt tightening, approximately 10%-20% of the applied torque generates preload, in the bolt. The remaining part is used to overcome friction in the bolted joint [8] [9]. Figure 1 illustrates the estimated torque distribution. It can be seen that friction plays an essential role in preload formation, and its variation in any region significantly impacts the bolt preload. In the literature, J. Drumheller [9] indicated that a 5% friction coefficient increase, under

the head or on the thread, could reduce the preload to half. Morgan and Henshall [10] investigated the effect of joint friction on the bolt preload of wheel bolts and nuts. They reported that repeated tightening processes caused up to a 50% reduction in bolt preload, while a constant state was reached when re-lubricated with engine oil.

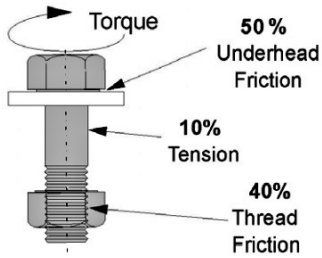


Figure 1
Tightening torque distribution [9]

Nassar et al. [11] examined the effects of the tightening speed and the repeated tightening on the wear pattern and torque-preload relationship for an M12x1.75 bolt of 8.8 grade. They found that for the non-lubricated case, the washer's surface roughness almost doubled after the fifth tightening. Also, they found no significant change in the friction coefficient when increasing the tightening speed after the fifth tightening cycle. However, when lubrication was used on the zinc coating material, the friction coefficient slightly decreased at increasing tightening speed. Eccles et al. [6] investigated the effect of repeated tightening on electro-zinc plated (EZP) nuts, bolts, and washers. They found significant wear on the bolt/nut thread contact surfaces and nut top face. In addition, they observed an increase of 100% in the friction coefficient while the preload decreased to 50% by the tenth tightening. They suggested a nonlinear empirical model for the tightening torque-preload relationship based on the number of tightening cycles. Z. Liu et al. [12] studied the frictional behavior of high-strength bolts during repeated tightening. Their observations are based on the so-called nut factor, discussed later in this paper. W. A. Grabon et al. [13] conducted a systematic tribological study on the threaded fastener. They linked the friction coefficient increase to the plastic deformation at the threads, which was affected by the presence of coating material and lubrication. B. Güler and K. T. Gürsel [14] used the torque-angle control tightening method to investigate a vehicle chassis zinc-coated joint. They concluded that the repetitive tightening process caused an increase in the friction coefficient and linked it to the large worn-out coating material. They summarized the main factors affecting the tightening process in a fishbone diagram.

Actually, manufacturers estimate a tightening torque appropriate for the first tightening in industrial applications. They frequently recommend installing a brand-new bolt-nut pair and washer after disassembly for safety reasons. In various applications, the everyday practice is that the disassembled fasteners are reused due to poor maintenance, cost saving, or lack of spare parts.

Most of the previous experimentally conducted research was conducted on bolts with known standardized material specifications. In everyday practice, there are cases when commercial bolts with unknown material specifications are used during the maintenance operation. The present work targeted such cases during the investigation of the bolts preload behavior. The bolts are taken from fastener shops, as the cheapest types. The bolts and nuts are from the same fastener box, but the manufacturer, material specifications, and the batch numbers are unknown.

Based on the literature background and everyday experience, this research aims to investigate the behavior of the bolt preload and the friction coefficient under repeated tightening-releasing cycles made on the same bolt, using the recommended tightening torque during the entire process. Moreover, we investigated the influence of the bolt/nut surface finish and the presence of lubrication on the generated preload. We conducted experiments to simulate typical tightening-releasing cycles. A perfect application of this idea is the vehicle wheel bolt. During the service time, the wheel bolts are periodically released and then tightened in case of seasonal tire changes or brake pad repairs. If only seasonal tire changes are considered, the bolt bears ten releasing/tightening cycles over five years of service. Due to its special shape, the wheel bolt or nut is not changed during the vehicle's lifetime.

NOMENCLATURE

D	Bolt nominal diameter (mm)	P	Pitch of the thread (mm)
d_2	Bolt pitch diameter (mm)	r_n	The effective bearing radius (mm)
d_h	Clearance hole diameter (mm)	r_t	The effective thread radius (mm)
D_o	Bearing surface outer diameter (mm)	T_{Pitch}	Torque to generate bolt tension (N.m)
F_i	Preload at the i^{th} tightening cycle	Vi	Percentage change of preload value relative to the initial tightening
F_P, F,	Clamping force, preload (kN)	X	Geometrical and frictional parameters of the joint
K	Nut factor	α	Thread lead angle (°)
T_{in}, T_{Input}, T	Input tightening torque (N.m)	β	Metric thread profile angle (°)
M_H, T_{Head}	Bearing surface friction torque (N.m)	μ_n, μ_b	Under the head friction coefficient
MoS_2	molybdenum disulfide powder	μ_t, μ_{th},	Thread friction coefficient
M_T, T_{Thread}	Thread friction torque (N.m)	ρ'	Computed angle of friction cone on thread surface (°)

2 Theoretical Background

Generally, in machine element theory, the bolt tightening torque equations are composed of three members, where two are linked to friction, and only one is linked to the preload:

$$T_{Input} = T_{Pitch} + T_{Threads} + T_{Head} \tag{1}$$

Motosh [15] introduced the following form in 1976:

$$T_{in} = F_P \left(\frac{P}{2\pi} + \frac{\mu_t r_t}{\cos(\beta/2)} + \mu_n r_n \right) \tag{2}$$

The standard DIN EN ISO 16047 also shows a similar variant:

$$T = F \left(\frac{1}{2} \cdot \frac{P + 1{,}154 \cdot \pi \cdot \mu_{th} \cdot d_2}{\pi - 1{,}154 \cdot \mu_{th} \frac{P}{d_2}} + \mu_b \cdot \frac{D_o + d_h}{4} \right) \tag{3}$$

As each member contains the preload force F as a parameter, a generalized form can be written as:

$$T = F \cdot X \tag{4}$$

Here T is the input torque, F is the bolt preload, and the constant X reflects the geometrical and frictional parameters of the joint. The problem with this structure is that it is based on a two-dimensional model (axial section cut) of the threaded fasteners. This model uses the assumptions of uniformly distributed contact pressure along the engaged thread surfaces [16] and a constant friction coefficient at every surface pair, as it is difficult to measure the friction under the head bolt/nut and on the thread during rotation. It is known that the real friction coefficients are different during each tightening, and it is not easy to have a constant guess value. From a practical point of view, another short expression can be used for the torque-tension relationship [11, 13, 17, 18], where the input tightening torque T is related to the bolt preload F in the function of the bolt diameter D. Here the constant containing all friction parameters is called the nut factor K (Torque coefficient ISO 16047). This gives the following simple equation:

$$T = K \cdot F \cdot D \tag{5}$$

The ASME Standard PCC-1 [19] states that "K is an experimentally determined, dimensionless constant related to the coefficient of friction." Equation (6) has a simple form and is easy to use, as it contains standard measurable data. Unfortunately, many experiments are required to get statistically firm data for each bolt diameter with an acceptably narrow confidence level. These statistically sound results are more accurate for industrial applications [7].

3 Methodology and Experimental Procedure

All experiments were performed according to ISO16047 in a closed room where the air conditioning kept the temperature and humidity at constant levels (25°C, 55%), to have a consistence clamping force results. since the clamping force can be affected by temperature and humidity variations [7].

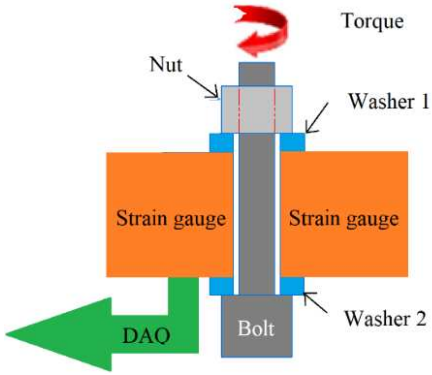


Figure 2
Representation of the tested bolted link

3.1 Preload Variation Measurement

In these experiments, M8x40 full-threaded bolts were used. The bolt head was clamped into a vise; then, a bolt force sensor was mounted between two special washers. The tightening element was an M8 nut (Figure 2). The torque was applied through a ½-inch mechanical torque wrench (Brüder Mannesmann Werkzeuge) of a type II class A with a range of (10-210 N.m, ± 4%). The input torque was set to a constant value of 20 N.m. Referring to ISO 898-1 and ISO 898-7 standards, for bolt of grade 10.9 the breaking torque is specified to be 40 N.m, and according to [20] the recommended tightening torque falls within the range of 50%-60% of the breaking torque, to generate tension in the bolt approximately 60% to 70% of elastic limit (yield strength) of the bolt material. After tightening, the preload force was noted, and then the nut was released till it became loose, and the preload became zero. This tightening cycle was repeated twenty times on the same bolt. For having a statistical base, twenty new bolts and nuts were used, with twenty tightening cycles each.

The effect of the tightening speed was not taken into consideration, as it has little effect on the nut factor [21-23]. Note that all the tightening cycles were conducted by the same operator (the author), using the same tool, in the same experimental environment. Figure 3 shows a sample of the conducted tightening-releasing cycle as a function of time. Here the colored area indicates the time duration needed to generate the preload. The calculated average time is of 1.06 seconds, per one tightening execution.

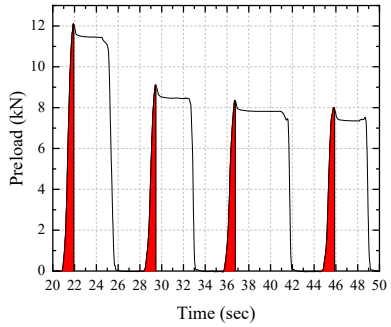


Figure 3
Bolt tightening-releasing cycle vs. time

Four cases of lubrication were studied to consider the effect of an eventual lubrication. In the first case "out of the box", we used the surfaces as they were obtained from the manufacturer, which named as “As-is” case in this article. In the second case, we applied drops of mineral-based 15W-40 motor oil (MOL MSE) on the bolt thread and under the bolt head surface before the first tightening. This is the “oiled” case. In the third case, all surfaces were cleaned with a degreaser (Loctite SF 7061) before the first tightening. This is the “dry” case. In the fourth case, solid powder lubricant (molybdenum disulfide, MoS_2) was applied on the threads and under the bolt head before the first tightening. For each lubrication case, brand-new bolts and nuts were used. Overall, 80 bolts and nuts were used in the study, and 1600 individual measurements were executed (Figure 4).

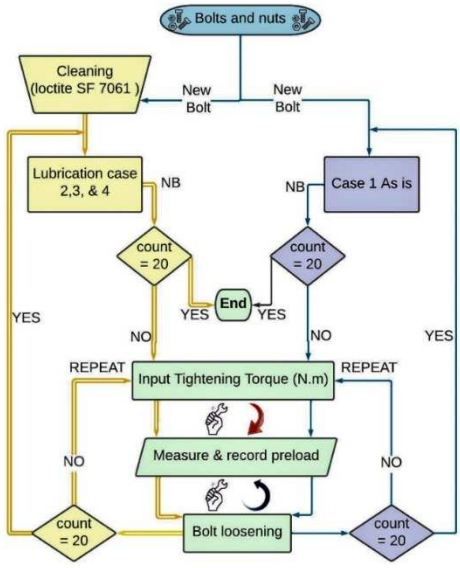


Figure 4
Experimental process flowchart

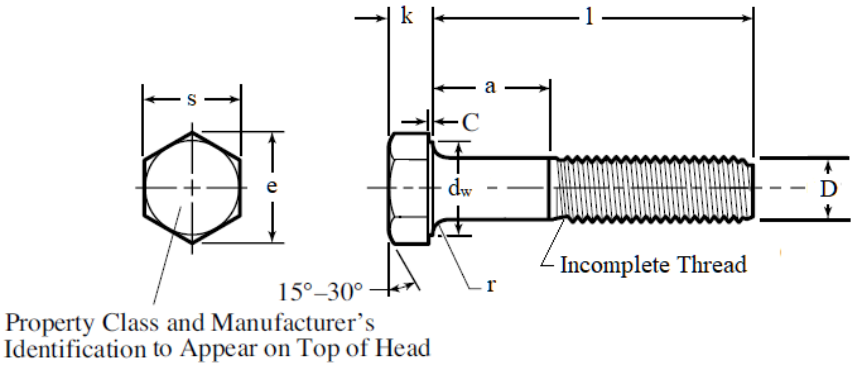


Figure 5
Tested bolt schematic[24]

The data acquisition system consisted of an HBM Quantum X data collector device, a computer, and an HBM KMR+/ 40kN bolt force sensor calibrated as per VDI/VDE 2638, with an accuracy class of 1.5 (±1%). The technical information and geometrical data of the tested bolt (Figure 5) are summarized in Table 1. Furthermore, Figure 6 shows the experimental setup layout.

Table 1
Bolt specifications

Surface finish		Black finish		
Grade	Bolt	10.9		
	Nut	8		
Size, d [mm]		8	Thread lead angle, α [°]	3.168
Thread pitch [mm]		1.25	Distance across flat, s [mm]	13
Metric thread profile angle, β [°]		60	Distance across corner, e [mm]	15
Tightening torque (N.m)		20	Head thickness, k [mm]	5.25
d_1[mm]		7.188	Unthreaded length, a [mm]	2.25
d_2[mm]		10.75	Washer face depth, c [mm]	0.4
Assumed μ_{th} and μ_n from		[25]	Washer face diameter, dw [mm]	12
Computed angle, ρ' [°]		6.587	Head junction radius, r [mm]	0.3

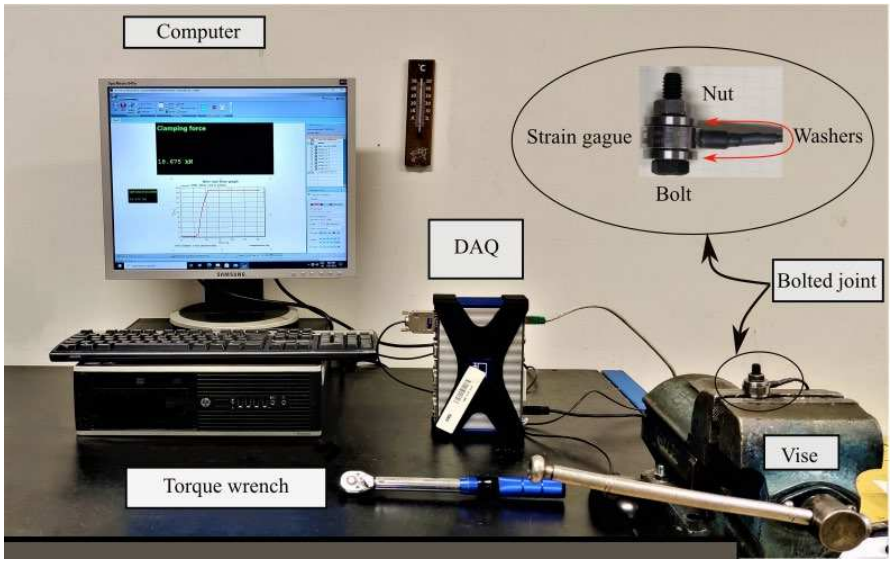


Figure 6
Experimental setup

4 Results and Discussion

The results of the experiments are discussed in two parts. The first part describes the bolt preload variations, and the second the nut factor variation.

4.1 Bolt Preload

In this section, the collected experimental preload data are presented. The effects of the presence of lubrication during the repeated tightening-releasing are shown. To validate the results, a statistical study was performed to show the significance of the achieved data.

The theoretical preload values were calculated using the information provided in Table 1. The values obtained from equation (1) and equation (3) were 17.369 kN and 17.365 kN, respectively. The assumed friction coefficient was used consistently for both equations, resulting in no significant difference between the values. Table 2 presents the mean of the measured preload achieved in the first tightening, sorted in descending order. It can be observed that the theoretical equation is validated for the oiled case. However, when the surfaces are dry, the preload values are lower than the theoretical predictions. Conversely, for the As-is and MoS_2 lubricant cases, the preload during the initial tightening exceeds the theoretical values.

Table 2
Mean of the measured preload for the first tightening cycle

Case	Measured Preload (kN)	Assumed μ_{th} and μ_n	Theoretical Preload (kN)
MoS_2	23.71	0.08 [25]	20.813
As-is	22.83	0.12 [25]	14.904
Oiled	17.71	0.1 [25]	17.369
Dry	12.95	0.13 [25]	13.916

A detailed representation of the distribution of the experimental data for each repetition is presented in the form of boxplots. The medians of the data are depicted by the centers of the boxes, while the interquartile range, represented by the height between the ends of the boxes, provides insight into the spread of the data within the middle 50% of the observations. The height of the whiskers further illustrates the dispersion of the data. All cases have the same vertical axis scale for ease of comparison.

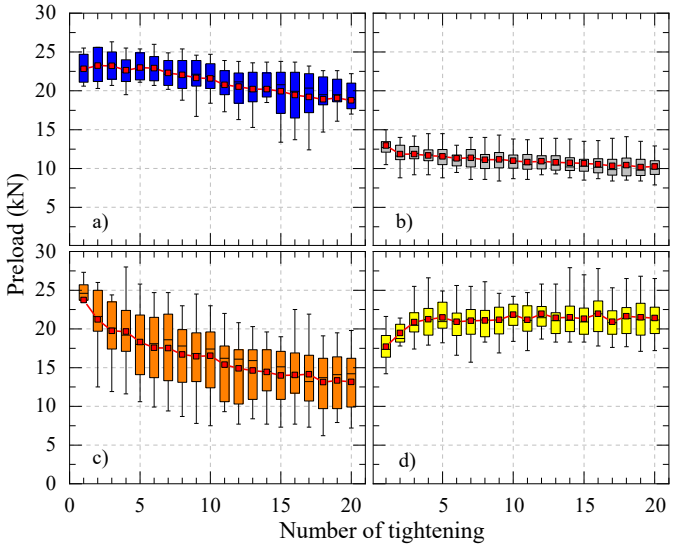


Figure 7
Preload box plot of the black bolt under lubrication conditions: a) As-is, b) dry, c) MoS_2, d) oiled

Let us consider the preload variation (Figure 7). Here, in the As-is case, the median of the preload decreases slightly when the number of tightening cycles increases. The range of the lower quartile increases at the last few tightening runs indicates that the distribution is negatively skewed, and there is a higher tendency to get a lower preload value as the number of tightening increases, see Figure 7. In the dry case, the median of the preload decreases almost linearly with the number of tightening cycles. The short box plots show a small median variation in the observed

data, with a platykurtic distribution (Figure 8). In the MoS_2 lubricated case, the result shows a significant scatter with a symmetrical distribution: this could be related to the presence of the rolling powder on the friction surfaces. The preload median gradually decreases as the number of tightening repetitions increases. Finally, in the oiled case, the median gradually increases during the first five tightening cycles and then stabilizes at a value around 20% higher than the initial preload.

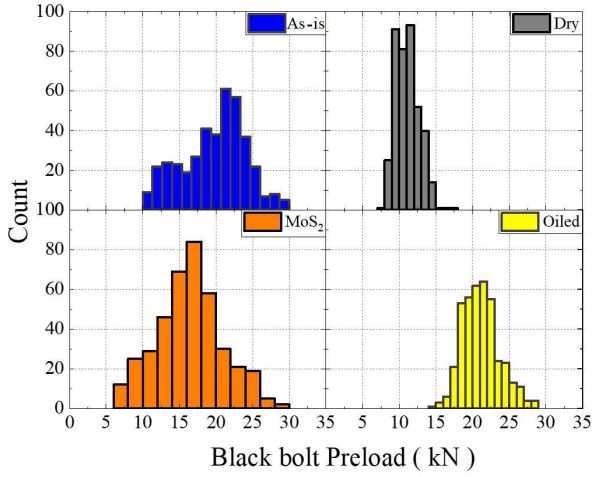


Figure 8
Histogram of black bolt preload for different lubrication conditions

Figure 9 shows the effect of repeated tightening on the mean of the measured preload for different lubrication conditions. The following remark can be made:

1) The preload values exhibited alterations in their relative order as the number of cycles increased. During the second cycle, the order of preload values from highest to lowest was As-is, MoS_2, oiled finally dry. This order changed the subsequent cycles. By the third cycle, the oiled condition had a higher preload value than the MoS_2, which persisted until the tenth cycle. The final order became Oiled As-is, MoS2, and finally Dry, which remained unchanged until the twentieth cycle.
2) For the As-is condition, the preload values decrease gradually from 22.8 kN with a few fluctuations until reaching its minimum value of 15.7 kN at cycle 20.
3) For the MoS_2 condition, the preload values decrease continuously with a few fluctuations until cycle 20, where it reaches its minimum value of 13.1 kN.
4) For the Oiled condition, the preload values show a slight increase from the first cycle to the fifth cycle and then stabilize for the remaining cycles.
5) For the dry condition the preload is the lowest among the tightening

repetitions, and the decreasing range is the smallest (from 12.95 kN to 10.26 kN).

6) Generally, the preload mean values change upon a bilinear curve as a function of the number of tightening cycles. The slope change happens around the fifth tightening for the As-is and the oiled lubrication condition. The first and the second slope depend on the type of lubrication. The trends are summarized in Table 3.

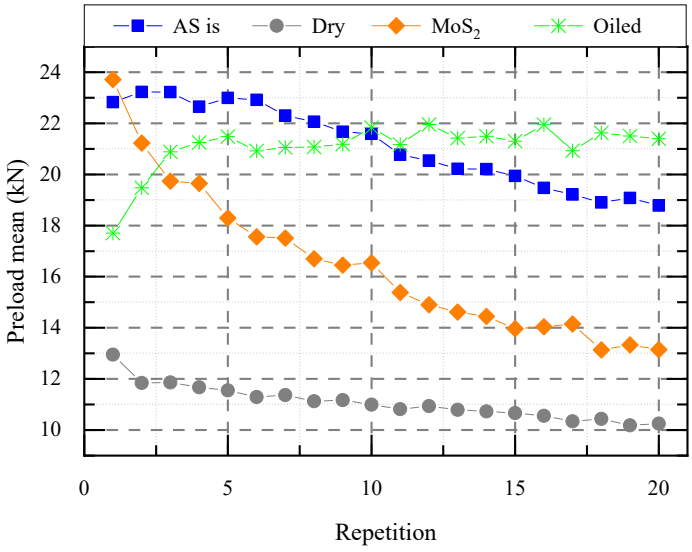


Figure 9
Effect of repeated tightening on the preload for the black bolt under different lubrication conditions

Table 3
Preload variations in the case of black bolts

Lubrication	Range (kN)	First slope	Second slope
As-is	22.8 to 15.7	0	decrease
Dry	13 to 10.3	Slightly decrease	Slightly decrease
MoS_2	23.7 to 13.1	Strong decrease	Strong decrease
Oiled	17.7 to 21.4	increase	0

The collected data is related to a randomly selected 80 bolts with mating nuts, visually inspected for any seen damage, and randomly assigned to each lubrication group. We have the same sample size n=20, over all cycles. For having statistical evidence about the suppositions made previously, a two-way mixed ANOVA design as described in [26] was performed after testing the ANOVA assumption's applicability (e.g., normality, homogeneity, the assumption of sphericity). The target was to compare the means of groups cross-classified by two different

types of factor variables, including:

A. Between-subjects factors, which have independent categories:
    1. Lubrication condition (As-is, dry, MoS_2, and oiled)

B. Within-subjects factors, which have related categories, also known as repeated measures:
    1. Cycle (1,…, 20)

The two-way mixed ANOVA tested three null hypotheses (at a significant level α= 0.05) are:

1) The means of preload force are equal for the four lubrication conditions (As-is, dry, MoS_2, and oiled).

   $\mu_{As\text{-}is} = \mu_{Dry} = \mu_{MoS2} = \mu_{Oiled}$

2) The means of preload force are equal over the 20 cycles.

   $\mu_1 = \mu_2 = ... = \mu_{20}$

3) There is no significant effect of the interaction between the lubrication conditions and the cycle.

The two-way mixed ANOVA test was performed using the R software environment, and the results are summarized in Table 4.

Based on the ANOVA result in Table 4, all the main effects of the two factors significantly affected the preload value at $\alpha = 0.05$ since (p-value$< 2e^{-16}$). which leads us to reject the three null hypotheses and accept the alternative hypotheses as follows:

1) The mean of the preload is different for different lubrication conditions.
2) The mean of the preload over the 20 cycles is different for the same lubrication condition.
3) There is a significant interaction effect between the cycle and lubrication conditions.

Table 4
Result table for the Two-way mixed ANOVA

Only between subject factor					
Source of variation	DF	Sum Sq	Mean Sq	F value	Pr(>F) = p-value
Condition	3	27229	9076	83.74	$< 2e^{-16}$
Residuals	76	8237	108		
Between subject factor over the within the factor					
Source of variation	DF	Sum Sq	Mean Sq	F value	Pr(>F) = p-value
Cycle	1	1841.2	1841.2	183.41	$< 2e^{-16}$

Condition : Cycle	3	2263.3	754.4	75.45	$< 2e^{-16}$
Residuals	76	762.9	10.0		

In what follows, we study the overall bolting performance with two more tools: lubrication control and cycle control.

As lubrication control, a Dunnett's test was used to examine the two factors' interactions. This test is used to compare a one-factor level's effect on the response when one of the levels is assumed to be controlled under the null hypothesis:

$$H_0: \mu_{group(i)} - \mu_{control} = 0$$

where $\mu_{group(i)}$ - the mean of the treatment group i, $\mu_{control}$ - the mean of the control group.

In this research, the level "As-is" is assumed to be the control level in the factor lubrication condition. Table 5. Present the outcomes of the Dunnett`s test. From the result, we can conclude the following. For the first cycle, the As-is lubrication was significantly better than the dry and the oiled lubrication, with no significant difference from the MoS_2 lubrication. After five cycles, the As-is lubrication was significantly better than the dry and the MoS_2 lubrications, with no significant difference from the oiled lubrication. After 10 and 15 cycles, the As-is lubrication is still significantly better than the dry and the MoS_2 lubrications, with no significant difference from the oiled lubrication. Finally, after 20 cycles, the oiled lubrication significantly becomes the best lubrication condition.

Table 5
Result of Dunnett's test of the lubrication conditions factor with control level "As-is" over five levels of the cycle factor (1, 5, 10, 15, and 20)

Cycle	Comparison	Difference	p-value
1	Dry – As-is	-9.880	$<2e^{-16}$
	MoS_2 – As-is	0.883	0.473
	Oiled – As-is	-5.125	$5.4e^{-10}$
5	Dry – As-is	-11.440	$<2e^{-16}$
	MoS_2 – As-is	-4.696	$1.5e^{-7}$
	Oiled – As-is	-1.510	0.149
10	Dry – As-is	-10.545	$<2e^{-16}$
	MoS_2 – As-is	-5.050	$7.4e^{-8}$
	Oiled – As-is	0.245	0.982
15	Dry – As-is	-9.275	$<2e^{-16}$
	MoS_2 – As-is	-5.973	$1.5e^{-9}$
	Oiled – As-is	1.360	0.279
20	Dry – As-is	-8.525	$4.9e^{-15}$
	MoS_2 – As-is	-5.633	$5.3e^{-8}$
	Oiled – As-is	2.620	0.0143

For the cycle control, to illustrate the effect of the cycle, let us consider the following expression for the percentage change of the preload force:

$$V_i = \left|\frac{F_1 - F_i}{F_1}\right| \times 100\% \quad (6)$$

Here *i=1..20* indicates the number of tightening cycles. The preload coming from the first tightening (F_1) is considered as a reference since the manufacturers often prescribe using a new bolt and washer after disassembly. The summary of the variations is plotted in Figure 10. The observations are the following:

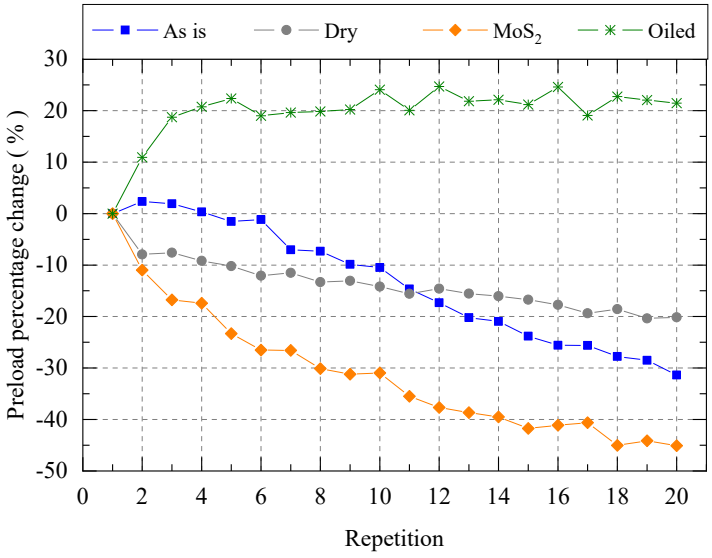


Figure 10
Preload percentage decrease relative to the first tightening

1) In the case of As-is lubrication, the bolt preload loss fluctuates around zero till the fifth tightening. As the number of cycles increases, the preload value loss at the 20th cycle is approximately 31% of the first cycle.

3) For the "MoS2" condition, the preload values decrease more steeply throughout the 20 cycles. The preload value at the 20th cycle was approximately 44.7% lower than that of the first cycle.

4) The presence of oil stabilizes the preload variation and has the best performance among the other lubrications. The preload values show a slight increase till the fifth cycle by 20% higher than the first value and then fluctuate around that percentage.

5) Preload percentage change in the function of tightening cycles for different lubrications is not the same. The oil lubrication gave the best performance; and the rank ordering is: oiled, As-is, MoS_2, and dry until the 11th cycle; then the dry became better than the As-is lubrication.

4.2 Nut Factor Variation

In this section, an analysis of the variation in the nut factor is presented. The nut factor (K) was calculated utilizing the collected experimental preload data and equation (5). The calculation of the nut factor was performed for each tightening instance, and subsequently, the mean value was computed.

For the first cycle, Table 6 shows the nut factor for each lubrication case, where the calculated nut factor value is calculated using the measured preload values, and the theoretical nut factor value is calculated using the theoretical preload values. The difference between these two values can indicate the accuracy in predicting the friction coefficient for the theoretical equation for each lubrication condition. The theory was valid only for the oiled lubrication condition.

Table 6
Theoretical and calculated nut factor for each lubrication condition at the first cycle

Lubrication	Calculated Nut factor K	Theoretical Nut factor K
MoS_2	0.105	0.12
As-is	0.110	0.171
Oiled	0.141	0.144
Dry	0.193	0.18

The mean values of the nut factor (K) are depicted in Figure 11. The nut factor is known to include the effect of all unknown factors that influence the relationship between the input torque and the output preload. Therefore, by using it, we can compare the behavior of each type of bolt preload under different lubrication conditions and have the following remarks:

1) A higher nut factor indicates higher losses in the input torque. It is very significant in the dry case.
2) The lowest initial nut factor means the highest preload value was reached. This happened when MoS_2 was applied.
3) Lubrication improves both the tightening process efficiency and its repeatability. The K value was less than 0.15 in the As-is and oiled cases. Note that the received black bolts might have a thin oil layer to avoid rusting during storage.
4) The lubrication type and the cycle number combinations have an important effect on the preload.

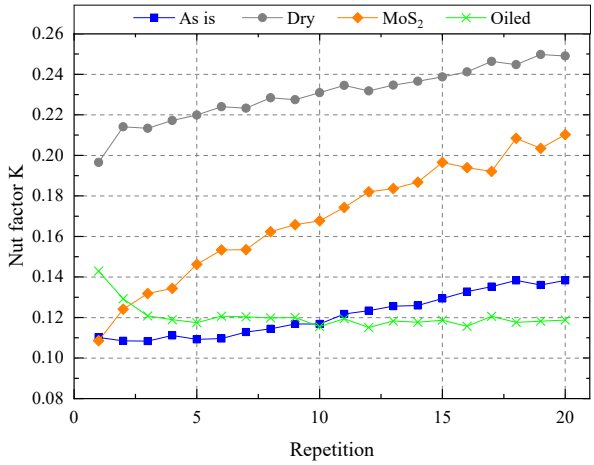


Figure 11
Calculated nut factor and its variation during repeated tightening

In the boxplot of Figure 12, the dispersion of the calculated nut factor is presented for all lubrication conditions throughout the twenty tightening cycles. It can be seen that oiling is good for keeping the nut factor dispersion small. Minimal nut factor values are present in the case of the oiled, followed by the As-is case due to the protection oil film. The wide K range for the MoS_2, despite the high initial achieved preload, can also be seen.

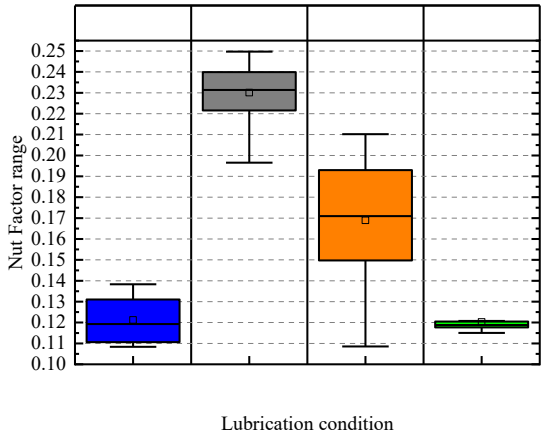


Figure 12
Nut factors range throughout the tightening repetitions

Conclusions

This study examined the variation of preload force, in a bolted joint, under repetitive tightening cycles. Initially, the preload force variation was investigated under different lubrication conditions. The nut factor is a measure of the influence of

various unknown factors on the relationship between the input torque and the output preload. This factor was utilized to compare the behavior of the bolt preload under different lubrication conditions. Subsequently, the nut factor variations were computed based on the measurements obtained. The results indicate that the preload force exhibits variations during successive tightening cycles, with a notable decrease.

The study's results validate the engineering practice of not reusing dismantled bolts, washers, and nuts, as the initial preload of new bolts and nuts was the highest in most lubrication cases.

The influence of lubrication on preload variation was observed, with a small amount (few drops) of oil lubricant leading to a stable repeated preload value. Application of a powder lubricant (MoS_2) resulted in higher preload force but also a wider distribution under repeated tightening. Similar preload behavior was observed between the as-is and oiled conditions, which can be attributed to using an oil film during manufacturing to prevent surface rust during storage.

References

[1] Rétfalvi, A.: Fixture design system with automatic generation and modification of complementary elements for modular fixtures. *Acta Polytechnica Hungarica*, **12** (7), 2015, pp. 163-182

[2] Rétfalvi, A., Stampfer, M.: The key steps toward automation of the fixture planning and design. *Acta Polytechnica Hungarica*, **10** (6), 2013, pp. 77-98

[3] Bíró, I., Fekete, G.: Approximate method for determining the axis of finite rotation of human knee joint. *Acta Polytechnica Hungarica*, **11** (9), 2014, pp. 61-74

[4] Jovanović, V. D. et al.: Determination of the load acting on the axial bearing of a slewing platform drive in hydraulic excavators. *Acta Polytechnica Hungarica*, **12** (1), 2015, pp. 5-22

[5] Martínez-parrales, R., Téllez-, A. C.: Vibration-based Fault Detection System with IoT Capabilities for a Conveyor Machine. **19** (9), 2022, pp. 7-24

[6] Eccles, W. et al.: Frictional changes during repeated tightening of zinc plated threaded fasteners. *Tribol Int*, **43** (4), 2010, pp. 700-707

[7] Bickford, J. H., Oliver, M.: *Introduction to the Design and Behavior of Bolted Joints*. Boca Raton: CRC Press, 2022

[8] Shoberg, R. S.: Engineering Fundamentals of Threaded Fastener Design and Analysis. *PCB Load & Torque, Inc.*, 2000, pp. 1-39

[9] Drumheller, J.: Fundamentals of Torque-Tension and Coefficient of Friction Testing. *White paper 21*, 2018

[10] Morgan, R. C., Henshall, J. L.: The torque-tension behaviour of 22 × 1.5 mm bolts for fixing spigot-located wheels on heavy commercial vehicles. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, **210** (3), 1996, pp. 209-214

[11] Nassar, S. A. et al.: Effect of tightening speed on the torque-tension and wear pattern in bolted connections. *Journal of Pressure Vessel Technology, Transactions of the ASME*, **129** (3), 2007, pp. 426-440

[12] Liu, Z. et al.: Changing behavior of friction coefficient for high strength bolts during repeated tightening. *Tribol Int*, **151** (June), 2020, p. 106486

[13] Grabon, W.A. et al.: Friction of threaded fasteners. *Tribol Int*, **118** (October 2017), 2018, p. 408-420

[14] Güler, B., Gürsel, K.T.: Experimental analysis of the friction coefficient effect of the zinc-lamella coated fasteners on bolt preload and tightening moment. *Materwiss Werksttech*, **50** (6), 2019, pp. 696-705

[15] Motosh, N.: Development of design charts for bolts preloaded up to the plastic range. *Journal of Manufacturing Science and Engineering, Transactions of the ASME*, **98** (3), 1976, pp. 849-851

[16] Nassar, S. A., Xianjie, Y.: Novel formulation of the tightening and breakaway torque components in threaded fasteners. *Journal of Pressure Vessel Technology, Transactions of the ASME*, **129** (4), 2007, pp. 653-663

[17] Nassar, S. A. et al.: The effect of coating and tightening speed on the torque-tension relationship in threaded fasteners. *SAE Technical Papers*, (724) 2006

[18] Shoberg, R. S.: Engineering Fundamentals of Threaded Fastener Design and Analysis. *RS Technologies*, 2000, pp. 1-39

[19] Hamilton, S.: *Bolt Lubricant and Torque: A Comprehensive Guide*. 2021

[20] Oberg, E. et al.: *Machinery's Handbook :A reference book for the Mechanical Engineer, Designer, Drafter, Metalworker, Toolmaker, Machinist, Hobbyist, Educator, and Student*. 2020

[21] Sakai, T.: The Friction Coefficient of Fasteners. *Bulletin of JSME*, **21** (152), 1978, pp. 333-340

[22] Jiang, Y. et al.: An Experimental Investigation on Frictional Properties of Bolted Joints. *American Society of Mechanical Engineers, Pressure Vessels and Piping Division (Publication) PVP*, **433**, 2008, pp. 59-66

[23] Zou, Q. et al.: Effect of lubrication on friction and torque-tension relationship in threaded fasteners. *Tribology Transactions*, **50** (1), 2007, pp. 127-136

[24] Franklin D Jones, Henry H Ryffel, Erik Oberg, Christopher J McCauley, R. M. H.: *Machinery's Handbook 27th Edition*. New York, NY: Industrial Press, Inc., 2004

[25] Lee, Y.-L., Ho, H.-C.: *Chapter 12 - Design and Analysis of Metric Bolted Joints: VDI Guideline and Finite Element Analysis*. In: Metal Fatigue Analysis Handbook (Editors: Y.-L. Lee et al.) Boston: Butterworth-Heinemann, 2012, pp. 461-513

[26] Kassambara, A.: *Practical Statistics in R II - Comparing Groups: Numerical Variables*. Datanovia, 2019

The Project and Risk Management Challenges of Start-ups

Tamás Bence Venczel[1], László Berényi[2] and Krisztián Hriczó[3]

[1]Institute of Mathematics, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Miskolc-Egyetemváros, 3515 Miskolc, Hungary, bence.venczel.tamas@uni-miskolc.hu

[2]Institute of Management Science, Faculty of Economics, University of Miskolc, Miskolc-Egyetemváros, 3515 Miskolc, Hungary, laszlo.berenyi@uni-miskolc.hu

[3]Institute of Mathematics, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Miskolc-Egyetemváros, 3515 Miskolc, Hungary, krisztian.hriczo@uni-miskolc.hu

Abstract: Start-up companies are essential to maintaining innovation in an economy. However, the high failure ratio of start-ups indicates that market, financial, and other risks require serious attention. As start-ups mostly evolved at the end of the 20th Century or the beginning of the 21st Century, the history of project and risk management practices for them has a shorter history. Overall, 90% of start-ups fail, 10% within the first year, and 70% within two and five years after foundation; therefore, understanding the underlying factors and how proper project and risk management can reduce the likelihood of failure, is worthwhile. This paper reviews the history of start-ups and the typical causes of failure based on a literature review. Finding the appropriate way and tools for risk management, a new approach is introduced. Considering start-ups as projects, a much more mature methodology is available for solving the problems. As a result of the diversity resulting from industry and other specificities, a two-level approach is suggested, including a risk-oriented management framework model and an additional flexible toolset.

Keywords: start-up; project management; risk management

1 Introduction

Due to the result of the latest technical improvements of the last century, there have been numerous innovations across the globe regarding new product and service developments. As start-up companies have a remarkable contribution to innovation processes regardless of the industry, the research interest is increasing in the field [1]. Start-up companies can generate relevant economic and social impact, but they

usually take higher risks to achieve success than other companies. Therefore, start-ups can be investigated as entities with a goal to generate profit as well as incubators of innovation, regardless of business outcome [2]. The competing goals of creating profit and maintaining a high level of creation must be achieved simultaneously. Global policy initiatives have emerged to fulfill this goal in the past few years to support the environment of new companies with local regulations and reduce the risks over the life cycle. Policymakers have also acknowledged the importance of start-ups in the economy, especially after the global recovery from the effects of the COVID-19 pandemic [3].

The highest number of start-ups are operating in the technological sector, especially in the Fintech industry (total 35.9%) [4]. The United States had the highest number of start-ups registered at the end of 2021 (70641), which is way above the second, India (12440) [5]. Despite the high number of start-ups, their success rate is relatively low. A study based on the analysis of 80 start-ups in March of 2021 found that only 10% of start-ups will make it through their first year. The most frequent reason for failure is the lack of product-market fit or market need (34%), and the second is the lack of funds (29%). Even if they survive the first year, only 40% of the companies will become profitable. There was no significant difference found between industries; the highest failure rate is within IT (63%), the lowest is within finance, insurance, and real estate (42%), and manufacturing represents the average (51%). Notably, about half of the start-up owners expect an acquisition by a larger corporation [6]. Regarding the invested amount of money, the most spectacularly growing start-up sectors are healthcare (41.2 billion USD), transportation (25.5 billion USD), and financial services (24.6 billion USD) [7].

Santisteban et al. [8] emphasize that the success of a start-up is influenced by applied project management and risk management strategies. A broad range of project management methods evolved in the 20th Century, mainly in parallel with the development of computer technologies [9] and continue to advance as organizations recognize the importance of conscious project management. However, these project management methods and frameworks were designed for large companies with a mature management system, and the question arises whether the practices are also applicable to start-ups or might require different approaches.

Risks are usually considered obstacles with a wide range of probabilities to achieve success. Due to worldwide globalization, several possibilities emerged for organizations, but in parallel, the number of risks to handle increased too [10]. The most effective approach to risk management usually considers economic aspects, as the main target of private companies is to achieve targeted profitability, which catalyzes the development of project risk management principles [11]. It later considers their applicability for start-ups [12].

2 History of Start-ups and their Research

A Stanford University professor and entrepreneur, Steve Blank, defined a start-up as a "temporary organization that aims to pursue a repeatable and scalable business model" [13]. The definition is quite similar to that of small and medium-sized enterprises (SMEs), but there is a relevant difference between them in the innovative approach. Start-ups are highly innovative companies and have become increasingly popular since technological inventions make innovation procedures easier and faster in basically every industry.

This kind of entrepreneurship can be considered a business model that can adapt to the rapidly changing environment with constant re-iteration to reach the target and create value. Also, the level of competition is significant between start-ups since the economic race is obviously increasing in line with the number of actors within an industry. Digitization, the internet with simple and quick access to knowledge, and global supply chain improvements all support the birth of new start-ups worldwide.

Start-ups are often linked with the rise of Silicon Valley, where most of the innovative technology, mainly semiconductor manufacturing, companies concentrated in this area in the 1970s. After this, a huge "boom" started in the 1990s with the development of Internet companies, which is considered the second phase of start-up history. Later, the technological improvements provided a base for other industries to get leverage from the latest developments. Nowadays, start-ups are among several technologies and are a worldwide trend, and no longer exclusively in the United States.

The first definition of the modern start-up was published by Forbes [14] in 1976, and afterward, in Business Week [15], the term start-up company was defined. In the 1980s, Van de Ven [16] analyzed the management framework of start-ups, followed by Dean [17] to discover the project management aspects of start-ups. Recently, a detailed classification of different start-up types was collected by Krishnan et al. [12] in 2020.

Finkelstein [18] targeted the general risks related to start-ups. Chang [19] in 2004 and Konecsny [20] in 2018 explained the applicability of one of the most frequently used risk management methods, venture capital financing, to mitigate the financial risks of start-ups. A similar approach was followed by Midler [21] in 2008; he investigated the importance of continuous learning. Blank and Dorf [13], Trimi et al. [22], Erzurumlu et al. [23], and Picken [24] all investigated the role of business models in start-ups. Several authors analyzed the risk management practices of start-ups like Erzurumlu et al. [23], Jaroslaw [25], and Halmosi [26]. Mantilla [27] and Santisteban et al. [8] called attention to the difficulties and success factors of start-ups.

Start-ups differ from other companies that use traditional business planning strategies [13] because their future predictions cannot be made based on past experience since a comprehensive operations database is not available [28].

Therefore, a key point to running a successful start-up business is managing knowledge to build lessons learned into the next loop of strategic planning and initiate a quickly adaptable system for fast changes. They improve through continuous changes and building a business model, for the actual situation [24] [29-31]. Teece stated that the success of an organization is highly dependent on its ability to adapt the business model dynamically [32].

All these contributing factors to the high failure rate of start-ups [19] [33] can be considered risks, and some of these can be traced back to internal managerial issues. Trimi and Berbegal-Mirabent [22] highlight that a major cause of start-up failure is the lack of a structured process to understand their markets better and validate theories in the early stages of the company. Learning about the related risks and explaining their reasons is critical to increasing the success rate. This article aims to map the frequent risks of start-ups to better understand current obstacles and explore options for implementing project management practices to mitigate risks effectively.

3 Risks of Start-ups

The literature on corporate risk management is broad. Competing information is available regarding the definition of risks, risk management, and its relation to project management [34]. Studies on risk management emerged in the 1980s, along with the appreciation of business strategies and project management. The motivation for developing risk management was that many projects were completed late, over budget, or did not perform as expected. A database from the 1980s showed that "many projects met their time-target – the average slippage was 17% – but there was a clear over-run-on cost – the average over-spend was 88%". Williams gave a detailed bibliography of the topic [35].

According to Giardino et al. [33], about cutting-edge technologies, just one failed project can destroy the start-up's future. Case studies show how inconsistent management strategy and execution lead to failure [36]. It is important to understand the importance of a fast and effective learning procedure, especially regarding the market, which requires information. The study reveals "inconsistency between the strategy of understanding and testing the problem/solution fit and the behavioral execution of pursuing the product/market fit." Early recognition and solution of problems lead to higher chances of start-up success. The analyzed, failed start-ups, showed a reluctant behavior to reflect customer needs appropriately [33].

Based on empirical investigation among young companies in their formative age (2-8 years old, across 10 EU countries and 18 sectors), some similarities can be seen in the risk management of these companies. Financial risks can be managed with the support of formal and informal networks. Market risks are usually not well

manageable by these companies. Firms in knowledge-intensive sectors (high-tech manufacturing) and companies with more formally educated leaders apply risk management more consciously. Technology and financial risks are positively related to internal risk mitigation and networking. Operational risk is positively related to internal risk mitigation but negatively to networking. Market risk is exactly the opposite of operational risk. The education of founders and new product introduction are positively related to all aspects of risk mitigation. Short life cycles are strongly related to market risk mitigation strategies across all sectors. Networking and technology risk management show a correlation in low-tech sectors. At the same time, it was found that the founders' previous employment was unrelated to risk mitigation activities [37].

Table 1
Summary of start-up risk categories

Source	Risk or reason of failure categorized
Giardino et al. [33]	Lack of Problem/Solution fit
Giardino et al. [33]	Neglected Learning Process
Janaji et al. [38]	Lack of fund
Cantamessa et al. [36]	Business model (e.g., no/wrong business model, product/market)
Cantamessa et al. [36]	Product (e.g., not feasible, bad quality)
Cantamessa et al. [36]	Environment (e.g., competitors, lack of funds)
Cantamessa et al. [36]	Customer/user (e.g., few customers)
Cantamessa et al. [36]	Organization (e.g., wrong leadership, wrong scaling)
Kim et al. [39]	Commercialization
Pisoni et al. [40]	Human capital
Pisoni et al. [40]	Financial resources
Pisoni et al. [40]	Strategic/managerial decisions
Pisoni et al. [40]	Product/service-related aspects
Pisoni et al. [40]	Contextual/environmental-related aspects

There are multiple approaches to categorizing the most common risks of start-ups (Table 1). The SHELL model (Figure 1) developed by Cantamessa et al. [36] in 2018 is a robust framework to provide a structural method to analyze possible risks of start-ups. The conclusion was that the top three reasons for start-up failures are "No/Wrong Business Model" (35%), "Lack of business development" (28%), and "Run out of cash" (21%).

A Brazilian study in 2017 regarding risk management behavior aimed to analyze similarities in the risk management of companies through correlation analysis. The main finding was that there is no unique way; the start-ups look at different ways of risk management. Their approach to risk management does not depend on the operation time and amount of investment, but a start-up with a more developed strategic framework has a better risk management process.

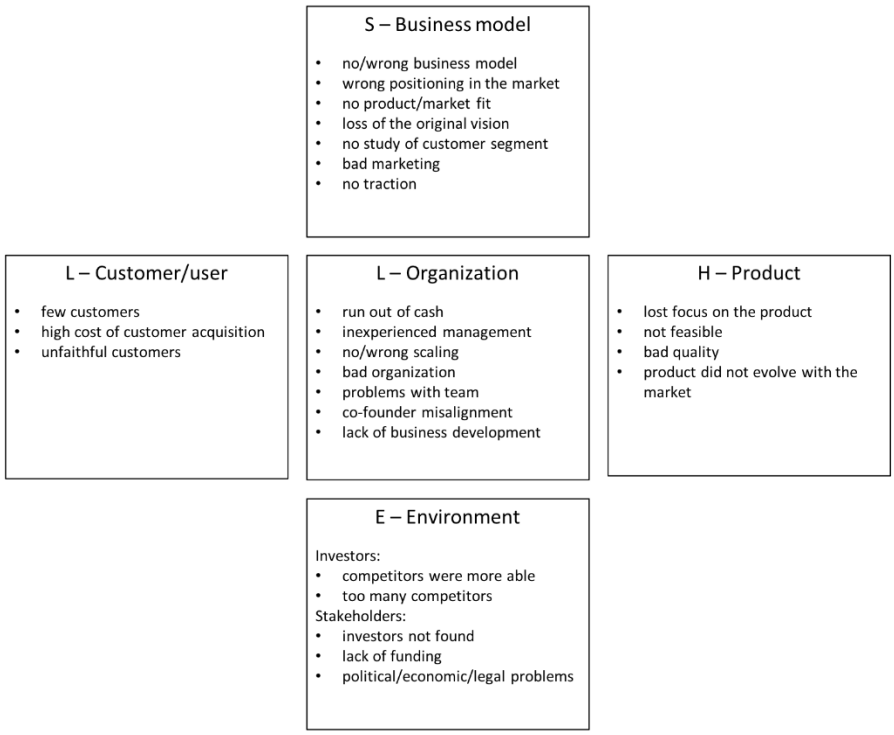


Figure 1
SHELL model based on [36]

Notably, managers who participated in the survey were interested in improving their risk management practices. The available risk management methodologies were too extensive and robust for the usage of start-ups, which clearly asks for the development of targeted and simplified methods in line with the start-up characteristics. The recommendation of the author is the ISO 31000 procedure (Figure 2), which is a simple and easy-to-implement procedure for start-ups. Besides, some start-ups followed the methodologies of Lean Start-up [41] or SCRUM [42]. These methodologies offer an incorporated toolset for risk management along the iteration cycles. The appreciation of the agile approach to project management emphasizes customer involvement at all project stages, including an improved feedback system. It could indicate that the analyzed companies tried to behave as companies in their stable enterprise phase, not in their initial phase [43].

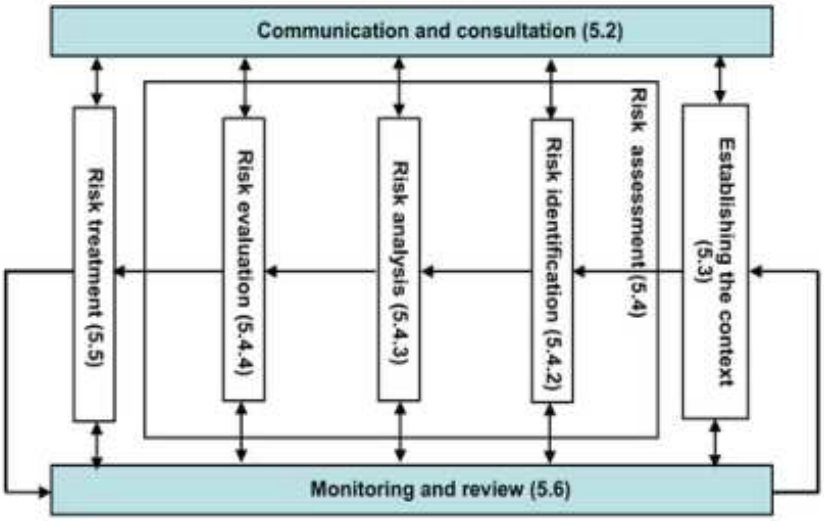


Figure 2
ISO 31000 risk management process [43]

Also, an exciting approach was introduced by Sanz-Prieto et al. [44] in 2021 called the Technical Due Diligence methodology. Due diligence is defined as a process that involves identifying and evaluating risks within a framework, including investments originated by commercialization activities, essentially the purchase and sale of companies, business units, and actions related to merger and corporate absorption mechanisms. Due diligence means rigorously investigating the possible operational risks and reducing them to the bare minimum expression. The methodology intends to perform a technical inspection of an asset, product, service, or process, including start-up ecosystems. The process is divided into phases (Kick-off call, Documentation review, Follow-up, and Report). The applicability of the method is restricted to start-up acquisition; therefore, it is a particular case among the available methods [44].

Filippetto et al. [45] offer a mathematical model developed among software development companies. Risk modeling requires analyzing historical data, then creating an algorithm to establish a tool for future risk prediction. They compose a computational model to reduce the probability of project failure based on the prediction of risks by using historical data (Figure 3). Since the method uses historical data, it is not applicable for starting companies, but accepting it as a framework, continuous data collection may support a quick introduction. Moreover, data management coordinated by an incubator organization may allow access to relevant information to a local or industrial community to boost the development of start-ups. The study considered 17 completed projects and considered 70% of their data to initiate a learning system to generate recommendations for future projects.

Additional 153 other projects from different companies were used as context histories. After the calculation by the algorithm, a comparison was made with an expert judgment regarding the predicted risks. The result showed a 73% acceptance rate by professionals and 83% accuracy compared to old projects. The model and study outline a possible future research field in risk management, as the development of artificial intelligence and big data could significantly support risk prediction models with a high amount of available data. Of course, historical data can only be obtained from the industry in the case of a new start-up, but the learning procedure can be more effective with this method [45].

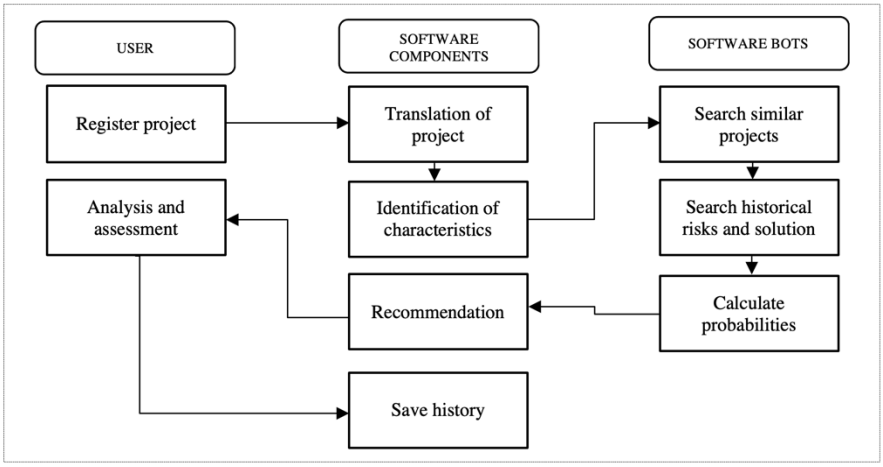


Figure 3
Risk recommendation flow in a project, based on Filippetto [45]

Ward [45] aimed to clarify the meaning of risk management and especially consider it rather a project uncertainty management than a management of purely "bad events". It seems the risk is usually considered an event that can negatively affect the project; however, approaching it as uncertainty could provide a better perspective, including opportunity management. The author argues that current risk management methodologies are not fulfilling their potential, as the perspective should also focus on opportunities beyond threats. Moreover, the event-based approach should be improved, as it can result in a lack of attention to several areas, like variability because of different knowledge levels or the basis of estimates. It is recommended to rename Project Risk Management to Project Uncertainty Management to move the focus toward the new approach. The author recommends applying this management approach earlier in the project life cycle [46].

SMEs (small and medium-sized enterprises) have some similarities with start-ups; in some cases, they cannot be easily separated. In 2014, Brustbauer analyzed the risk management practices of SMEs based on a questionnaire. He suggests that companies should apply a passive (defensive strategy) or active (offensive strategy)

risk management method. The chosen method should be based mainly on company size, sector affiliation, and ownership structure. Risk management is a significant issue for SMEs, mainly because of the lack of resources for this activity, and about two-thirds of the analyzed companies have a passive risk management approach. Also, larger companies have a greater affinity for implementing risk management strategies. The author interprets that applying risk management increases competitiveness and success. A key factor for effective risk management is the awareness of the company regarding possible risks. If a company is not ready to define the risk in itself and its surroundings, it is not possible to create an effective action plan for risk mitigation [47].

It is essential to emphasize the development of start-up policies, which can also be considered a risk reduction approach for start-ups. As governments recognized the appreciation of start-ups in social and economic aspects, they started to create policies to support the growth of start-ups and secure their economic environment to increase the probability of their success. However, Mason [48] also raises the question of "huge internal inequalities" based on Silicon Valley studies, which could be reconsidered in further studies.

4 Project Management Considerations for Risk Mitigation

Some start-up companies try to apply traditional project management methods [49], but these might not be suitable for start-ups. The maturity of the management, the immature structure, and the level of accumulated experience require a different approach. On the one hand, risks are not selective according to maturity, financial, market, and operational issues; these are the same for start-ups and other companies. On the other hand, a less developed organization is also a source of risk. Considering start-ups as projects can open new opportunities to build a toolset that supports risk mitigation, among other purposes.

Mantilla [27] performed qualitative research on how different start-ups implement project management methodologies. According to Santisteban and Mauricio [50], 21% of start-ups last more than five years. The study revealed that 40% of companies used Agile methods (like Kanban, Lean Start-up, Trello), 30% used traditional methods (e.g., WBS, PERT, and GANTT), and 30% only planned to use any project management methodology in the future. Four out of ten start-ups used Microsoft or Google Office products, and five used online products supporting project management and communication (Asana, Jira, or Trello). The author assumed traditional PM methods are harder to implement in start-ups, and these companies naturally tend towards agile methods. [27].

One of the first articles dealing with start-up project management is related to Dean (1986). He shows the "principal results obtained by applying the project management approach to strategic planning and operations management of innovative start-up firms' key activities" [17]. The approach implements entrepreneurship as a systematic principle and suggests considering innovation as one of the systematic principles. During the birth of a start-up company, several activities should be performed in an uncertain environment and with limited resources. He concluded that "without a centralized, cohesive, and logical systems approach, the entire start-up operation can quickly become a hopeless tangle of unrelated jobs" [17]. The study found that the lack of a generally effective business and project management plan and arbitrary decision-making, based on feelings and intuition rather than strategic planning leads to unpredictable outcomes. It is worth considering a start-up company as a project since the toolset applied (task definition, precedence relations, durations, milestones, throughput time planning) is common. Project management tools supply relevant initial point tools because the body of knowledge about them is extensive, including case studies and covering concepts or standards. These can moderate the risk of missing experience at start-ups. Dean proposed a basic project management approach for an innovative start-up, as summarized in Figure 4.

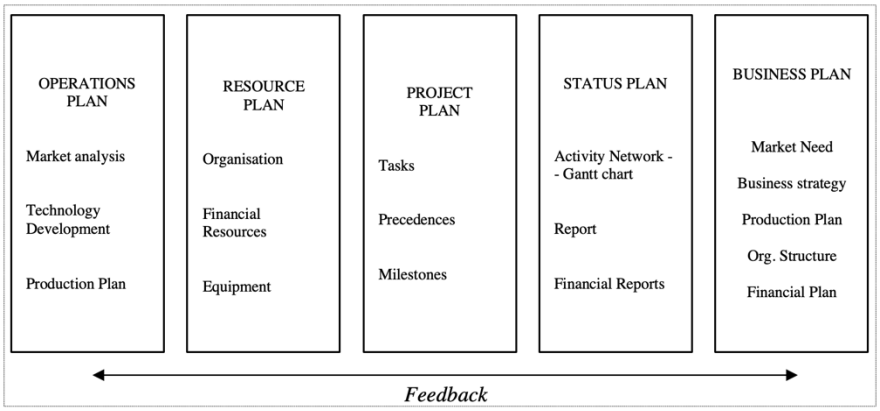


Figure 4
Project management approach to managing the innovative start-up firm (based on Dean) [17]

Midler and Silberzahn [21] present cases about the importance of learning through the product development process of start-ups. Learning efficiency was found to be a critical factor in this context. They analyzed three theoretical aspects, project management, organizational learning, and entrepreneurship. The study concluded that exploration and learning are key with these start-ups, as the cumulative learning method seems more successful [21]. Another study [51] analyzed how companies handle feature innovation on a strategic level. The ability of a company to successfully deploy feature innovations is a critical capability that allows car

manufacturers to be competitive in the market. The study investigated nine general car manufacturers and 26 feature innovation cases, showing a "clear trend towards the structure of autonomous "advanced engineering" units and processes responsible for exploring innovative features and transferring them to multiple products". They found that automotive companies separated the product development process from the innovation development process by called "Advanced Engineering" departments. The study proves that the innovation implementation process has a "direct impact on the competencies and routines of the carmaker" and, therefore, it might be a major driver for dynamic capability [51].

The increasing attention paid to sustainability also offers lessons learned. Projects in this field face high-risk situations and require specialized know-how. A study [23] analyzed 207 clean technology projects in the US to compare how operation design affects risk and enhances project valuation. A positive correlation has been confirmed between deployment feasibility and project valuation.

Yudine [52] proposed a four-dimensional thinking methodology to develop start-up projects. The method contains three stages of development and five phases of milestones. The study does not provide evidence of the applicability of the proposed method based on empirical data.

1st stage: "Chaotic" thinking broadwise on a two-dimensional plane of interdisciplinary links

2nd stage: The thinking in the time-dimension

3rd stage: Thinking in the vertical direction

1st phase: Developing an idea for the startup-project

2nd phase: The Business-plan of the startup-project

3rd phase: The search for the financial resource

4th phase: The implementation of the project

5th phase: The assessment of the startup-project efficiency

The model proposes to simplify the business procedure for a new start-up as the informational overload can affect managerial judgment and, consequently, the efficiency of the company. Phases of the method have been detailed but kept to a simple level to make it "user-friendly". The author recommends for future studies the implementation and effectiveness checks of the proposed method [52].

Conclusions

Start-ups are the beating heart of economic growth, through sustaining the dynamics of new products and novel solutions. The high failure ratio among start-ups is a call to address targeted actions. The fact that the failure ratio has been high for a long time indicates a lack of effectiveness in the proposed solutions, and a fundamentally new approach is needed to handle the risks. Handling a start-up as a project allows

may be an initial step. A start-up is not a project, but projects have a decisive role in their life cycle, project management tools may have a practical extension to their organizational support in the field of risk management.

Start-ups are not mature and not experienced organizations that can introduce complex management systems, covering project or risk management. Some try to use traditional models; others focus on agile practices, depending on the knowledge level and the industrial specifications. The high failure ratio also suggests that no common practice is available, and the case studies around lessons learned, are appreciated in finding unique solutions. The analysis of causes for failure, emphasize a misalignment between products or services and market demand, the incompetence to improve through quick iteration and the lack of structured business management processes.

A similar characteristic of a start-up and a project is embodied in project-based organizations. Implementing conscious risk management practices into their business management system is advised as part of the project management practices. The most frequent risks are related to financial, market and continuous learning implementation. Therefore, these items should be handled separately in the project phase planning activities. There are also global policies and standards initiated in the past few years for start-ups that can reduce risks.

Further research aims to explore industrial differences among start-ups and seeks common characteristics, if they exist. In line with an agile environment, developing a two-level method is recommended, including a principle-fold, risk management framework model and a flexible toolset, as a supplement.

References

[1] J.-C. Spender, V. Corvello, M. Grimaldi, and P. Rippa, 'Startups and open innovation: a review of the literature', *EJIM*, Vol. 20, No. 1, pp. 4-30, Jan. 2017, doi: 10.1108/EJIM-12-2015-0131

[2] D. Audretsch, A. Colombelli, L. Grilli, T. Minola, and E. Rasmussen, 'Innovative start-ups and policy initiatives', *Research Policy*, Vol. 49, No. 10, p. 104027, Dec. 2020, doi: 10.1016/j.respol.2020.104027

[3] A. Kuckertz *et al.*, 'Startups in times of crisis – A rapid response to the COVID-19 pandemic', *Journal of Business Venturing Insights*, Vol. 13, p. e00169, Jun. 2020, doi: 10.1016/j.jbvi.2020.e00169

[4] 'Distribution of global startups by industry', *Statista*. https://www.statista.com/statistics/882615/startups-worldwide-by-industry/ (accessed Jan. 02, 2022)

[5] 'Countries - With the top startups worldwide | Startup Ranking', *StartupRanking*. https://www.startupranking.com/countries (accessed Jan. 02, 2022)

[6] 'Startup Failure Rate: Ultimate Report + Infographic [2021]' https://www.failory.com/blog/startup-failure-rate (accessed Jan. 02, 2022)

[7] 'Top 5 fastest-growing Industries of 2019 by money invested', *YouTeam*, Jan. 14, 2020, https://youteam.io/blog/top-fastest-growing-industries-2019/ (accessed Jan. 02, 2022)

[8] J. Santisteban, D. Mauricio, and O. Cachay, 'Critical success factors for technology-based startups', p. 25, 2021

[9] T. Seymour and S. Hussein, 'The History Of Project Management', *IJMIS*, Vol. 18, No. 4, p. 233, Sep. 2014, doi: 10.19030/ijmis.v18i4.8820

[10] F. L. Oliva *et al.*, 'Risks and critical success factors in the internationalization of born global startups of industry 4.0: A social, environmental, economic, and institutional analysis', *Technological Forecasting and Social Change*, p. 121346, Nov. 2021, doi: 10.1016/j.techfore.2021.121346

[11] M. J. Pennock and Y. Y. Haimes, 'Principles and guidelines for project risk management', *Syst. Engin.*, Vol. 5, No. 2, pp. 89-108, 2002, doi: 10.1002/sys.10009

[12] S. N. Krishnan, L. S. Ganesh, and C. Rajendran, 'Characterizing and Distinguishing "Innovative Start-ups" Among Micro, Small and Medium Enterprises (MSME)', *Journal of New Business Ventures*, Vol. 1, No. 1-2, pp. 125-156, Jun. 2020, doi: 10.1177/2632962X20964418

[13] S. G. Blank and B. Dorf, *The startup owner's manual: the step-by-step guide for building a great company*. Pescadero, Calif.: K & S Ranch, 2012

[14] 'The unfashionable business of investing in startups in the electronic data processing field.', *Forbes*, Aug. 15, 1976

[15] 'An incubator for startup companies, especially in the fast-growth, high-technology fields.', *Business Week*, Sep. 05, 1977

[16] A. H. Van de Ven, R. Hudson, and D. M. Schroeder, 'Designing New Business Startups: Entrepreneurial, Organizational, and Ecological Considerations', *Journal of Management*, Vol. 10, No. 1, pp. 87-108, Apr. 1984, doi: 10.1177/014920638401000108

[17] B. V. Dean, 'The project-management approach in the "systematic management" of innovative start-up firms', *Journal of Business Venturing*, Vol. 1, No. 2, pp. 149-160, Mar. 1986, doi: 10.1016/0883-9026(86)90011-X

[18] S. Finkelstein, 'Internet startups: so why can't they win?', *Journal of Business Strategy*, Vol. 22, No. 4, pp. 16-21, Apr. 2001, doi: 10.1108/eb040180

[19] S. J. Chang, 'Venture capital financing, strategic alliances, and the initial public offerings of Internet startups', *Journal of Business Venturing*, Vol. 19, No. 5, pp. 721-741, Sep. 2004, doi: 10.1016/j.jbusvent.2003.03.002

[20] Konecsny J., 'Decision-making processes and project evaluation criteria for venture capital funds in Hungary', 2018, doi: 10.14751/SZIE.2018.017

[21] C. Midler and P. Silberzahn, 'Managing robust development process for high-tech startups through multi-project learning: The case of two European start-ups', *International Journal of Project Management*, p. 8, 2008

[22] S. Trimi and J. Berbegal-Mirabent, 'Business model innovation in entrepreneurship', *Int Entrep Manag J*, Vol. 8, No. 4, pp. 449-465, Dec. 2012, doi: 10.1007/s11365-012-0234-3

[23] S. Erzurumlu, J. Davies, and N. Joglekar, 'Managing Transformational Start-Up Risks: Evidence from ARPA-E Program', *SSRN Journal*, 2012, doi: 10.2139/ssrn.2130288

[24] J. C. Picken, 'From startup to scalable enterprise: Laying the foundation', *Business Horizons*, Vol. 60, No. 5, pp. 587-595, Sep. 2017, doi: 10.1016/j.bushor.2017.05.002

[25] State Higher School of Technology and Economics in Jarosław (Jarosławm, Poland), R. Pukala, E. Sira, University of Presov in Presov (Presov, Slovakia), R. Vavrek, and University of Presov in Presov (Presov, Slovakia), 'Risk management and financing among Start-ups', *MMI*, No. 3, pp. 153-161, 2018, doi: 10.21272/mmi.2018.3-13

[26] P. Halmosi, 'The Interpretation of Industry 4.0 by Hungarian Technology-Oriented Startups', *Timisoara Journal of Economics and Business*, Vol. 12, No. 2, pp. 149-164, Dec. 2019, doi: 10.2478/tjeb-2019-0008

[27] I. Mantilla, 'The Difficulty With Introducing Project Management Techniques in Digital Startups', p. 34, 2020

[28] R. G. McGrath, 'Business Models: A Discovery Driven Approach', *Long Range Planning*, Vol. 43, No. 2-3, pp. 247-261, Apr. 2010, doi: 10.1016/j.lrp.2009.07.005

[29] G. Fisher, 'Effectuation, Causation, and Bricolage: A Behavioral Comparison of Emerging Theories in Entrepreneurship Research', *Entrepreneurship Theory and Practice*, Vol. 36, No. 5, pp. 1019-1051, Sep. 2012, doi: 10.1111/j.1540-6520.2012.00537.x

[30] T. Baker and R. E. Nelson, 'Creating Something from Nothing: Resource Construction through Entrepreneurial Bricolage', *Administrative Science Quarterly*, Vol. 50, No. 3, pp. 329-366, Sep. 2005, doi: 10.2189/asqu.2005.50.3.329

[31] S. D. Sarasvathy, 'Causation and Effectuation: Toward a Theoretical Shift from Economic Inevitability to Entrepreneurial Contingency', *AMR*, Vol. 26, No. 2, pp. 243-263, Apr. 2001, doi: 10.5465/amr.2001.4378020

[32] D. J. Teece, 'Business Models, Business Strategy and Innovation', *Long Range Planning*, Vol. 43, No. 2-3, pp. 172-194, Apr. 2010, doi: 10.1016/j.lrp.2009.07.003

[33] C. Giardino, X. Wang, and P. Abrahamsson, 'Why Early-Stage Software Startups Fail: A Behavioral Framework', in *Software Business. Towards Continuous Value Delivery*, C. Lassenius and K. Smolander, Eds., in Lecture Notes in Business Information Processing, Vol. 182. Cham: Springer International Publishing, 2014, pp. 27-41, doi: 10.1007/978-3-319-08738-2_3

[34] Fekete I., 'Integrated risk management in practice', *Veztud*, pp. 33-46, Jan. 2015, doi: 10.14267/VEZTUD.2015.01.03

[35] T. Williams, 'A classified bibliography of recent research relating to project risk management', *European Journal of Operational Research*, Vol. 85, pp. 18-38

[36] M. Cantamessa, V. Gatteschi, G. Perboli, and M. Rosano, 'Startups' Roads to Failure', *Sustainability*, Vol. 10, No. 7, p. 2346, Jul. 2018, doi: 10.3390/su10072346

[37] Y. Kim and N. S. Vonortas, 'Managing risk in the formative years: Evidence from young enterprises in Europe', *Technovation*, Vol. 34, No. 8, pp. 454-465, Aug. 2014, doi: 10.1016/j.technovation.2014.05.004

[38] S. A. Janaji, K. Ismail, and F. Ibrahim, 'Startups and Sources of Funding', Vol. 02, No. 08, 2021

[39] B. Kim, H. Kim, and Y. Jeon, 'Critical Success Factors of a Design Startup Business', *Sustainability*, Vol. 10, No. 9, p. 2981, Aug. 2018, doi: 10.3390/su10092981

[40] A. Pisoni, E. A. Aversa, and A. Onetti, 'The Role of Failure in the Entrepreneurial Process: A Systematic Literature Review', *IJBM*, Vol. 16, No. 1, p. 53, Dec. 2020, doi: 10.5539/ijbm.v16n1p53

[41] R. F. Bortolini, M. Nogueira Cortimiglia, A. de M. F. Danilevicz, and A. Ghezzi, 'Lean Startup: a comprehensive historical review', *MD*, Vol. 59, No. 8, pp. 1765-1783, Aug. 2021, doi: 10.1108/MD-07-2017-0663

[42] V. Mahnic and S. Drnovscek, 'Agile Software Project Management with Scrum', p. 7, 2005

[43] B. V. Todeschini, A. S. Boelter, J. S. D. Souza, and M. N. Cortimiglia, 'Risk Management from the Perspective of Startups', *European Journal of Applied Business Management*, Vol. 3, No. 3, pp. 40-54, 2017

[44] I. Sanz-Prieto, L. de-la-fuente-Valentín, and S. Ríos-Aguilar, 'Technical due diligence as a methodology for assessing risks in start-up ecosystems: An advanced approach', *Information Processing & Management*, Vol. 58, No. 5, p. 102617, Sep. 2021, doi: 10.1016/j.ipm.2021.102617

[45] A. S. Filippetto, R. Lima, and J. L. V. Barbosa, 'A risk prediction model for software project management based on similarity analysis of context histories', *Information and Software Technology*, Vol. 131, p. 106497, Mar. 2021, doi: 10.1016/j.infsof.2020.106497

[46] S. Ward and C. Chapman, 'Transforming project risk management into project uncertainty management', *International Journal of Project Management*, Vol. 21, No. 2, pp. 97-105, Feb. 2003, doi: 10.1016/S0263-7863(01)00080-1

[47] J. Brustbauer, 'Enterprise risk management in SMEs: Towards a structural model', *International Small Business Journal*, Vol. 34, No. 1, pp. 70-85, Feb. 2016, doi: 10.1177/0266242614542853

[48] C. Mason and D. R. Brown, 'Entrepreneurial Ecosystems and Growth-Oriented Entrepreneurship', presented at the OECD LEED Programme and the Dutch Ministry of Economic Affairs, The Hague, 2013, p. 38

[49] E. Pollman, 'Startup Governance', *SSRN Journal*, 2019, doi: 10.2139/ssrn.3352203

[50] J. Santisteban and D. Mauricio, 'Systematic literature review of critical success factors of information technology startups', Vol. 23, No. 2, p. 24, 2017

[51] R. Maniak, C. Midler, R. Beaume, and F. von Pechmann, 'Featuring Capability: How Carmakers Organize to Deploy Innovative Features across Products: Featuring Capability in the World Auto Industry', *J Prod Innov Manag*, Vol. 31, No. 1, pp. 114-127, Jan. 2014, doi: 10.1111/jpim.12083

[52] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" (Kyiv, Ukraine) and N. Yudina, 'Methods of the startup-project developing based on "the four-dimensional thinking" in information society', *MMI*, No. 3, pp. 245-256, 2017, doi: 10.21272/mmi.2017.3-23

Semantic Composition of Data Analytical Processes

Peter Bednár, Juliana Ivančáková, Martin Sarnovský

Technical University of Kosice, Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence
Letna 9, 042 00 Kosice, Slovakia
peter.bednar@tuke.sk, juliana.ivancakova@tuke.sk, matrin.sarnovsky@tuke.sk

Abstract: This paper presents the semantic framework for the description and automatic composition of the data analytical processes. The framework specifies how to describe goals, input data, outputs and various data operators for data pre-processing and modelling that can be applied to achieve the goals. The main contribution of this paper is the formal language for the specification of the preconditions, postconditions, inputs and outputs of the data operators. The formal description of the operators with the logical expressions allows automatic composition of operators into the complex workflows achieving the specified goals of the data analysis. The evaluation of the semantic framework was performed on the two real-world use cases from the medical domain, where the automatically generated workflow was compared with the implementation manually programmed by the data scientist.

Keywords: data science; data mining; semantic technologies; ontology

1 Introduction

Analysis of data and application of machine learning and artificial intelligence methods become very important approach successfully applied to the research or business problems in the various domains. However, the application of these methods is not always straightforward and requires extensive knowledge exchange between data scientists and domain experts. Additionally, the implementation of these applications is rather complex, with many constraints coming from the definition of the goal, properties of input data and constraints of the algorithms applied to generate the results. The result is that implementation of the data analytical approach can be time and resource-consuming. In this paper, we propose the semantic framework for automatization of the data-analytical processes based on the application of ontologies and logical inference. The proposed framework allows to automatically compose data-analytical workflow based on the semantic description of the goals.

The paper is organized as follows. The first chapter describes the current state-of-the-art and our motivation. In the following Chapter 3, we introduce our semantic model for data-analytical processes, followed by the main contribution of this paper: semantic description of the data analytical processes for the automatic composition of workflows and additional types of data operators covering various data and model visualization techniques. Last Chapter 4 then presents the evaluation of the proposed approach on the real application cases from the medical domain.

2 Semantic Description of Data Analytical Processes

A few semantics have been proposed to describe data-analytical process models formalized as ontologies [1-2]. One of the main proposals is the OntoDM ontology [3], which consists of 3 modules. The first module deals with the specification of input and output data and is based on the ISO standard for describing data types (atomic and composite). The main module characterizes the concepts that describe the data, the data analysis tasks (such as classification, clustering, etc.), the data mining algorithms, and the analysis outputs in the form of general data mining models. The third module uses concepts from the first two modules to formalize the different phases of the overall process according to the CRISP-DM methodology [4].

The other proposed ontologies extend the definition of some concepts introduced in OntoDM. DMOP ontology [5] deals mainly with the detailed description of the data mining algorithms, including their internal principle, e.g., a description of the numerical optimization method used, the error function, or the regularization of learning. DMOP covers 3 phases of CRISP-DM. DMWF ontology [6] defines data operators, which are applicable in data pre-processing, modelling, and evaluations. Operators are described similarly to services by definition of their inputs, outputs, assumptions, and effects. These input and output conditions are defined as logical expressions in the SWRL language as far as possible use of automatic derivation in the composition of workflows at data analysis.

One of the most general proposals that extend OntoDM is ontology Exposé [7], which extends the CRISP-DM phase of the data analysis process to the level of experiments to ensure, e.g., reusability of procedures in data analysis and reproducibility of results. In addition to conceptualization, it also provides a language for describing experiments based on XML [8], which allows you to publish and share a description of experiments on the web in a machine-readable format.

Panov et al. describe OntoDT, ontology to represent knowledge about data types. This ontology defines basic entities such as datatypes and their properties, specifications, characterizing operations, and datatype taxonomy [9].

In [10] Tianxing et al. presents a meta-mining ontology, which is used for building a domain-oriented ontology. The main goal of creating INPUT ontology is to understand data and business goals better and use it as an input interface for user queries.

The conceptualization of these semantic models is based on the description interfaces of existing software tools used for data analysis, such as R environment or sci-kit-learn library. Other relevant technologies can also include formats for exchanging data-analytical models when deploying them, such as XML standard PMML, PFA format based on JSON notation or currently the most supported ONNX format focused mainly on the exchange of models of deep learning.

Data Science Ontology (DSO) is a data science knowledge base focusing on computer programming. DSO is a way of organizing and classifying the concepts and entities within the field of data science. It helps define the relationships between different aspects of data science work and provides a framework for understanding and communicating about the field. One important aspect of data science ontology is the classification of data types and sources. This includes things like structured data, unstructured data, and semi-structured data, as well as data sources such as databases, APIs, and text files. Another important aspect is the classification of data analysis and modelling techniques. This can include things like statistical methods, machine learning algorithms [11], data visualization techniques, and natural language processing. The concepts for this ontology are gleaned from statistics, machine learning [29], and software engineering for data science. In addition to concepts, ontology also provides semantic annotations for data science. The annotations map the types and functions of the libraries to the universal concepts of ontology [12]. Data science ontology also includes the different roles and responsibilities within a data science team. This can include roles such as data engineer, data analyst, data scientist, and machine learning engineer, as well as the specific tasks and responsibilities associated with each role. Data science ontology can also include the models, frameworks and methodologies that are used in the field. These can include things like CRISP-DM for data mining, SEMMA for data mining and statistics, and the OSEMN framework for data science.

Data visualization and models should not be forgotten in the preprocessing and modelling framework; there is a VISO ontology [13] for describing such concepts, which formally models concepts and facts specific to visualizations. Visualization ontology is an important area of study, as it provides a structured and consistent way of understanding and discussing the field of data visualization. By understanding the different types of visualizations, design elements, and ways of representing data, we can create more effective and engaging visualizations that can help to communicate complex data in a more understandable way. The advantages of this ontology, also achieved thanks to well-established semantics standards such as RDFs [14] and OWL [14-16], are technical interoperability, support for a common understanding between interdisciplinary

parties in the visualization process, and the ability to derive new knowledge from existing facts. Viso is characterized by being composed of 7 modules: Graphic - formalizes concepts related to graphical relations and representations; Data - defines data variables and structures; Facts - formalizes constraints, rankings and defaults; Activity - deals with the human aspect within the visualization; System - this module covers HW and SW; User - characterizes user extensions and Domain - describes the domain specifications.

3 Semantic Framework for Automatization of Data-Analytical Processes

In our previous work [17-18], we have defined the semantic framework for the description of the data analytical processes, which is divided into the following modules:

- Domain Concepts - concepts for the description of entities and known relationships in the domain under investigation used for data analysis methods.
- Data Items and Performance Indicators - concepts for the description of key performance indicators formalising business and research requirements and goals of data analysis and concepts to describe input and output data attributes and data sets.
- Algorithms and Data Mining Models - concepts for describing methods and data mining algorithms and their settings, and concepts for describing data mining models (main outputs of data analysis).

The first module is mainly designed from the point of view of a domain expert, using domain concepts to formalise the description of a given domain. The second module contains concepts that are shared between the domain expert and the data analyst and is used to formally describe the goals of the analysis and the data. The last module was mainly proposed for the semantic documentation of the existing data-analytical processes to achieve interoperability and reproducibility of the processes. The main contribution of this paper is in the extension of the framework, with the concepts which will allow the automatic composition of the workflows and automation of the data-analytical processes. The following subchapters will describe the proposed modules in detail.

3.1 Domain Concepts

Concepts for domain formalisation are specified using the SKOS metamodel [19], which allows the specification of title, narrative description and definition of

concepts localised in several natural languages. Concepts can be arranged hierarchically in the form of a thesaurus/taxonomy by the relations skos#broader/skos#narrower. It is also possible to define polyhierarchical schemes. In addition to the hierarchical arrangement, terms can also be linked associatively by the relation skos#related.

A defined common dictionary of domain terms can also be used as a classification scheme for organising different types of documents that enter into data analysis as domain documentation, created by data analysts to document the analysis process itself, the data and the results achieved. The document types themselves can be specified as a classification taxonomy in SKOS.

In addition to the narrative description, in some domains, it is also appropriate to explain existing concepts using various diagrams, schemes and other types of graphical notations (e.g., using BPMN diagrams for modelling business processes or process diagrams for visualising production processes in the field of Industry 4.0, etc.). In this case, the individual graphical elements (e.g., an activity block in a BPMN diagram) are described as separate SKOS concepts that are linked to a given element to allow bi-directional navigation between the graphical notation and a set of semantic concepts. However, the proposed formalism does not define how these links are represented either in a semantic representation or inserted directly into the graphical notation format.

3.2 Data Elements and Performance Indicators

Figure 1 illustrates the basic concepts that represent the input and output data. Data elements are specified on two levels: logical and physical. Logical data elements are utilized to describe every input data attribute that the domain expert recognizes as relevant for resolving the given task of data analysis or for describing all output data produced during the analysis process, including prediction of the data-analytical models or values of domain and technical performance indicators.

The logical representation defines the metadata for each data attribute, which includes its name and definition in natural language, logical data type (for example, whether it is nominal, ordinal, numeric data, scalar, vector quantity, spatially-arranged data, time series, etc.), the commonly used physical unit of measurement, and the role of the data element in the data analysis process (such as whether it is input, output, or input-output data).

Logical data attributes can be connected by interdependencies, which are represented by the Dependency class. The Dependency class defines a relationship between one dependent data element and one or more independent elements. The dependency can be described in text or specified mathematically using a known physical or economic model. Dependencies can be further specified by

basic type, for example, expressing whether the value of a dependent element is derived from an independent element through transformation or is an aggregation of multiple independent elements, etc.

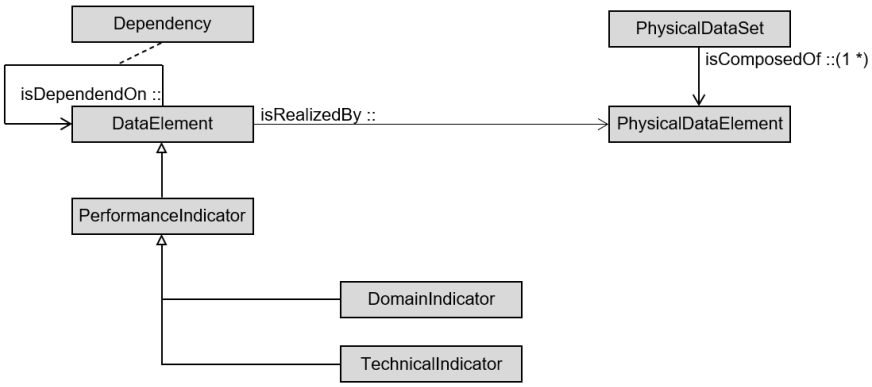


Figure 1
Data elements and performance indicators

The physical location of the data during analysis is represented by the physical data elements, and one logical attribute can have multiple physical realizations. Physical data elements are assigned physical data formats and types. The physical type is based on the ISO standard for describing data formats in software systems and can be atomic (real/integer, Boolean value, string) or composite (record with data fields, list, or unordered set). The specific location of the data is defined using IRI, which represents a unique identifier in a software environment for data analysis or a URL for placement on the web. Multiple physical data elements can be combined into a single data element set referenced through IRI.

The results of the data analysis are quantitatively described by the measurable performance indicators, which are defined as the special type of data elements. Similarly, to the description of input data, the performance indicators are divided into two subclasses: domain indicators and technical indicators. Domain indicators are commonly introduced by the domain expert for the evaluation of the results from the business perspective. They include indicators which specify, for example, financial costs/savings, energy or resource consumption, environmental impact, etc. The technical indicators include statistics expressing the performance of the data-mining model, such as accuracy, specificity, sensitivity, etc., estimated on the test or validation data set or using the cross-validation. In addition to performance metrics, technical indicators also cover the various metrics expressing the complexity or interpretability of data-mining models (e.g., the number of numerical parameters, the number of classifications or association rules, the number of clusters, etc.). Technical indicators are formally mapped by the data analyst to domain indicators by specifying the dependencies represented by the Dependency class, which allows to retrospectively determine how the

technical quality of data-mining models quantitatively affects the required quality of business goals.

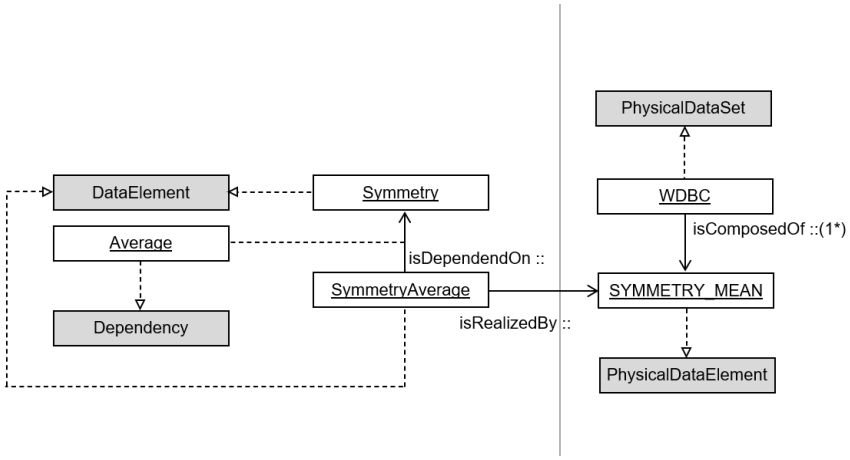


Figure 2
Example of the physical and logical data elements

Figure 2 presents an example which demonstrates the relations between logical and physical data elements. The physical dataset is represented with the WDBC instance, which points to the data stored in the comma-separated value file on the disk. The file consists of multiple columns, which are represented as instances of the Physical Data Element class. The example column is represented with the SYMMETRY_MEAN instance, which corresponds to the logical data element represented by the SymmetryAvrg instance. Another example of the logical element is Symmetry instance from which the SymmetryAvrg element is derived by arithmetic averaging. The dependency between the two logical elements is formally represented by the Average instance of type Dependency. In this case, the Average instance can specify directly by a structured formula for the computation of the arithmetic average over the source element.

3.3 Algorithms and Data-Analytical Models

The use of the algorithm in the analysis is determined by the definition of the data mining task as specified by the data analyst (Figure 3). The task is defined by a set of constraints (Constraint class) that use logical expressions to outline the desired properties of the resulting data mining model. These constraints cover the characteristics of the input data (such as data type attributes and the presence of missing or extreme values), the type of desired model (such as classification, regression, clustering, association rules, anomaly detection, etc.), and the quality and interpretability of the model by limiting technical performance indicators.

These constraints are further divided into hard constraints, which must be fully met in the solution of the task, and soft constraints, which should be taken into consideration in the solution but may not necessarily be met unconditionally.

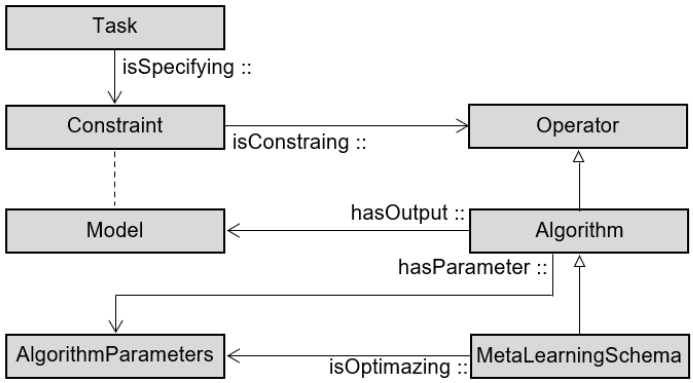


Figure 3
Data mining tasks, algorithms, and models

The main concept which covers all operations over the data is the Operator class. The machine learning algorithms are represented with the Algorithm concept, which is the special type of the operator with the data input and output in the form of the data mining model. The Model itself is the Operator, which can be applied to the input data for scoring. The output of the Model operator are data elements with predicted values and with the optional additional metadata such as confidence scores, identifiers of the classification rules, or weights of the contributing input values. In addition to the output model, an algorithm can have a set of (hyper) parameters that need to be set before it can be executed. Each parameter has a defined data type and a predetermined value. A special type of algorithm is the MetaLearningSchema class [20], which internally optimises parameter settings for a given training set and basic learning algorithm or selects the best algorithm from a group of algorithms for a given training data.

Figure 4 presents an example of the formalisation of the predictive machine learning algorithm and model. The RandomForest algorithm is the instance of the Algorithm concept, which is constrained for the classification task. The classification task specifies the constraint for the output predicted value (it must be of ordinal or nominal type). The algorithm is the data operator, which can be applied to input data represented as the physical dataset and which outputs the data mining model represented as the RFModel instance. An algorithm can have multiple parameters (settings which must be specified by the data scientist or automatically optimised), which are represented as the instances of type AlgorithmParameter. The figure shows a parameter for the number of trees included in the random forest model. Parameters can be semi-automatically optimised using the meta-learning schema. In the presented example, the number

of trees parameter is optimised by GridSearch meta-algorithm, which iteratively learns and evaluates model using the given algorithm (e.g., RandomForest) and selects the optimal value of the selected parameter.

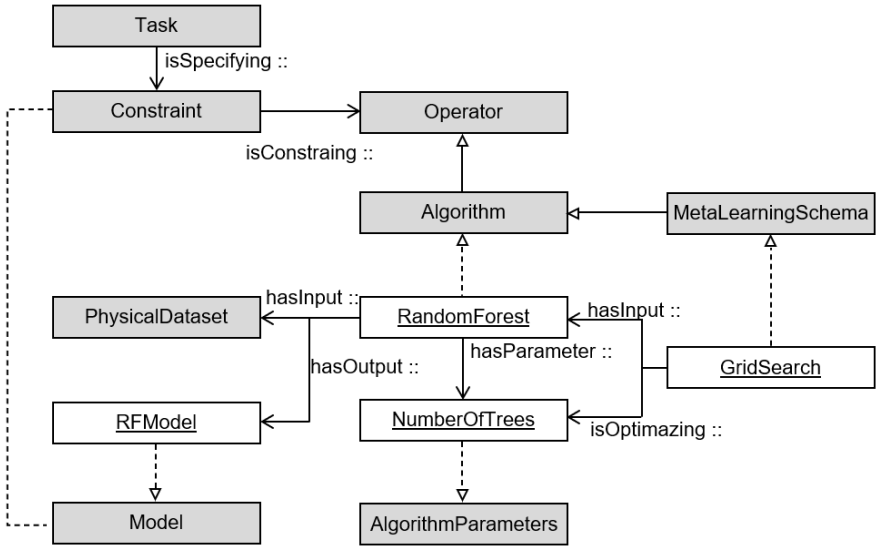


Figure 4
Example of the algorithms and data-analytical models

3.4 Process Model for Data Analytical Workflows

The proposed process model can be used to automatically generate workflows for data analysis tasks and also to formally describe existing data analysis scripts, ensuring their replicability and reusability. It is designed similarly to a process model for choreography and orchestration of web services, with the state represented by instances assigned to the shared variables. The process consists of nodes (Figure 5) that represent individual operations (Operator class) for data preprocessing, modelling, and evaluation, or control blocks such as branches, cycles, parallel execution, and synchronization (ControlNode class). The nodes are connected in a workflow by edges represented by the GuardedTransition class, which represents conditional transitions between nodes. Operators are described functionally, with inputs, outputs, assumptions, and effects defined as logical expressions. The flow of the process is determined by backward chaining of the effects and assumptions, taking the desired target effects from the task specification of the data mining. Some restrictions may be imposed by the algorithms used in the workflow, such as the ability to only work with certain types of attributes or to handle missing values.

An important part of our semantic framework is the formalism used for the description of the logical expressions in the specification of the operators inputs, outputs, preconditions and postconditions. The formalism is divided into variants with gradually increased expressiveness, which allows for choosing a trade-off between complexity and expressiveness and simplify the implementation of the automatic planning. The full specification was based on the Web Service Modelling Language (WSML) formalism [21], which combines constructs from descriptive logic and logic programming. WSML expressions consist of logical variables, functional symbols, logical operators (and, or, classical negation and negation as a failure) and quantifiers (existential and universal). Our current proposal substantially reduced the expressiveness of the WSML expressions in order to even further simplify formalism and implementation of the automatic method for the process composition.

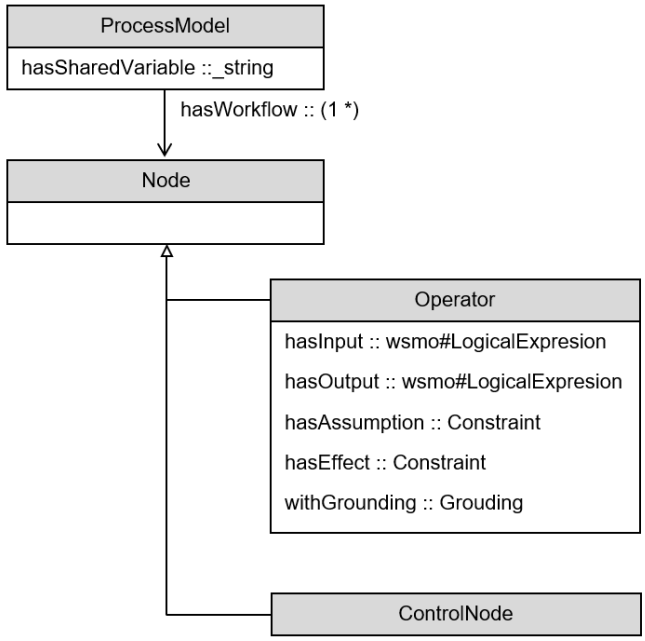


Figure 5
Process model for data analytical workflows

The current specification corresponds to the WSML light variant with the following constraints:

- All logical variables in the expressions are universally quantified.
- If α, β and γ are terms (identifiers, data values or variable symbols): α **subConceptOf** γ, α **memberOf** γ and $\alpha[\beta$ **hasValue** $\gamma]$ are atomic formulas where *sub-concept of* predicate specifies that the type α is the

sub-type of the γ, *member of* predicate constrains the type of the term α and *has value* predicate defines that the term α has value γ for the property β.

- Atomic formulas can be further combined with the logical operators **and**, **or** and negation as a failure (denoted as **not**).

Examples

The following example describes inputs, outputs, preconditions and postconditions for the operator, which replaces missing values of all numerical attributes in the input dataset.

Inputs:

?x **memberOf** PhysicalDataset

Preconditions:

?x[hasDataElement **hasValue** ?y] **and** ?y **memberOf** NumericalAttribute

Outputs:

?x **memberOf** PhysicalDataset

Postconditions:

not hasMissingValues(?y)

The operator defines the physical dataset as an input with the precondition that all data elements (columns in the input physical dataset) have a numerical type (note that all variables are implicitly universally quantified). The output of the operator is the same dataset with the postcondition that all attributes are without missing values (tested with the hasMissingValues predicate). All variables within the definition are shared between the inputs/precondition and outputs/postconditions, so for example, it is not necessary to bind variable ?y in the postconditions since it was already constrained in the preconditions.

The following example defines the operator for the logistic regression machine learning algorithm.

Inputs:

?x **memberOf** PhysicalDataset

Preconditions:

?x[hasDataElement **hasValue** ?y] **and**

?y **memberOf** NumericalAttribute **and**

not hasMissingValues(?y)

Outputs:

?z **memberOf** LogisticRegressionModel

The input to the operator is the physical dataset with numerical attributes without the missing values, and the output is the logistic regression machine learning model represented with the type LogisticRegressionModel (which is subsequently sub-concept of ClassificationModel, etc.).

Finally, the process of creating an executable workflow in a specific software environment involves anchoring the operators with the Grounding class. Grounding is a customizable text template that generates code for the operator's function in the environment when filled with variables. Operators can have multiple anchors, each designed for a different programming language (e.g. R or Python) or version of the software library used. The task specification can also include constraints on the anchoring, ensuring that only operators compatible with the given environment are used in the workflow.

3.5 Visualization Concepts

These concepts describe visualization in data mining processes. The key concepts are Algorithm, which is a type of operator that depends on input variables, and VisualizationMethod, which is an operator that takes data or a model as an input and outputs a visualization in the form of a graph.

Figure 6 illustrates the concepts of data visualization. As an example, we have selected the visualization of the PDP method for explaining machine learning models. In the beginning, it is important to define the type of visualization that is required. In principle, any visualization requires data as input. Either the data can be visualized as part of the data understanding or preprocessing phase, or the output of the visualization can be in the form of graphical representations of the models used for data mining. The concept of Algorithm is treated as a class operator, which is represented as a generic operation that transforms an input of a specific type into the desired output. The output of algorithms is then represented by the Model class. This concept is used together with data as input for the PDP (Partial Dependence Plot) [22], a graphical tool that helps understand the relationship between the input and the predicted variable. The outputs of the PDP operator are used to compute PDPStats, defined as TechnicalIndicator, which is a technical indicator specified based on domain indicators. The values obtained from PDPStats are inputs to the PDPVisualMethod. The final desired output is a PDP graph, which is a subclass of the VisualizationMethod operator and is specified in this case as a Graph.

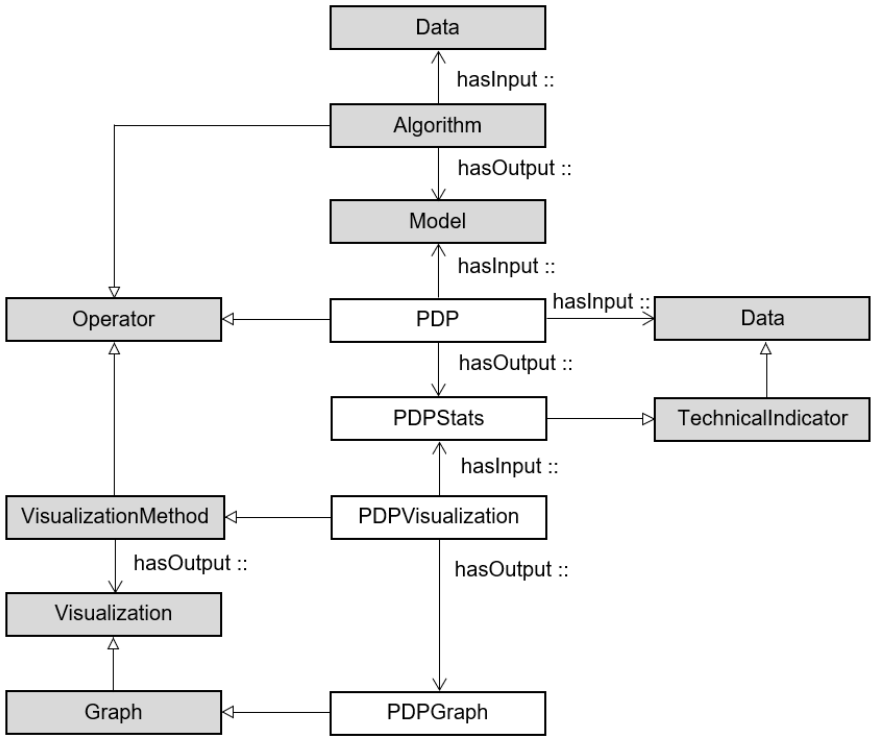


Figure 6
Example of the visualization concepts

4 Experiments

The proposed semantic model is intended for the formal description of the data analytical processes, ensuring their reproducibility and interoperability and for the automation of the analytical processes. To evaluate the proposed approach, we have applied the semantic model to two real-world case studies from the medical domain [23]. In the evaluation, we have at first manually annotated all scripts used for the pre-processing, modelling and evaluation of the machine learning models with the concepts from the proposed semantic model; and, second, compared code manually created by the data scientists with the code automatically generated from the knowledge graph.

For the comparison of the code, we have defined mainly metrics based on the code coverage [24], namely the number of lines of the code (including control statements for branching and cycling), the number of exact operator matches and

the number of operators with the partially matched parameters. We have defined the semantic constraints for the final data mining model to exactly match the results of the expert's code without the additional automatic optimization using the meta-learning schema. In addition to evaluating the coverage of the code, the accuracy of the learned models was also tested, and there were no significant differences in the quality of the learned models between the generated and the original code.

5 Evaluation

The first task was a binary classification for the diagnosis of breast cancer. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image, such as radius, texture, perimeter, area, etc. All features were real-valued numerical attributes without the missing values. Some attributes were derived from the source attribute but aggregation functions (e.g., overall average symmetry computed from the symmetry of each annotated cell, etc.). Together the dataset [25] contains 32 attributes. The code manually created by the data scientists [26] was in the R language and covered pre-processing, modelling, and evaluation for the six data mining algorithms for the classification ranging from decision trees, random forest, k-nearest neighbours, naïve Bayes classifier (with the normal distribution of the attribute probabilities) and support vector machines with the linear and polynomial kernels. The models were evaluated using the standard metric for the overall accuracy and precision/recall for the positive class.

For the evaluation, we have defined the semantic constraints for the final data mining model to exactly match the results of the expert's code without the additional automatic optimization using the meta-learning schema. The comparison of the code coverage between the experts' code and automatically generated code is summarized in Table 1.

Metric	**Expert code**	**Generated code**	**Coverage %**
Number of code lines	328	-	-
Number of code lines with operators	126		
Number of algorithms	6	6	100
Number of visualization methods	6	4	67
Number of evaluation metrics	1	1	1
Number of variables	98	80	82
Number of operator arguments	26	21	81
Number of functions	2	-	-
Number of branchings	8		

The scripts for this case were developed as the Shiny application, where a large part of the code is related to the programming of the user interface, which is not relevant to the data analytical task. From the remaining code, 126 lines contain the operations over data and models. The automatically generated code covers all 6 evaluated algorithms (decision tree, random forest, Naïve Bayes, k-nearest neighbour, neural network and SVM). Some parameters (i.e. number of decision trees in the random forest and the number of neighbours for kNN) were optimised in the original code with the simple cycle. This part of the optimisation was replaced in the semantic model with the equivalent grid-search operator. The models were evaluated with the same set of technical KPIs for binary classification (accuracy and contingency table) [27]. The coverage of the visualisation operators was 67%, where two visualisation methods were not covered by our current model. Non-covered visualisation methods show dependency between the selected technical KPI (e.g. precision) and one of the algorithm parameters (e.g. number of trees in the random forest algorithm). The coverage of variables also includes all variable aliases in the initial code (i.e. when the same data value is assigned to the two variables with different names). Overall, the semantic model has a much lover and consistent set of unique variables. All branchings in the original code were covered since they were related to the selection of the model. Additionally, the original code contains the definition of two helper functions. Both were used as data preprocessing operators, and both were replaced in the generated code with the equivalent functions from the standard R packages.

The second use case was also from the medical domain for the diagnosis of Acute lymphoblastic leukaemia (ALL). The dataset [28] was preprocessed from the set of images with a convolutional neural network, and the extracted features were reduced with the ANOVA method and with the importance of weighting based on the random forest tree algorithm. The final set of features includes 584 numerical attributes without the missing values. The code manually created by the data scientists was implemented in Python and covers preprocessing, modelling and evaluations.

Metric	**Expert code**	**Generated code**	**Coverage %**
Number of code lines	211	-	-
Number of code lines with operators	88		
Number of algorithms	4	4	100
Number of visualization methods	4(2)	4	100
Number of evaluation metrics	5	5	100
Number of variables	75		
Number of operator arguments	20		
Number of functions	5	-	-
Number of branchings	3		

In this case, the generated code covers all classification algorithms (random forest, support vector machine, naïve Bayes and k-nearest neighbours). Code also covers all five technical key performance indicators (accuracy, precision, recall, F1 score and confusion matrix). The visualisation methods cover mainly the visualisation of the confusion matrix with the heatmap. Additionally, scripts contain two visualisations of input image data before and after cropping, but these visualisations serve just for the visual checking for data scientists and are not relevant to the automatically generated code. An interesting case which was not covered by our current semantic framework is the usage of machine learning models for feature extraction. The framework just specifies that each predictive machine learning model can be used as a data operator for scoring (i.e. computation of the output prediction data element from the input data). However, in the presented use case, the internal state of the model (convolutional deep learning neural network) is used to extract input data features. Annotated extension, which was not covered in our current framework, is the case where the output of the predictive machine learning model is used for the feature importance weighting for feature selection. The latter case was a straightforward extension, but the former case requires better specification of the internal structure of the models (especially for deep learning models), which will be the goal of future work.

Conclusions

In this paper, we have presented the application of semantic technologies for automatization of the data-analytical processes. We have demonstrated that it is possible to semantically describe the goals of the data mining tasks and data analysis and automatically orchestrate the data mining workflows to find a solution to the goals. Additionally, the proposed semantic description can be used to formally document existing data analytical scripts for their reproducibility. In our experiments, we have demonstrated good coverage of the proposed semantic framework on the various real-case scenarios of data analysis in the medical domain. During the experiments, we have also identified some corner cases which were not initially covered in the framework, namely, visualization of dependencies between technical KPIs (performance metrics) and algorithm's parameters, usage of the internal state of the predictive models as the feature extraction method and usage of the output of the predictive models for feature selection.

Our experiments also showed that besides the overall quality of the final model (i.e., optimal business KPIs), an important part of the data scientist's work is also data understanding and post-analysis of results for explainability of the model. This is also the motivation for our future research, where we are planning to extend our semantic model to cover also constraints for better data understanding and explainability of the machine learning algorithms.

Acknowledgement

This work was supported by the grant APVV-16-0213, APVV-20-0232 and VEGA 1/0685/21.

References

[1] N. Guarino and P. Giaretta. “Formal ontology, conceptual analysis and knowledge representation”. International Journal of Human - Computer Studies 43, 625-640, 1995

[2] P. Panov. “A modular ontology of data mining”. Doctoral dissertation. Jožef Stefan International Postgraduate School, Ljubljana. 2012

[3] P. Panov, S. Dzeroski and L. N. Soldatova. “OntoDM: An Ontology of Data Mining”. In: 2008 IEEE International Conference on Data Mining Workshops. 2008. pp. 752-760

[4] M. Muchova, J. Paralic and M. Nemcik. “Using Predictive Data Mining Models for Data Analysis in a Logistics Company”. Information systems architecture and technology, PT I. Springer International Publishing AG, Gewerbestrasse 11, CHAM, Switzerland. pp. 161-170, 2018, doi: 10.1007/978-3-319-67220-5_15

[5] M. Hilario, P. Nguyen, H. Do, A. Woznica and A. Kalousis. “Ontology-Based Meta-Mining of Knowledge Discovery Workflows”. In: Meta-Learning in Computational Intelligence. 2011

[6] J. Kietz, F. Serban, A. Bernstein and S. Fischer. “Towards Cooperative Planning of Data Mining Workflows”. In: Proceedings of the Third Generation Data Mining Workshop at the 2009 European Conference on Machine Learning. 2009

[7] J. Vanschoren and L. Soldatova. “Exposé: An ontology for data mining experiments”. In: International workshop on third generation data mining: Towards service-oriented knowledge discovery. 2010. pp. 31-46

[8] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann and I. Horrocks. “The Semantic Web: the roles of XML and RDF”. IEEE Internet Computing 4, 2000. pp. 63-73

[9] P. Panov, L. N. Soldatova, and S. Džeroski. “Generic ontology of datatypes”, Information Sciences. 2016. pp. 900-920, doi: https://doi.org/10.1016/j.ins.2015.08.006

[10] M. Tianxing, N. Zhukova, A. Vodyaho, and T. Aung. “A Meta-Mining Ontology Framework for Data Processing”. International Journal of Embedded and Real-Time Communication Systems. 2021. pp. 37-56, doi: 10.4018/IJERTCS.2021040103

[11] J. Palša, J. Hurtuk, M. Chovanec and E. Chovancová. “Using Machine Learning Algorithms to Detect Malware by Applying Static and Dynamic Analysis Methods”. Acta Polytechnica Hungarica. 2022, pp. 177-196

[12] E. Patterson, I. Baldini, A. Mojsilovic and K. Varshney. “What is the Data Science Ontology?” [online] [cit. 31.01.2023] doi: <https://www.datascienceontology.org/help>

[13] J. Polowinski and M. Voigt. “VISO: a shared, formal knowledge base as a foundation for semi-automatic infovis systems”. In: CHI ’13 Extended Abstracts on Human Factors in Computing Systems, Paris France. 2013, pp. 1791-1796, doi: 10.1145/2468356.2468677

[14] G. Antoniou and F. Van Harmelen. “A semantic Web primer”. 2nd ed. Cambridge, Mass: MIT Press (Cooperative information systems) 2008, pp. 264

[15] Ontologies and Semantic Web: Working with Ontologies, [online] [cit. 2023-01-13] doi: <https://www.obitko.com/tutorials/ontologies-semantic-web/working-with-ontologies.html>

[16] A. Gomez-Perez and O. Corcho. “Ontology languages for the Semantic Web”. IEEE Intelligent Systems 17. 2002, pp. 54-60

[17] P. Bednár, J. Ivančáková and M. Sarnovský. “Semantic automatiozation of the data-analytical processes”. In: International Symposium on Applied Computational Intelligence and Informatics. 2022

[18] M. Sarnovsky, P. Bednar, and M. Smatana. “Cross-Sectorial Semantic Model for Support of Data Analytics in Process Industries”. Processes 7, No. 5: 281, 2019, doi: https://doi.org/10.3390/pr7050281

[19] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference. W3C Recommendation. World Wide Web Consortium. 2009, doi: https://www.w3.org/TR/skos-reference/

[20] M. Sarnovsky and J. Marcinko. “Adaptive Bagging Methods for Classification of Data Streams with Concept Drift”. Acta Polytechnica Hungarica. 2021. pp. 47-63, doi: 10.12700/APH.18.3.2021.3.3

[21] J. Bruijn, H. Lausen, A. Polleres, and D. Fensel. “The web service modelling language WSML: An overview”. 2006, pp. 604, doi: 10.1007/11762256_43

[22] Ch.Molnar, “Interpretable Machine Learning”, A Guide for Making Black Box Models Explainable, 2023, [online] [cit. 2023-03-02] doi: <https://christophm.github.io/interpretable-ml-book/>

[23] P. Šatala, P. Butka, A. Samaiev and P. Levická. “Cueing of Parkinson’s Disease Patients by Standard Smart Devices and Deep Learning Approach”. Acta Polytechnica Hungarica. 2023, pp. 165-184

[24] S.Pittet. “What is code coverage?” [online] [cit. 2023-01-29] doi: <https://www.atlassian.com/continuous-delivery/software-testing/code-coverage>

[25] Breast Cancer Wisconsin (Diagnostic) Data Set. [online] [cit. 2023-02-25] doi: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

[26] J. Ivančáková, F. Babič and P. Butka. “Comparison of different machine learning methods on Wisconsin dataset”. 2018 IEEE 16th World Symposium on Applied Machine Intelligence and Informatics. Kosice and Herlany, Slovakia, 2018, pp. 173-178, doi: 10.1109/SAMI.2018.8324834

[27] P. P. Bakucz and J. Z. Szabó. “Determining the Embedded Key Performance Indicator (KPI), based on a Fuzzy FxLMS Algorithm”. Acta Polytechnica Hungarica. 2021, pp. 127-139, doi: 10.12700/APH.18.9.2021.9.8

[28] A. Gupta and R. Gupta. ALL Challenge dataset of ISBI 2019 The Cancer Imaging Archive. Data set [online] [cit. 2023-02-25] doi: <https://doi.org/10.7937/tcia.2019.dc64i46r>

[29] V. Diaz and G. Rodríguez. “Machine Learning for Detection of Cognitive Impairment”. Acta Polytechnica Hungarica. pp. 195-213, 2022, doi: 10.12700/APH.19.5.2022.5.10

Similarity Measure Supported Fuzzy Failure Mode and Effect Analysis

Edit Laufer

Bánki Donát Faculty of Mechanical and Safety Engineering, Óbuda University, Bécsi 96/B, H-1034 Budapest, Hungary, laufer.edit@bgk.uni-obuda.hu

Abstract: Nowadays in the various engineering fields quality requirements are continuously increasing. There is also a need to develop flexible and highly adaptive systems to meet current requirements. At the same time, it is also essential to predict possible system failures and to address the arising problems appropriately. A widely used approach for predicting and preventing system failures is the Failure Mode and Effect Analysis (FMEA), which accompanies the entire development process and is able to adapt to changes in the system. The conventional method can be improved if fuzzy logic is incorporated into the evaluation. In this way the often arising subjectivity and uncertainty can be handled to ensure a more reliable result. In this paper, the author proposes a Fuzzy-FMEA (F-FMEA) based approach supported by similarity measures, for the system level. In the evaluation fuzzy arithmetic operations are applied to determine the Probability of Failure for the different failure codes. In addition to the single-expert F-FMEA system, the evaluation method that takes into account the opinions of multiple experts is also presented.

Keywords: risk assessment; expert system; Fuzzy Failure Mode and Effect Analysis; similarity measures

1 Introduction

As a consequence of the rapid development of technology, not only the opportunities in the engineering field are expanding, but at the same time the quality requirements are also increasing, while continuous availability must be ensured as well. These criteria call for a flexible, highly adaptive system that can be operated with high reliability. In order to ensure reliable operation, it is not only necessary to choose the right manufacturing method, but continuous failure-free operation, and quick identification and management of any failures that may arise are also indispensable [1], [2]. One of the most frequently used methods is the Failure Mode and Effect Analysis (FMEA), which is suitable for predicting and preventing system errors already in the planning phase, and can be continued throughout the entire life of the product or service. During the analysis, all possible events that could cause failure in the system during the process are classified and ranked. In the traditional

crisp FMEA method, the level of risk can be specified with numerical values between 1-1000. However, these values are difficult to quantify since tasks of this nature are full of uncertainty and subjectivity. This problem can be addressed using the fuzzy approach, as it uses linguistic terms and can handle the subjectivity, inaccuracy and uncertainty that arise during evaluation [3].

Due to the aforementioned advantageous properties of fuzzy logic, the reliability of the model can be significantly increased. In the Fuzzy FMEA (F-FMEA), instead of numerical risk values, fuzzy sets are used in the model. In the literature several papers are available related to the F-FMEA-based failure predicting and preventing method. N. Chanamool and T. Naenna developed a fuzzy FMEA model suitable for prioritizing and evaluating possible failures in the work processes of the emergency department to choose the appropriate action and increase the confidence on hospitals [4]. G. Jin, Q. Meng and W. Feng proposed an AHP (Analytic Hierarchy Process) supported F-FMEA method to analyze the causes of failure of the logistics system. In the system the weight of the risk indicators was determined using the AHP method [5]. In the paper of X. Hu, J. Liu and Y. Wang an ontology-based F-FMEA model is introduced, in which the rating based on entropy weight and fuzzy TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) [15], [7].

The above systems can work effectively if the evaluation of possible failures has to be compiled based on the unified opinion of a single group of experts. However, there may be disagreement within the group; and in order to make the assessment more reliable, it is worth asking for the opinions of several groups of experts, which may also differ. Handling different opinions properly is a considerable challenge since a consensus has to be arrived at [8]. In the literature several different methods are presented to address this problem, such as Ordered Weighted Averaging to aggregate expert preferences [9], consistency-based algorithms using fuzzy preference relations [10], or similarity measures.

Similarity measures are widely used in risk assessment based on its advantageous properties. This approach has favourable computational requirements because it can be calculated by comparing simple features of fuzzy sets [11], [12].

In this paper, the author makes a general, flexible proposal for similarity measure-based Fuzzy Failure Mode and Effect Analysis (SF-FMEA) model to specify the probability of the potential failures (PoF) focusing on the system level. Due to the fuzzy approach FMEA components are represented by fuzzy sets taking the advantage of using linguistic terms, and the manageability of uncertainties. In the system Consequence of Failure (CoF) is also considered for each potential failure codes to determine the overall system result. The author made a proposal both for the case when the opinion of a single unified group of experts is available, and when the potentially different opinions of several different groups must be taken into account. The current PoF values were determined using fuzzy arithmetic operators, then the result was compared to the reference fuzzy sets using similarity measures

to determine the current result. Furthermore, a similarity measure-based method was introduced to specify the magnitude of the consensus in the multiple-expert case to define a weight factor, which can be used in the aggregation of the expert groups' opinion.

The paper is organized as follows: In Section 2 the basic concepts related to fuzzy set theory, fuzzy operations, Fuzzy Failure Mode and Effect Analysis and similarity measures are defined. Section 3 presents the proposed similarity supported F-FMEA in two subsections: Subsection 3.1 introduces the case when the F-FMEA is prepared based on the opinion of a single expert group, while Subsection 3.2 considers the case when the opinion of multiple expert group is available, which may differ. In Section 4 a method is proposed to define the magnitude of the consensus between the different experts' opinion, and based on this value a weight factor is defined to calculate the aggregated experts's opinion. Then, in the Conclusions section the results are summarized.

2 Applied Methods

2.1 Preliminaries

This section outlines the definitions of concepts essential for the presentation of the methods.

Generalized fuzzy set: Fuzziness can be represented by a fuzzy set, which is devoted to specify the extent to which the element belongs to the set (membership degree). The fuzzy set, $A(a,b,c,d,h_A)$ is determined by a continuous mapping (membership function) from R to the closed interval [0,1]. Trapezoidal membership function is represented by (1).

$$\mu_A(x)=\begin{cases} h_A\dfrac{x-a}{b-a} & \text{if} \quad a\le x\le b \\ h_A & \text{if} \quad b\le x\le c \\ h_A\dfrac{d-x}{d-c} & \text{if} \quad c\le x\le d \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $a \neq b, c \neq d$ and h_A is the maximum value of the set, $h_A \in]0,1]$ [13].

Normality of a fuzzy set: The normality of the fuzzy set A is basically determined by its highest value (h_A). In most cases normal sets, i.e. $h_A = 1$, are used. However, in the case of generalized fuzzy sets lower value is also allowed, i.e. $0 < h_A \leq 1$ [13].

Defuzzification: Defuzzification is a process when a fuzzy set (e.g. system result) should be represented by a suitable crisp value. The most commonly used method is the Centre of Gravity (CoG), which can assign a crisp value to any shape set properly. However, the greatest disadvantage of this approach is the high computational requirement. To handle this drawback the Simplified Centre of Gravity (SCoG) method is used in this study. The basis of this approach is the center curve of the trapezoidal fuzzy set [11], [14] which can be calculated for trapezoidal sets, $A(a, b, c, d, h_A)$ according to (2), (3).

$$y_{SCoG_A} = \frac{h_A\left(\frac{c-b}{d-a}+2\right)}{6} \tag{2}$$

$$x_{SCoG_A} = \frac{y_{SCoG_A}(c+b)+(d+a)\left(h_A - y_{SCoG_A}\right)}{2h_A} \tag{3}$$

Fuzzy arithmetic operations: In order to be able to perform operations with generalized fuzzy numbers, arithmetic operations should be defined. In this study, Chen's operators are used for the above defined fuzzy sets $\left(A_1(a_1, b_1, c_1, d_1, h_{A_1});\ A_2(a_2, b_2, c_2, d_2, h_{A_2})\right)$ as follows [15]:

Addition: $(A_1 \oplus A_2) = \left(a_1 + a_2, b_1 + b_2, c_1 + c_2, d_1 + d_2, min\left(h_{A_1}, h_{A_2}\right)\right)$ (4)

Subtraction: $(A_1 \ominus A_2) = \left(a_1 - d_2, b_1 - c_2, c_1 - b_2, d_1 - a_2, min\left(h_{A_1}, h_{A_2}\right)\right)$ (5)

Multiplication: $(A_1 \otimes A_2) = \left(a_1 \times a_2, b_1 \times b_2, c_1 \times c_2, d_1 \times d_2, min\left(h_{A_1}, h_{A_2}\right)\right)$(6)

Division: $(A_1 \oslash A_2) = \left(\frac{a_1}{a_2}, \frac{b_1}{b_2}, \frac{c_1}{c_2}, \frac{d_1}{d_2}, min\left(h_{A_1}, h_{A_2}\right)\right)$ (7)

2.2 Fuzzy Failure Mode and Effect Analysis

Failure Mode and Effect Analysis (FMEA) is an effective technique for predicting and preventing system failures. It is a commonly used approach in manufacturing systems, mainly in those that produce safety-critical products and contain advanced electronic and mechanical equipment based on system analysis [16]. The essence of the method is to qualify and prioritize the random and natural events occurring in the system during the process, which can cause damage. Each failure mode is characterized by three metrics: Consequence of Failure (*CoF*), Probability of Failure (*PoF*) and Detectability of Failure (*DoF*). These three aspects are often referred to in the literature as Severity (here *CoF*), Occurrence (here *PoF*), and Detectability (here *DoF*). The crisp FMEA method is based on a numerical scale, ranging from 1 to 10, where 1 is the lowest risk and 10 is the highest. Taking into

account these three characteristics together, Risk Priority Number (*RPN*) has to be calculated, using (8) to be able to rank the particular risk scenarios.

$$\mathrm{RPN}_i = \mathrm{CoF}_i \cdot \mathrm{PoF}_i \cdot \mathrm{DoF}_i \tag{8}$$

where $i \in [1, n]$, n is the number of the different failure codes.

These kinds of tasks are full of uncertainties and subjectivity. The fuzzy approach is a good solution for this problem because it is able to handle subjectivity, imprecision and uncertainty in the evaluation. In this way the reliability of the model can be significantly increased. In order to fuzzify the process *CoF*, *PoF* and *DoF* should be reprezented by fuzzy sets instead of crisp numbers. These sets have to be a partition of the [0,10] interval. In this study, [0,1] interval is used proportionally due to later calculations. Fuzzy sets representing *CoF*, *PoF* and *DoF* values are illustrated in Fig. 1 [17].

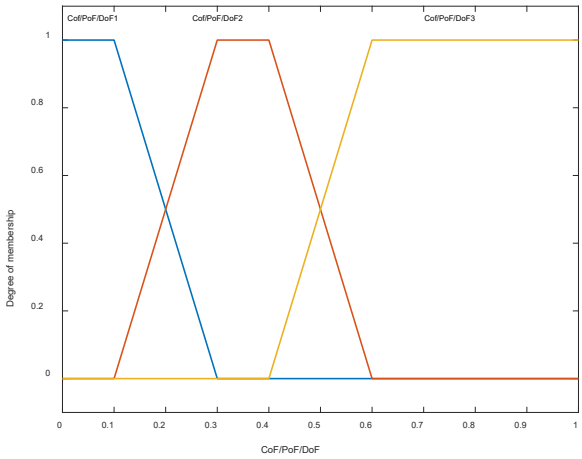


Figure 1

CoF, PoF and DoF fuzzy sets (Linguistic terms for CoF/PoF/DoF : CoF/PoF/DoF1= Low/Improbable/EasilyDetectable, CoF/PoF/DoF2= Medium/Occasional/Detectable, CoF/PoF/DoF3= High/Probable/HardlyDetectable, respectively)

In the Fuzzy FMEA (F-FMEA) the *RPN* value is determined by a fuzzy inference system, where the evaluation is based on a rule base [4]. In this study, the Mamdani-type inference is used, i.e. the output domain is also covered by fuzzy sets (see Fig. 2).

The input of the F-FMEA can be a crisp value or even a fuzzy number. However, fuzzy set-represented expert knowledge is more informative. Consequently, in this study, the fuzzy number type opinions are considered. Similarly to the crisp FMEA, in its fuzzy version each failure code has to be evaluated using the above method.

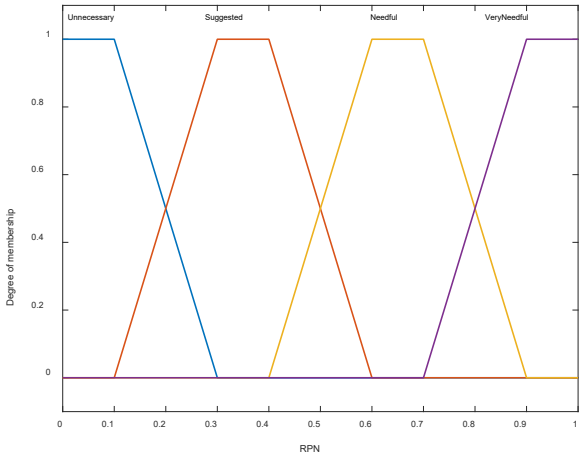


Figure 2
RPN fuzzy sets representing the necessity of action

2.3 Similarity Measures

Similarity measures are used to compare fuzzy sets and numbers calculating the degree of similarity, $0 < S(A_1, A_2) \leq 1$, where A_1, A_2 are fuzzy sets or numbers. If the similarity value is 1, the fuzzy sets are the same. The lower its value the greater the difference between the sets. Similarity is determined based on the characteristics of fuzzy sets, there are many different approaches. In this study, the fuzzy set parameters and the defuzzified value are used for the comparison, calculating the SCoG value for each fuzzy set by (2), (3). Similarity calculation can be performed using (9), (10), (11).

$$S(A_1, A_2) = c\left(1 - \frac{|a_1 - a_2| + |b_1 - b_2| + |c_1 - c_2| + |d_1 - d_2|}{4}\right)\Delta x_{SCoG} M \tag{9}$$

where

$$\Delta x_{SCoG} = 1 - \left| x_{SCoG_{A_1}} - x_{SCoG_{A_2}} \right| \tag{10}$$

$$M = \frac{\min\left(y_{SCoG_{A_1}}, y_{SCoG_{A_2}}\right)}{\max\left(y_{SCoG_{A_1}}, y_{SCoG_{A_2}}\right)} \tag{11}$$

c is a constant to specify the direction of the deviation, if needed. If the direction is not relevant, or $x_{SCoG_{A_1}} \geq x_{SCoG_{A_2}}$ then $c = 1$. If $x_{SCoG_{A_1}} < x_{SCoG_{A_2}}$ then $c = -1$.

3 Fuzzy Failure Mode and Effect Analysis using Simiarity Measures

In the F-FMEA method the potential failures of the system can be represented by three main components, such as the Probability of Failure (*PoF*), the Consequence of Failure (*CoF*), and the Detectability of Failure (*DoF*). In the fuzzy approach these components are characterized by fuzzy sets taking the advantage of the use of linguistic terms and the manageability of uncertainties. In this study, the focus is on the *PoF* value. The main goal is to determine its overall value taking into account all failure codes determined by the experts. In this paper, the failure codes are not specified, as this is a general suggestion that can be flexibly applied to different specific systems, and failure codes.

3.1 Single Expert Case

In the F-FMEA process, the fuzzy reference sets shown in Figure 1 are used both during the expert classification of individual error codes, and the overall system output is compared with them.

Let the failure codes be $C_1, C_2, \dots, C_n$, where n is the number of the potential failures, and each C_i is characterized by its corresponding PoF_i and CoF_i. The overall PoF of the system is determined by the fuzzy weighted average calculated using fuzzy arithmetic operations (see 2.1) as follows:

$$\mathrm{PoF_O} = \frac{\sum_{i=1}^{n} \mathrm{CoF_i} \otimes \mathrm{PoF_i}}{\sum_{i=1}^{n} \mathrm{CoF_i}} \tag{12}$$

Based on the calculated *PoF* parameters, which represent a normal fuzzy set, the probability of an error occurring in the system can be determined using similarity measures. The overall PoF set (PoF_O) and reference PoF sets (Fig. 1) should be compared by (9), (10), (11). Based on the highest similarity value, it can be determined which of the reference sets the overall *PoF* is closest to.

Let the number of the failure codes be 5, for which the expert opinion is defined according to Table 1.

Using (12), the overall *PoF* value is as follows:

$\mathrm{PoF_O} = (0.21, 0.35, 0.58, 0.67)$ as illustrated in Fig 3.

Table 1
Expert's opinion for the different failure codes

Failure code	CoF	$CoF_i(a_1, b_1, c_1, d_1)$	PoF	$PoF_i(a_1, b_1, c_1, d_1)$
Failure1	Low	(0,0,0.1,0.3)	Improbable	(0,0,0.1,0.3)
Failure2	Medium	(0.1,0.3,0.4,0.6)	Probable	(0.4,0.6,1,1)
Failure3	High	(0.4,0.6,1,1)	Probable	(0.4,0.6,1,1)
Failure4	High	(0.4,0.6,1,1)	Improbable	(0,0,0.1,0.3)
Failure5	Medium	(0.1,0.3,0.4,0.6)	Occasional	(0.1,0.3,0.4,0.6)

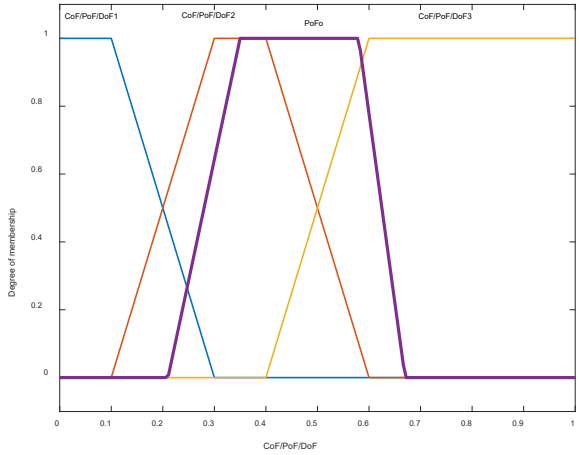


Figure 3
Comparison of the overall PoF set to the reference sets (CoF/PoF/DoF1, CoF/PoF/DoF2, CoF/PoF/DoF3 represent the reference sets, while PoFo is the overall PoF set)

After the PoF_O value is available, one has to compare it to the reference fuzzy sets (see Fig. 1) to obtain the final result. Similarity values for all reference sets are listed in Table 2. The highest value determines which linguistic variable can be assigned to the *PoF* value of the overall system. It can be seen that the highest value is 0.714, and the associated fuzzy set is PoF_2 representing the linguistic term *Occasional*. This result means that intervention may be necessary to avoid the occurence of a potential failure.

Table 2
Similarity values of the overall PoF and reference sets

PoF_i	$S(PoF_i, PoF_O)$
PoF_1 (Improbable)	0.402
PoF_2 (Occasional)	0.714
PoF_3 (Probable)	0.466

3.2 Aggregated Experts' Opinion-based Evaluation

In order to compile an effective FMEA, one should consider the opinions of several experts. However, these opinions may differ, requiring great care to be handled appropriately. In this section, the author proposes the multiexpert version of the similarity measures supported FMEA to address the above problem.

This method is an extension of the similarity measures supported evaluation process (see 3.1). In the above described case only one expert's opinion is available, therefore, normal fuzzy sets can be used effectively. However, in the multiexpert version, the opinions of several experts, which may differ, should all be taken into account. For this reason, these opinions have to be weighted based on the degree of confidence of the experts using subnormal fuzzy sets. The height of the generalized fuzzy set is used to represent the degree of confidence (DoC) of each expert. First, the problem is reduced by averaging the different opinions for each failure code by (13), (14) determining the average *PoF* value $\left(PoF_{avg_i}\right)$.

$$\mathrm{PoF}_{\mathrm{avg_i}} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{PoF}_{\mathrm{ij}}}{\mathrm{m}} \tag{13}$$

$$\mathrm{PoF}_{\mathrm{avg_i}} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{a}_{\mathrm{ij}}}{\mathrm{m}}, \frac{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{b}_{\mathrm{ij}}}{\mathrm{m}}, \frac{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{c}_{\mathrm{ij}}}{\mathrm{m}}, \frac{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{d}_{\mathrm{ij}}}{\mathrm{m}}, \frac{\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{h}_{\mathrm{ij}}}{\mathrm{m}} \tag{14}$$

$$\mathrm{PoF}_{\mathrm{avg_i}} = \mathrm{a}_{\mathrm{avg_i}}, \mathrm{b}_{\mathrm{avg_i}}, \mathrm{c}_{\mathrm{avg_i}}, \mathrm{d}_{\mathrm{avg_i}}, \mathrm{h}_{\mathrm{A_{avg_i}}} \tag{15}$$

where $j \in [1, m]$, m is the number of the expert teams, $i \in [1, n]$, n is the number of the different failure codes, $a_{ij}, b_{ij}, c_{ij}, d_{ij}, h_{ij}$ are the generalized fuzzy set parameters for failure code i, and expert j, while $a_{avg_i}, b_{avg_i}, c_{avg_i}, d_{avg_i}, h_{A_{avg_i}}$ represent the average fuzzy set parameters for failure code i.

Following the aggregation, the process is the same as in the original (single expert) case, but instead of the opinion of the single expert, the above calculated average *PoF* value $\left(PoF_{avg_i}\right)$ is used. The next step is the overall *PoF* value calculation by (16), then the obtained generalized fuzzy number $\left(PoF_{avg_O}\right)$ compared to the reference fuzzy sets specified in Fig 1. Comparison is performed by similarity measure using (9), (10), (11) and the reference set for which the largest value obtained represents the system result.

$$\mathrm{PoF}_{\mathrm{avg_O}} = \frac{\sum_{i=1}^{n} \mathrm{CoF_i} \otimes \mathrm{PoF}_{\mathrm{avg_i}}}{\sum_{i=1}^{n} \mathrm{CoF_i}} \tag{16}$$

where $i \in [1, n]$, n is the number of the failure codes.

Let the number of the failure codes be 5, and the number of the different expert groups be 3. The opinion of the groups are presented in Table 3, where the Degree of Confidence (DoC) of the groups are represented by the height of the fuzzy sets.

Table 3
Expert groups' opinion for the different failure codes

Failure code	Group1 (DoC=0.9)	Group2 (DoC=0.7)	Group3 (DoC=0.8)
Failure1	Improbable	Occasional	Improbable
Failure2	Probable	Occasional	Occasional
Failure3	Probable	Occasional	Occasional
Failure4	Improbable	Improbable	Occasional
Failure5	Occasional	Probable	Probable

First, average *PoF* value should be calculated by taking into account the opinion of all expert groups using (13), (14), (15). These average values are summarized in Table 4.

Table 4
Average PoF and CoF values for each failure code

Failure code	$\mathrm{PoF}_{\mathrm{avg_i}}$	$\mathrm{CoF_i}$
Failure1	(0.033,0.1,0.2,0.4)	(0,0,0.1,0.3)
Failure2	(0.2,0.4,0.6,0.733)	(0.1,0.3,0.4,0.6)
Failure3	(0.2,0.4,0.6,0.733)	(0.4,0.6,1,1)
Failure4	(0.033,0.1,0.2,0.4)	(0.4,0.6,1,1)
Failure5	(0.3,0.5,0.8,0.867)	(0.1,0.3,0.4,0.6)

After the average values are available, the PoF value can be calculated for the overall system in the same way as in the single expert case, and the overall PoF set is as follows:

$$\mathrm{PoF}_{\mathrm{avg_O}} = (0.143, 0.317, 0.476, 0.632)$$

The final step of the process is to compare the overall PoF value to the reference fuzzy sets (see Fig. 1). The degree of similarities are presented in Table 5.

Table 5
Similarity values of the overall PoF and reference sets

PoF_i	$S(PoF_i, PoF_O)$
PoF_1 (Improbable)	0.406
PoF_2 (Occasional)	0.777
PoF_3 (Probable)	0.292

The highest value determines which linguistic variable can be assigned to the PoF value of the overall system. It can be seen that the highest value is 0.777, and the associated fuzzy set is PoF_2, representing the linguistic term *Occasional*. This result means that intervention may be necessary to avoid the occurence of a potential failure.

4 Consensus-based Similarity Supported FMEA Model

In this section a comparison method is introduced, whose main purpose is to represent the magnitude of the consensus between the different experts' opinion. Then, based on the obtained value a weight factor is defined, by which the aggregated experts's opinion can be calculated.

In this study, the comparison is performed taking into account the *PoF* value based on the experts' opinion, represented by fuzzy sets. During the evaluation subnormal fuzzy sets $A(a, b, c, d, h_A)$ are applied, where the height of the set (h_A) represents the degree of confidence associated with each expert.

To determine the degree of consensus, one should perform the following process for each failure code:

1. Fuzzy set $A(a, b, c, d, h_A)$ creation based on the experts' opinion
2. SCoG (x_{SCoG}, y_{SCoG}) value is calculated for each fuzzy set by (2), (3)
3. Similarity calculation to compare the sets by (9), (10), (11)

The result of this process is the magnitude of the consensus between the different experts. The greater the obtained values the higher the consensus. Its maximum value is 1, which means that the different experts completely agree on the specific error code. Based on the magnitude of the consensus a weight factor can be specified to use when the experts' opinion are aggregated. Fuzzy sets with identical or nearly identical parameters can be represented by a single set. Then, the number of these kinds of sets is used to calculate the weight factor (w_j) of this single set. This weight factor ensures the work with normal fuzzy sets, i.e., instead of the height of the set, a weight factor is used. In this case, the average *PoF* value can be calculated as follows:

$$PoF_{wavg_i} = \frac{\sum_{j=1}^{l} w_j PoF_{ij}}{\sum_{j=1}^{l} w_j} \tag{17}$$

$$PoF_{wavg_i} = \frac{\sum_{j=1}^{m} w_j a_{ij}}{\sum_{j=1}^{l} w_j}, \frac{\sum_{j=1}^{m} w_j b_{ij}}{\sum_{j=1}^{l} w_j}, \frac{\sum_{j=1}^{m} w_j c_{ij}}{\sum_{j=1}^{l} w_j}, \frac{\sum_{j=1}^{m} w_j d_{ij}}{\sum_{j=1}^{l} w_j} \tag{18}$$

$$PoF_{wavg_i} = a_{wavg_i}, b_{wavg_i}, c_{wavg_i}, d_{wavg_i} \tag{19}$$

where fuzzy sets are represented by their basic parameters $A(a,b,c,d)$, $j \in [1,l]$, l is the number of the different fuzzy sets, w_j is the weigth factor of fuzzy set j, $i \in [1,n]$, n is the number of the different failure codes, $a_{ij}, b_{ij}, c_{ij}, d_{ij}$ are the normal fuzzy set parameters for failure code i, and fuzzy set j, while $a_{wavg_i}, b_{wavg_i}, c_{wavg_i}, d_{wavg_i}$ represent the average fuzzy set parameters for failure code i.

Let the number of the failure codes be 5, and the number of the different expert groups be 3. The opinion of the groups are presented in Table 3, but normal fuzzy sets are used. First, the similarity degree is calculated for expert groups in pairs for each failure code separately. These values represent the magnitude of the consensus. Based on these results fuzzy sets with identical or nearly identical parameters can be represented by a single set, whose weight is determined accordingly.

Table 6
Similarity values for each failure code

Failure code	S(Group1,Group2)	S(Group1,Group3)	S(Group2,Group3)
Failure1	0.538	1.000	0.538
Failure2	0.300	0.300	1.000
Failure3	0.300	0.300	1.000
Failure4	1.000	0.538	0.538
Failure5	0.300	0.300	1.000

Based on Table 6 it can be seen which fuzzy sets are identical (similarity values are 1). These sets are represented by a single set and their weight is doubled. The resulting sets are then averaged using (17), (18), (19) as illustrated in Table 7.

Table 7
Weighted average of PoF and CoF values for each failure code

Failure code	PoF_{avg_i}	CoF_i
Failure1	(0.02,0.06,0.16,0.36)	(0,0,0.1,0.3)
Failure2	(0.16,0.36,0.52,0.68)	(0.1,0.3,0.4,0.6)
Failure3	(0.16,0.36,0.52,0.68)	(0.4,0.6,1,1)
Failure4	(0.02,0.06,0.16,0.36)	(0.4,0.6,1,1)
Failure5	(0.34,0.54,0.88,0.92)	(0.1,0.3,0.4,0.6)

After the average values are available, the PoF value can be calculated for the overall system in the same way as in the single expert and DoC-based multiexpert case. The overall PoF set is as follows:

$$PoF_{wavg_O} = (0.122, 0.29, 0.433, 0.602)$$

The final step of the process is to compare the overall PoF value to the reference fuzzy sets (see Fig. 1). The degree of similarities are presented in Table 6.

Table 8
Similarity values of the overall PoF and reference sets

PoF_i	$S(PoF_i, PoF_O)$
PoF_1 (Improbable)	0.382
PoF_2 (Occasional)	0.720
PoF_3 (Probable)	0.446

The highest value determines which linguistic variable can be assigned to the PoF value of the overall system. It can be seen that the highest value is 0.720, and the associated fuzzy set is PoF_2, representing the linguistic term *Occasional*. This result means that intervention may be necessary to avoid the occurence of a potential failure.

Comparing the results of the DoC-based and consensus-based approach (see Table 9), it is clear that the highest similarity can be seen with reference set 2 in both cases. However, for the other two reference sets, the similarity is reversed. In the consensus-based model, the result is shifted to the PoF_3, which means that it makes the occurence of a potential failure in the system more likely.

Table 9
Comparison of the DoC-based and consensus-based models

PoF_i	$S(PoF_i, PoF_O)$	$S(PoF_i, PoF_O)$
PoF_1 (Improbable)	0.406	0.382
PoF_2 (Occasional)	0.777	0.720
PoF_3 (Probable)	0.292	0.446

Conclusions

In engineering systems, it is not only necessary to apply the technologically appropriate method, but also to continuously avoid any unwanted events. Failure Mode and Effect Analysis is one of the most commonly used approaches suitable for the quick identification and management of potential failures in the system. The extension of this method with fuzzy logic (F-FMEA) makes it possible to handle uncertainties, subjectivity and imprecision in the evaluation.

In this paper a similarity measure-based F-FMEA model was proposed for determining the overall Probability of Failure in the system. Similarity measures are very popular in risk assessment applications because of their favourable computational properties. In this study, different potential failures were considered characterized by their PoF and CoF values. During the evaluation fuzzy arithmetic operators were used to determine the overall system result. Then, the results were interpreted based on the comparison with the reference fuzzy sets. The basic method takes into account a single expert's opinion. However, in order to make the results of the system more reliable, the opinions of several experts must be taken into account. In this case, the greatest challenge is that the opinion of the expert groups can often be different. For this reason. author also proposed a multiexpert vesion of the similarity supported F-FMEA to address the above problem. In this DoC-based model the Degree of Confidence for each expert groups was considered, which is represented by the height of the generalized fuzzy sets. Furthermore, the magnitude of the consensus between the different expert groups was also calculated using similarity measures. Then, based on the obtained result, a weight factor was defined, which was used in the overall PoF value calculation.

The methods were illustrated by numerical examples and the results of the DoC-, and consensus-based methods were compared. The comparison resulted in the same linguistic term as the system result. However, for the other two reference sets, the similarity was reversed.

Acknowledgement

This work was supported by the Fuzzy Systems Scientific Group at the Bánki Donát Faculty of Mechanical and Safety Engineering of Óbuda University.

References

[1] F. Bognár, P. Benedek: A Novel Risk Assessment Methodology – A Case Study of the PRISM Methodology in a Complience Management Sensitive Sector, Acta Polytechnica Hungarica, Vol. 18, No. 7, pp. 89-108, 2021

[2] D. Macuna, M. Laketic, D. Pamucar, D. Marinkovic: Risk Analysis Model with Interval Type-2 Fuzzy FMEA – Case Study of Railway Infrastucture Projects in the Republic of Serbia, Acta Polytechnica Hungarica, Vol: 19, No. 3, pp. 103-118, 2022

[3] L. Pokorádi: Application of Fuzzy Set Theory for Risk Assessment, Journal of Konbin, Vol. 14-15, No. 1, pp. 187-196, 2010, doi: 10.2478/v10040-008-0177-5

[4] N. Chanamool, T. Naenna: Fuzzy FMEA Application to Improve Decision-Making Process in an Emergency Department, Applied Soft Computing, Vol. 43, pp. 441-453, 2016, doi: 10.1016/j.asoc.2016.01.007

[5] G. Jin, Q. Meng and W. Feng: Optimization of Logistics System with Fuzzy FMEA-AHP Methodology, Processes, Vol. 10, No. 10, paper id: 1973, 2022, doi: 10.3390/pr10101973

[6] X. Hu, J. Liu and Y. Wang: Multi Ontology-Based System-Level Software Fuzzy FMEA Method, Proc. of 6th International Conference on Dependable Systems and Their Applications (DSA), Harbin, China, 2020, doi: 10.1109/DSA.2019.00015

[7] M. Chackraborty: TOPSIS and Modified TOPSIS: A Comparative Analysis, Decision Analytics Journal, Vol. 2, paper id: 100021, 2021, doi: 10.1016/j.dajour.2021.100021

[8] C. Lu, J. Lan, Z. Wang: Aggregation of Fuzzy Opinions Under Group Decision-Making Based on Similarity and Distance, Journal System Science and Complexity, Vol: 19, 2006, pp. 63-71

[9] E. Herrera-Viedma et. al: A Consensus Model for Group Decision Making with Incomplete Fuzzy Preference Relations, IEEE Transactions on Fuzzy Systems, Vol. 15, No. 5, pp. 863-877, 2007, doi: 10.1109/TFUZZ.2019.2893307

[10] F. Y. Meng, J. Tang, H. Fujita: Consistency-based Algorithms for Decision-Making with Interval Fuzzy Preference Relations, IEEE Transactions on Fuzzy Systems, Vol. 27, No. 10, pp. 2052-2066, 2019, doi: 10.1109/TFUZZ.2019.2893307

[11] S-J. Chen, S-M. Chen: Fuzzy Risk Analysis Based on Similarity Measures of Generalized Fuzzy Numbers, IEEE Transactions on Fuzzy Systems, Vol. 11, No. 1, pp. 45-56, 2003

[12] Zs. Cs. Johanyák, Sz. Kovács: Distance based Similarity Measures of Fuzzy Sets, SAMI 2005, 3rd Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence, Herl'any, Slovakia, January 21-22 2005, ISBN: 963 7154 35 3, pp. 265-276

[13] D. Dubois, H. Prade: New Results about Properties and Semantics of Fuzzy Set-Theoretic Operations, Fuzzy Sets, Springer, pp. 59-75, 1980, https://doi.org/10.1007/978-1-4684-3848-2_6

[14] E. Tóth-Laufer, I. Z. Batyrshin: Similarity-based Personalized Risk Calculation, Proceedings of 16th IEEE International Symposium on Applied

Computational Intelligence and Informatics, Timisoara, Romania, 25-28 May 2022, pp. 129-133

[15] S. H. Chen: Operations on Fuzzy Numbers with Functional Principal, Journal of Management Science, Vol. 6, No. 1, pp. 13-26, 1985

[16] Stamatis and D. H.: Failure Mode and Effect Analysis: FMEA From Theory to Execution, Vol. 38, No. 1, 1996, doi: 10.1080/00401706.1996.10484424

[17] S. Kocak, L. Pokorádi, E. Tóth-Laufer: Fuzzy Hierarchical Failure Mode and Effect Analysis, Proc of the 15th IEEE International Symposium on Applied Computational Intelligence and Informatics SACI 2021, Timisora, Romania, pp. 311-316, 2021

Mathematical Modeling of a Hydraulic Lifting System for the Autodock Docking Leveler

Bogdan Dorel Cioroagă[1], Vasile George Cioată[2], Imre Kiss[3]

[1] Politehnica University Timişoara, Doctoral School, 5 Revoluției Street, 331128 Hunedoara, Romania, bogdan.cioroaga@student.upt.ro

[2] Politehnica University Timişoara, Department of Engineering and Management, 5 Revoluției Street, 331128 Hunedoara, Romania, vasile.cioata@fih.upt.ro

[3] Politehnica University Timişoara, Department of Engineering and Management, 5 Revoluției Street, 331128 Hunedoara, Romania, imre.kiss@fih.upt.ro

Abstract: This paper focuses on aspects regarding the design of a universal lifting system (actuation) of the platform of a product such as hydraulically operated docking levelers, with example on the Autodock leveler, which can be found as a component of the external docking stations of logistics warehouses. Aspects regarding the configuration of this product are treated as well as functional aspects reflected through a mathematical model that describes the interdependence between the parameters that describe the movement of this mechanical system. It also describes the situation of the stresses that occur in this system during operation and how they vary depending on certain key parameters. Summarizing, an image is created on a design method that imposes a set of conditions and universal steps valid for the design of the leveler lifting system for any special configuration found in the required configuration limits. The aim of the paper is to form a clear picture of the interdependence of the parameters that shape the designed lifting system.

Keywords: docking; model; hydraulic; system; design; leveler

1 Introduction

Docking levelers are special devices installed in logistics warehouses with the aim of making the connection between the entrance to the warehouse through the docking station and the vehicle visiting the warehouse [1], [2], [3]. Loading ramps provide maximum loading speed and efficiency by building a permanent bridge between the loading area and the vehicle in areas where there is a lot of entry and exit goods shipment. The hinged loading ramp consists of two main parts:

— the main platform, which includes a ramp (hinged along its rear edge), and

— the lip, hinged at the front of the ramp [1], [2], [3].

Thanks to this simple but well–designed system, it is possible to control standard hinged ramps with a single button. Equipment and logistics machinery transits over these docking levelers, which can be found in various types and sizes.

The hydraulic oil pump, valves, pistons, and the control board move the main platform and lip properly. The paper deals with the design of the lifting system of an Autodock hydraulic leveler (1), a leveler that is installed outside the warehouse building, which supports on the edges of its frame a metal loadhouse construction (2) on which is installed a sealing device called shelter (3).

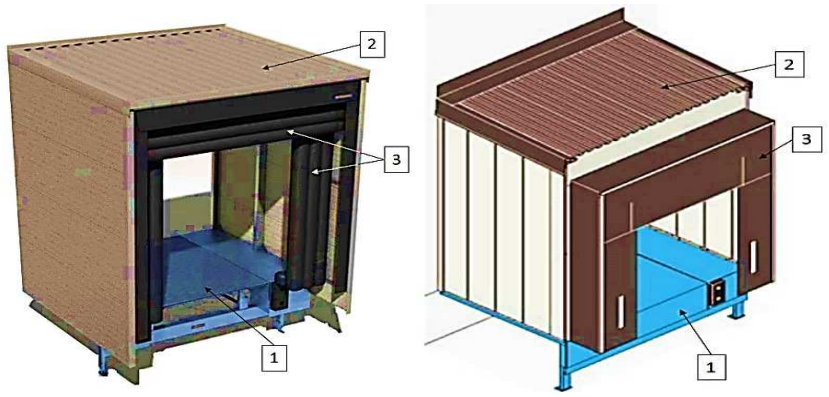


Figure 1
Docking station located outside the logistics warehouse, equipped with Autodock leveler

This is a table example: The problem of designing a docking platform lifting system is necessary due to the character of this product being found as a series product in standard configurations for newly built warehouses and as a product with special configuration for existing long–term warehouses (especially in Europe) which they have not been communized with the new regulations and standards.

At present, we used a set of linear hydraulic motors as drive systems for Autodock levelers but not only. Due to the wide range of configuration of the leveler dimensions, this platform lifting system is also influenced, which creates a need for the implementation of rules and algorithms in the design process for positioning and choosing the correct linear hydraulic motors found in this system. This paper will define a series of parameters and conditions necessary to obtain a universal solution in terms of designing this lifting system for docking levelers with example on the case of Autodock type leveler [1], [2], [3].

These automatic loading dock levelers and manual dock plates work with most types and sizes of vehicles to keep your operations running safely and smoothly [1]. Hydraulic telescopic–lip dock levellers have a movable telescopic lip, which provides a larger contact area between vehicle bed and dock leveller. Thanks to that, they can be precisely positioned on the vehicle bed for optimal load utilization and improved safety [1]. Preventive maintenance is easy and fast to secure functionality. All hydraulic telescopic–lip dock levellers are easy to operate [1].

2 Leveler Description

The Autodock docking leveler is composed of 3 large subassemblies that are found everywhere in the construction of hydraulically operated docking levelers, namely: frame, platform and lip. Figure 2 shows the main components of the docking Autodock type leveler, they can be found in several combinations of types and sizes, depending on the configuration of the leveler, imposed by the operating situations and other technical factors derived from the docking process.

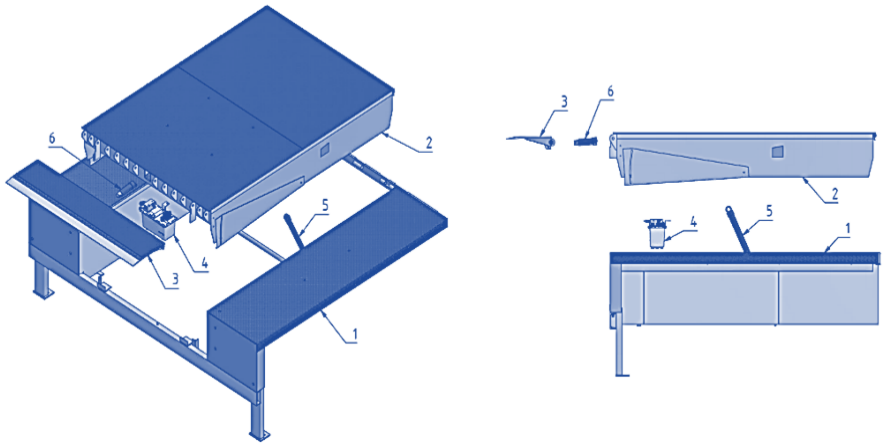


Figure 2
Autodock docking leveler components

According to Figure 2, the following constructive components that make up the leveler can be described:

1. The frame – represents the support structure both for the other components of the leveler and for the loadhous that is installed on its side wings. The frame is supported at the front by two support legs and at the other end by a profile behind the frame, which is welded at installation by a metal support embedded in the foundation of the building outside which the leveler is installed.
2. The platform – is made of striped sheet plate reinforced by L profiles, made of sheet metal, arranged longitudinally below it. The role of the platform is to be transited by logistics equipment, at the same time it supports other elements and accessories that are found in the composition of the leveler. It is installed in the frame by hinges located on the back of it and at the other end being installed the lip and two side guards to cover the space between the platform and the frame that appeared during the operation of the leveler.
3. The lip – has a double role, makes the transition between the leveler platform and the logistics vehicle visiting the docking station, the second role is supporting the platform during the resting position of the leveler when the platform is supported in the rear hinges and lip. The lip can be of two types, the

case of the articulated lip performing a rotational movement (with platform hinges) or telescopic, in which case it performs a translational movement and support on the support legs.

4. The hydraulic unit – consists of a hydraulic pump, motor and oil tank, it has the role of supplying and taking oil from the hydraulic system.
5. Hydraulic lifting motors – are linear hydraulic motors with simple action and are mounted in the supports on the frame and those on the platform, having the role of raising and lowering the platform to bring it into operation. These are found one on the right and one on the left inside the frame.
6. Hydraulic lip drive motor – is a single linear hydraulic motor with single action in the case of the articulated damper and with double action in the case of telescopic lip, has the role of extending or retracting the lip to bring the leveler into operation or of resting position.

The docking levelers are generally found provided with 3 operating positions and in the case of the Autodock leveler, they are represented in Figure 3. The leveler is considered at rest (Figure 3, A) when the platform is found in the horizontal plane, the articulated lip resting on the supports provided on the transverse front profile of the frame. During the resting position the leveler cannot be crossed by the logistics equipment and the hydraulic systems are in a depressurized state [1], [2], [3].

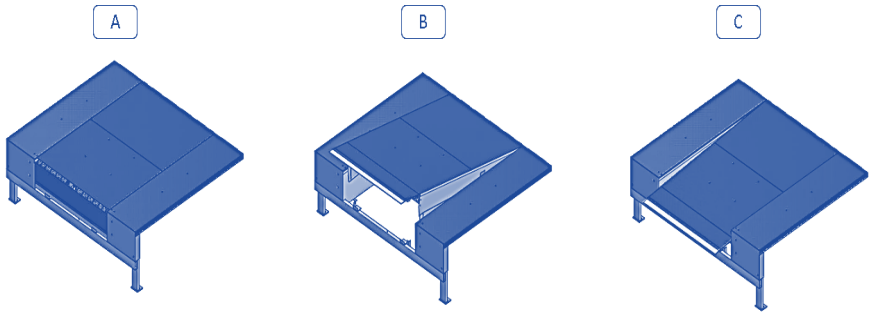


Figure 3
Autodock leveler operating positions

The operating position of the leveler is considered when the lip is in the extended position and supported on the platform of the vehicle visiting the docking station. Two operating positions are considered in the upper position (figure 3, B) when the vehicle level is higher than that of the leveler and lower position (figure 3, C) when the vehicle level is lower than the leveler level. During the operating position, the hydraulic systems are also depressurized, which facilitates the continuous adaptation of the platform according to the level variations of the vehicle platform due to its suspensions.

The hydraulic systems are pressurized in the transition phases of the ramp from the resting position to the operating position and vice versa. It is possible to self–pressurize the hydraulic system due to the accidental fall of the platform due to the disappearance of the possibility of supporting the damper which may be due to the accidental advancement of the vehicle. In the case of self–pressurization, the system is provided with a safety valve that stops the discharge of hydraulic fluid from the two linear hydraulic motors intended to drive the platform.

The operating sequences of the leveler starting from the rest position are raising the platform to the upper maximum point, operating the lip to full extension and lowering the platform until the lip comes into contact with the platform of the vehicle in the docking station. The retraction to the rest position is performed in reverse order.

3 Defining the Lifting System

Taking into account the configuration parameters of the Autodock docking leveler, the aim is to determine the technical characteristics that define the linear hydraulic motors with simple action intended for lifting the platform.

The first technical specification considered is the useful force developed F [kN] its magnitude is given by formula (1):

$$F = \frac{\text{the load capacity of the leveler}}{2} \cdot k_s \cdot k_d \tag{1}$$

where: $k_s - \text{safety factor, considered } 1.5;$

$k_d - \text{dynamic load coefficient, considered } 1.4.$

The coefficients k_s and k_d are in accordance with standard EN 1398:2009 [4].

It is also necessary to determine the dimensional limits of the hydraulic motor such as the minimum length L_{min} [mm], the maximum length L_{max} [mm] and the stroke of the piston St [mm] given in formula 2:

$$St = L_{max} - L_{min} \tag{2}$$

An important feature of the hydraulic motor is also the inner diameter D_i, which is required by the design input data to be 40mm and the outer diameter D_e in the value of 50mm. Another feature to consider is the speed of the piston V_p which can be adjusted from the hydraulic unit by varying the flow of the pump [5], [6].

$$V_p = Q_p \cdot \frac{\pi \cdot D_i^2}{4} \tag{3}$$

where Q_p [l/min] is the pump flow from the hydraulic unit.

The main parameters used in the configuration of the docking ramps are presented in Table 1 where are shown the intervals between which they can be found.

Table 1

Autodock docking leveler configuration parameters

Description	Parameter notation	UM	Range	
			Minim	Maxim
Load capacity of the leveler	–	kN	60	150
Nominal length	NL	mm	2000	4500
Nominal width	NWAD	mm	3300	3750
Nominal height	DH	mm	800	1500
Frame thickness	LH	mm	600	950
Lip width	NW	mm	2000	2500
The length of the lip	LL	mm	350	1000
Maximum opening	Ls	mm	500	200
Minimum opening	Li	mm	400	150

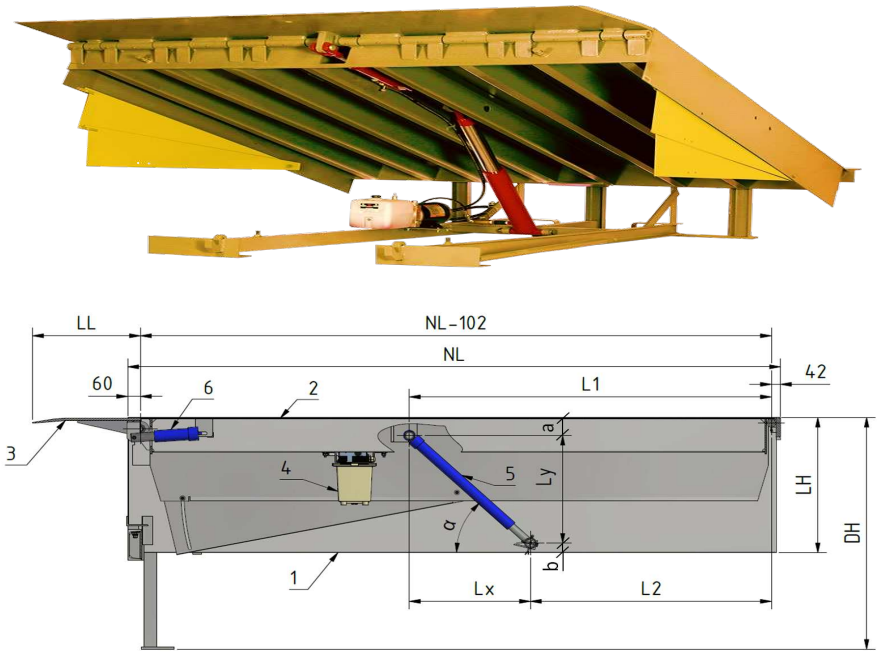


Figure 4

Constructive parameters of the lifting system designed for the Autodock docking leveler

Figure 4 shows the parameters that influence the lifting system of the hydraulic ramp, it can be seen that the hydraulic motor 5 is fixed in two articulated supports forming the angle of inclination α which is considered 45 degrees always when the

leveler is stationary, in this case the other 2 parameters L_1 and L_2 describe the position of the fixing articulated supports with respect to the center of the hinge that forms the rear joint of the leveler. The parameters L_x and L_y are influenced by the length of the hydraulic motor during the rest position.

$$L_x = L_y = LH - a - b \tag{4}$$

Considering the position of the center of gravity extracted according to the 3D model of the platform and lip assembly as 55% of the NL parameter, measured from the center of the rear hinge to the lip, the formulas for determining the fixed dimensions L_1 and L_2 can be considered:

$$L_1 = 0{,}55 \cdot NL \tag{5}$$

$$L_2 = L_1 - L_x \tag{6}$$

It is also possible to calculate the length L of the hydraulic motor in the resting position state of the leveler, this being given by formula 7:

$$L = \frac{L_y}{\sin \alpha} = \frac{L_x}{\cos \alpha} \tag{7}$$

When choosing the type of hydraulic motor it is necessary to respect the condition (8) in order to have a sufficient stroke of the piston necessary when lowering the platform of the docking leveler below the level of the horizontal.

$$L > 0{,}35 \cdot L_{max} \tag{8}$$

At the same time, a 50 mm reserve of the piston stroke is required in the minimum operating position of the leveler. This condition is due to the fact that the fluid supply of the hydraulic motor is made without the existence of an ante filling chamber with fluid, the supply being made by an orifice with a section much smaller than the active surface of the piston. In this case, the possibility of locking the hydraulic motor may occur due to its inability to develop the useful force required to lift the platform.

These dimensional parameters being determined it can be considered that the lifting system of the platform are completely defined from a constructive point of view in the resting position of the docking leveler [7].

4 Mathematical Modeling of the Lifting System

The determination of the maximum opening L_s according to Figure 5, which the linear hydraulic motor can offer for the different configurations of the Autodock leveler is made by elaborating a mathematical model describing the relations between all the parameters that vary during the actuation of the platform.

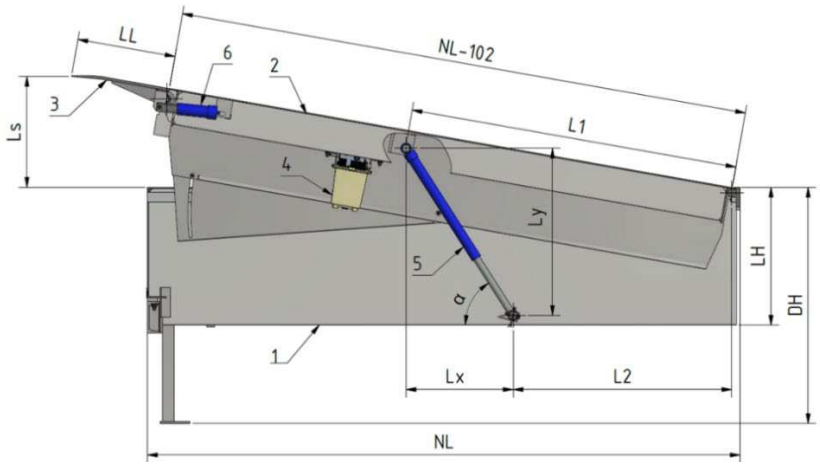


Figure 5

The parameters of the platform lifting system in the upper operating position of the leveler

The main parameters that describe this path of the platform movement are the inclination angle of the hydraulic motor α, the inclination angle of the platform β and the elevation that describes the vertical movement of the lip tip L_s.

Figure 6 shows the diagram describing the kinematic parameters of the route traveled by the platform and the extension of the hydraulic motors that drive it.

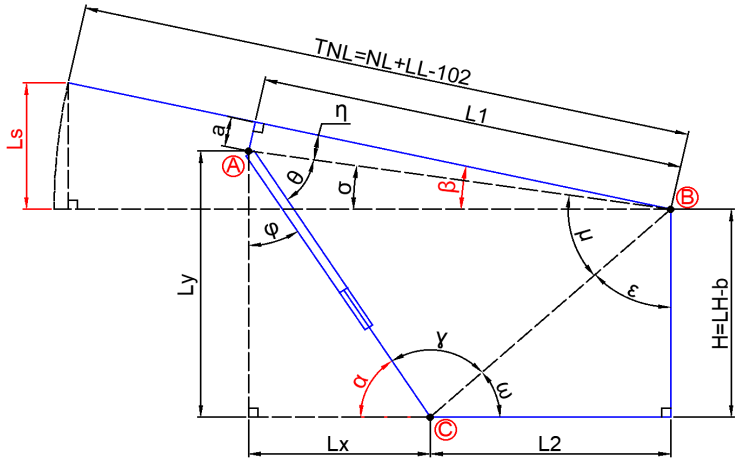


Figure 6

The scheme of the parameters involved in the mathematical model of the platform movement

Points A, B and C are joints that allow the rotational movement, points B and C being fixed and mobile point A, also the constructive parameters H, L1, L2, a and TNL are considered the parameters established according to the ramp configuration and are not considered variables in this analysis of movement.

The order of the scheme for the mathematical determination of each parameter is as follows:

$$BC = \sqrt{L_2^2 + H^2} \tag{9}$$

$$\omega = \cos^{-1}\left(\frac{BC^2 + L_2^2 - H^2}{2 \cdot BC \cdot L_2}\right) \tag{10}$$

$$\varepsilon = 90 - \omega \tag{11}$$

$$\gamma = 180 - \alpha - \omega \tag{12}$$

$$\varphi = 90 - \alpha \tag{13}$$

$$BA = \sqrt{L_1^2 + a^2} \tag{14}$$

$$\theta = \sin^{-1}\left(\frac{BC \cdot \sin\gamma}{BA}\right) \tag{15}$$

$$\mu = 180 - \gamma - \theta \tag{16}$$

The parameter AC represents the length of the extended linear hydraulic motor for a certain value of the parameter α. The additional parameters L_x and L_y can also be determined:

$$AC = BA + BC - \sqrt{2 \cdot BA \cdot BC \cdot \cos\mu} \tag{17}$$

$$L_x = AC \cdot \sin\varphi \tag{18}$$

$$L_y = AC \cdot \sin\alpha \tag{19}$$

$$\eta = \cos^{-1}\left(\frac{L_1^2 + BA^2 - a}{2 \cdot L_1 \cdot ba}\right) \tag{20}$$

$$\sigma = \varepsilon + \mu - 90 \tag{21}$$

Knowing the parameter β that characterizes the inclination of the platform can determine the maximum opening of the leveler L_s.

$$\beta = \sigma + \eta \tag{22}$$

$$L_s = TNL \cdot \sin\beta \tag{23}$$

Knowing all these relationships, it can be considered that the lifting system of the platform has a path that is clearly defined from a mathematical point of view, covering the configuration range of the Autodock docking leveler product [8], [9].

In case of lowering the platform below the horizontal level, the lower limit must be $\alpha > 30^o$. This limitation is due to the avoidance of the appearance of a too small vertical component of the useful force developed by the hydraulic motors, necessary to lift the platform, which leads to the impossibility of lifting the platform. The second risk that may occur is that the horizontal component of the force (much greater than that required to lift the platform) developed by the hydraulic motors has a destructive character by inducing stresses in various frame components or even in the support joints of hydraulic motors.

5 The Balance of the Forces Developed within the System

Regarding the demand developed within the lifting system of the docking platform, the situation is represented schematically in Figure 7. The main force acting during the lifting of the platform is the gravitational force G which is divided on the two linear hydraulic motors, acting on each one separately the component G_L.

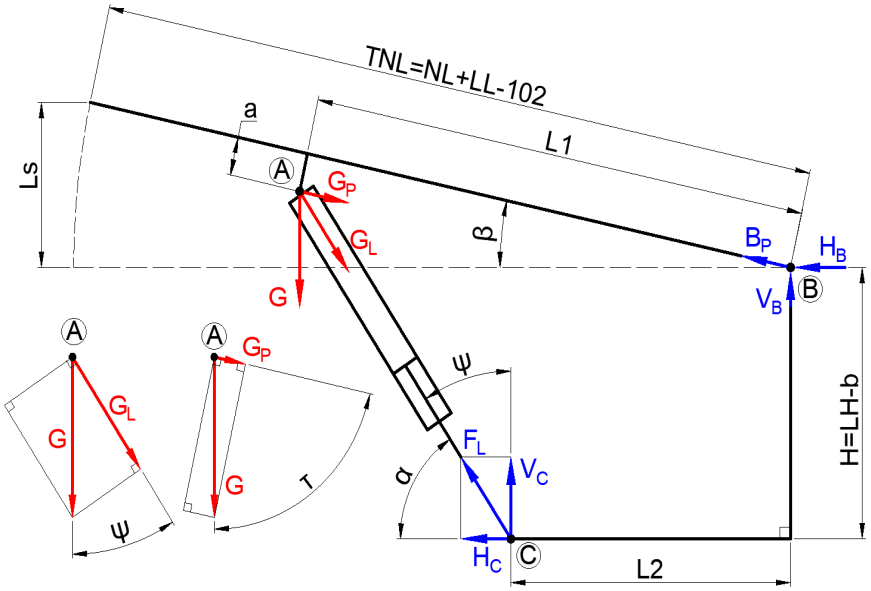


Figure 7
Distribution of forces within the designed leveler lifting system

The reaction forces present in the joints with fixed position in the system, represented by points B and C are highlighted in Figure 7 in blue, among which are components in the vertical direction V_B and V_C respectively components in the horizontal direction H_B and H_C.

The component G_P acting along the platform is compensated by the reaction force B_P which develops in the fixed joint B.

In the case of the component G_L of the force of force G acting on the linear hydraulic motors this is canceled by the reaction force F_L during the stationary stage of the leveler in stationary position when $G_L = F_L$. If the platform performs a lifting movement then the condition $G_L < F_L$ is met and during the descent of the levelers the reverse $G_L > F_L$ takes place [10].

The mathematical determination of the stresses is performed by a series of successive determinations starting with the angles appeared due to the decomposition of the main forces on components:

$$\Psi = 90 - \alpha \tag{24}$$

$$\tau = 90 - \beta \tag{25}$$

Knowing these angles can determine the components of the weight force acting on a single hydraulic motor G_L and G_P acting on the platform on a single hinge of the joint located in the back of a the leveler:

$$G_L = \frac{G \cdot \sin(90 - \Psi)}{2} \tag{26}$$

$$G_P = \frac{G \cdot \sin(90 - \tau)}{2} \tag{27}$$

For the steady state of the platform it is considered that the mobile joint A becomes fixed by respecting the following condition:

$$F_L = G_L \tag{28}$$

The relations by which the magnitude of the vertical and horizontal reaction component forces related to the fixed joint C can be determined are:

$$V_C = F_L \cdot \sin\alpha \tag{29}$$

$$H_C = F_L \cdot \sin\Psi \tag{30}$$

In the case of the fixed joint B the relations that describe the reaction forces developed in it are the following:

$$B_P = G_P \tag{31}$$

$$V_B = B_P \cdot \sin\beta \tag{32}$$

$$H_B = B_P \cdot \sin(90 - \beta\beta) \tag{33}$$

According to these relations, the pressure required in the hydraulic installation for lifting the platform can be determined according to the relation:

$$P_L = \frac{F_L}{D_i};\ F_L > G_L \tag{34}$$

These relations presented in the balance of forces are necessary for the strength sizing calculations of the different constructive components of the joints and other constructive components [11].

The gravitational force G acting on the platform can be found under two scenarios:

1) When the leveler is not subject to external stresses (it is not crossed), the situation of maneuvering the leveler to bring it into different operating positions. In this case, the weight force G is specific to the weight of the platform and the other components that are supported by it.

2) When the leveler fall safety system is activated and it is crossed by the machines. In this case, the force G is considerably higher than in the first case, having also a dynamic stress character where the leveler and the joints together with the lifting system must withstand the stress F in the relation (1).

The dimensioning of the leveler and the joints of the lifting system is performed for scenario 2 and the lifting capacity of the hydraulic motors must be dimensioned according to scenario 1 taking into account a safety factor between 1.2–1.4 in order not to encounter problems in operation.

Conclusions

The efficient flow of products in and out of facilities is critical in today's highly competitive world. Special attention must be given to the loading dock area design for this to happen. A number of factors must be considered when coordinating dock heights and door sizes, and when selecting the proper loading dock equipment. Hydraulic swing–lip dock levellers are designed to enable a safe and efficient loading and unloading while reducing downtime to a minimum. The result is exceptionally high safety for the transfer of goods, preventing any injuries or damage to equipment. Dock levelers bridge the gap and height difference between the dock and the trailer. They also compensate for the up and down float of the trailer bed during loading. They use fully powered raise and lip extension functions with hydraulic cylinders and hydraulic pump and motor stations. Hydraulic levelers are considered the safest loading dock choice.

The process of designing the Autodock docking leveler lifting system starts from knowing the destination of the product and defining its configuration range. For a good functioning of the system, the following constructive conditions are established and imposed:

— The platform of the leveler must be driven by linear hydraulic motors with simple action in the platform center of gravity.

— The angle of inclination α of the hydraulic motors while the leveler is in the resting position must be worth 45 degrees.

— For the sustainability of the lifting system and the leveler, it is necessary to have a reserve of the piston stroke of 50mm for the hydraulic motors that operate the platform.

A mathematical model was developed to determine the variation of the L_s dimension and the β angle as a function of the α angle. The balance of the forces appeared in the lifting system of the platform was made, depending on these, the sizing calculations of the different components of the leveler and from the lifting system will be carried out.

Today's workplace will not tolerate unsafe work practices. Planners must ensure that the loading dock area is not just efficient, but also safe. Installing loading dock safety equipment is just the first step towards minimizing hazardous and costly

accidents. A dock leveler is a fixed bridge designed to permit the safe and efficient flow of goods into and out of a building. In order to accomplish this, a dock leveler must be able to support extremely heavy loads and service a wide range of truck heights. Hydraulic units, although initially more costly, require less routine maintenance than mechanical units, and offer many long–term benefits. Heavy load, high usage, and severe condition applications are best suited to hydraulic dock levelers. Besides increasing efficiency and safety, it also generates energy savings by restricting thermal losses through open doors, ultimately improving hygiene and working conditions.

The innovative and unique docking control system implemented in our unit offers complete control of the dock leveller, dock shelter and door, all in one control unit. With very few self–explanatory buttons on the control unit, the system is easy to operate for all operators. Separate steering (control) units and complex wiring systems are no longer needed. The innovative and unique docking control system offers complete control of the dock leveller, dock shelter and door, through a single control unit.

Lifting system for the docking leveler are designed to enable a safe and efficient loading and unloading while reducing downtime to a minimum. It is the standard solution in general industry applications. Its swing lip safely bridges the gap between the ramp and the vehicle bed. The result is exceptionally high safety for the transfer of goods, preventing any injuries or damage to equipment. Maintenance is easy and fast to secure functionality.

References

[1] Product datasheet dock leveler, ASSA ABLOY, DL6010SA

[2] https://www.hormann.co.uk/industry–commerce–and–public–authorities/loading–technology/dock–levellers, accessed on: 02.02.2021.

[3] https://www.assaabloyentrance.co.uk/en/products/loading–bay–equipment/dock–levelers/, accessed on: 20.11.2021.

[4] European standard, NF EN 1398, Dock levellers safety requirements, 2009

[5] Parr, A.: Hydraulics and Pneumatics: A technician's and engineer's guide, Second Edition, Oxford: Butterworth–Heinemann, 1998

[6] Chapple, P.: Principles of Hydraulic Systems Design, Second Edition, New York: Momentum Press, 2015

[7] Ullman, D. G.: The Mechanical Design Process, Fourth Edition, New York: McGraw–Hill, 2010

[8] Whitney, E. L.: Math Handbookof Formulas, Processes and Tricks, Version 2.3, Editure Reno NV, 2021

[9] Nwokah, O. D. I.; Hurmuzlu, Y.: The Mechanical systems design handbook: modeling, measurement, and control, Editure CRC Press, 2002

[10] Meriam J. L.; Kraige L. G.: Engineering Mechanics Statics, Volume 1, Seventh Edition, Editure John Wiley & Sons, Inc., 2011

[11] Marghitu, D. B.: Mechanical Engineer's Handbook, San Diego: Academic Press, 2001

Optimization Methodology of Thermoelectric Peltier-Modules, for Structural Design and Material Selection, using MCDM and FEM Modelling

Judit Albert and Ágnes Takács

Institute of Machine and Product Design, University of Miskolc
Miskolc-Egyetemváros, H-3515 Miskolc, Hungary
judit.albert@student.uni-miskolc.hu; agnes.takacs@uni-miskolc.hu

Abstract: Monitoring the temperature of medical products is essential, to ensure their safety and stability. The thermoelectric Peltier-modules can be used for either cooling or heating in this application. In this work, different geometries are applied and the resulting performance of the Bi2Te3-based module, characterized via finite element simulations and materials, is analyzed, where the primary objective was to reduce shear and torque stresses, which also depends operating conditions. In the first phase, Finite Element Analysis (FEM) analysis of the thermo-mechanical test unit for the Peltier-modules, is conducted. In the second phase, the best design concept selection phase, using multi-criteria decision theory is introduced, where the ranking of the criteria is weighted by the characteristics assigned to the selected values. The object is to choose the best design concept, from the alternatives generated during the first phase.

Keywords: MCDM methods; VIKOR algorithm; FEM

1 Introduction

Nowadays, the issue of thermoelectric materials is very important, due to the increase of global energy prices, but one of the main problems related to thermoelectric devices is the low efficiency of applications, where their performance is usually determined by the material properties of the constituent materials [1]. Improving the electrical performance, it is also important to extend the life and improve the reliability of the module. Thermal stress is a big problem in electronic systems. Thermal stresses and deformations are the main reasons for the limited lifetime of electronic devices [2]. Among other things, local differences in the coefficients of thermal expansion (CTE) of different materials, can be the cause of failure [3]. Regarding the mechanical failure criteria of Thermoelectric modules (TE-module) [4], the literature widely considers the von Mises stress, as

the general stress, to determine the failure, it uses a FEM to calculate such stresses [5] [6]. Others highlighted the bending and shear stress responsible for mechanical failure [7]. The results of the examination show that the life expectancy can be improved by reducing the maximum thermal stress in the pellets. To verify the mechanical reliability of electronic devices, finite element software is now widely used to save analysis time. Looking at previous literature, most studies discussed the effects of changing parameters on mechanical and electrical performance. In Gao's research [8], he investigated the thermal stress distribution, mechanical properties, and thermoelectric properties of a TE-module using the finite element method. He found that the performance and efficiency of TE-module are not only influenced by material properties (Seebeck coefficient, electrical resistance, and thermal conductivity), but it is also determined by the mechanical properties of TE-module. However, few of these provided a practical way to take into account the values of mechanical and thermal stress on the entire structure, and to use them in the evaluation of new TE-modules constructions, in the initial design phase. Therefore, the aim of this work is to increase the lifetime of TE-module by exploring the potential stress-relieving effects of the geometry design at the initial design stage of TE-module by investigating the overall stresses and deflections. Using the values describing the behavior of the TE-module determined during the simulation of the finite element method (FEM) –the efficiency of the design concepts– was ranked with an interval-based target value VIKOR algorithm.

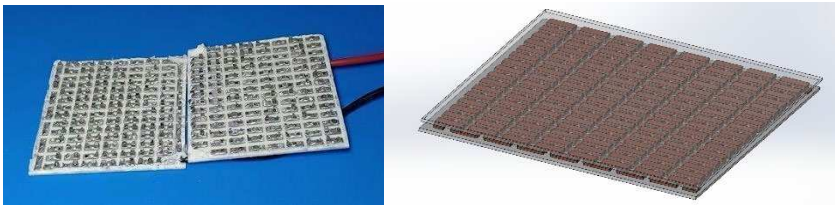


Figure 1

Structure of the TE-module (40x40x3.4 mm) (left side) and 3D model (right side) of the internal structure of the tested TE-module (Peltier 40x40 sealed Quick-Cool Quick-Ohm QC-127-1.4-8.5MS)

2 Research Methodology

The research discusses the extension of the evaluation of design concepts by taking into account weighted performance values, in the early stages of the product development process. In case of conflicting requirements, it is important to analyze the selection of materials, the selection of the design concept, the selection of production processes and the life cycles already at this design stage, since compromise solutions can be found based on the information obtained [9, 10, 11]. During the conceptual design of the development of TE-module, the best design concept is selected by using VIseKriterijumska Optimizacija I Kompromisno

Resenje (VIKOR) [12], which is a multi-criteria decision-making (MCDM) method. In order to compare the design concepts, we recorded the data obtained during the finite element analysis (FEA). The performance of the design concepts was predicted and estimated using the MCDM method, namely the interval-based target value VIKOR method, in order to obtain the concrete values of the individual sub-characteristics. The performance characteristics are the values of the maximum von Mises stress, the maximum deformation, the total deformation energy and the height, volume, surface and mass of the pellets in the TE-module. Finally, the collected data were compared using the VIKOR method, alternatively seven sub-criteria and ten design concepts. The details of each step are described in the following section.

3 Modelling of Thermoelectric Modules

In the early stages of design, numerical modelling techniques can be effectively applied to provide important information about the structural strength of new construction designs. Many literatures use the von Mises stress as a general stress to determine the failure of a TE-module, while others highlight the bending and shear stresses responsible for mechanical failure [6, 7, 8]. In this case, this work begins with the modelling of the TE-module, with the help of the SolidWorks three-dimensional (3D) computer design software and then with the finite element solver of the program, we model new designs with the operating conditions and analyses the obtained results. In addition to primary finite element analysis confirmed that soldered joints and pellets are components of the structure that suffer from plastic deformation and non-linear effects.

3.1 Thermoelectric Modules

The modelled thermoelectric module (Peltier 40x40 sealed Quick-Cool Quick-Ohm QC-127-1.4-8.5MS) contains 254 pellets of size 1.4x1.4x0.6 mm, the material is bismuth telluride (Bi_2Te_3). Based on the preliminary tests of the deformation and stress contours of this module, new pellet geometric designs were developed. Where the effect of increasing the height of TE-module pellets was investigated to determine their thermomechanical behavior under the same operating conditions. Initially analyzed results showed that the geometry of the pellets, and thus the change in their load, can have a significant effect on the von Mises stress, shear stress and total displacement arising in the TE-module. The distance between adjacent pairs of rows and columns between the pellets of the model is 0.5 mm. Figure 1 shows the structure and model of the TE-module, and Table 1 shows the dimensions of the elements for the TE-module. The 3D models of both the original and new designs were created to determine stresses (von Mises and shear) and total displacement. Previously analyzed results have shown that changing the load of

semiconductor thermocouples, “pellets”, has a significant effect on the von Mises stress, shear stress and total displacement. The distance between adjacent pairs of rows and columns between the pellets of the model is 0.5 mm. Figure 2 shows a model of the TE-module.

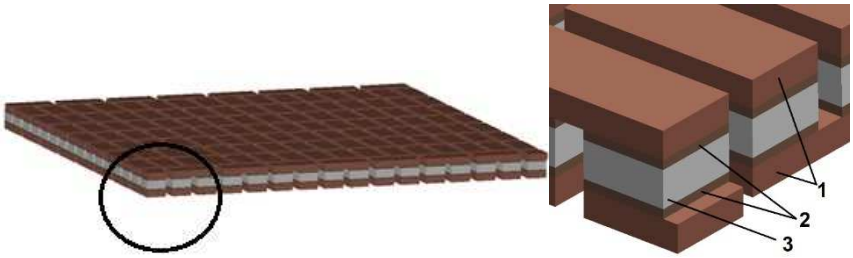


Figure 2
The structural elements of the TE-module: without the ceramic cover plate (left side), the elements of the TE-module and their enlarged detail (right side): 1: Cu elements, 2: SnPb solder layer, 3: Bi_2Te_3 semiconductor pellet

3.2 Material Properties

The modelled TE-module contain top and bottom ceramic cover plates, copper solder connectors and bismuth telluride (Bi_2Te_3) pellets. Material properties of the components at room temperature are listed in Table 2 and Table 3.

3.3 Boundary Conditions and Parameters used in the Numerical Simulation

Numerical simulation was performed using the finite element functions of SolidWorks. For the numerical simulation of the module behavior, the TE-module was placed between two rigid surfaces to perform the calculations, since under operating conditions, the load in the pellets continues to increase due to temperature changes, which hinders the thermal expansion of the TE-module, caused by the rigid surfaces under the effect of pressure. In the case of the examined TE-modul, a compressive load of 1200 kPa was applied as a load [13], specified on the surface of the aluminum oxide ceramic carrier on the warm side. During the constructional modelling of the TE-module, due to the sliding contact between the aluminum oxide ceramic and the interlocking surfaces, we characterized the contact with a friction coefficient of 0.6. For the numerical simulations, a temperature difference of 25°C was prescribed on the external aluminum oxide surfaces of the module, keeping the surface on the cold side at 20°C and applying 45°C on the warm side. Figure 3 illustrates the fit into the main structure. We present all the analyzed results for the TE-module to illustrate all the mechanical behavior and displacements during the application of the prescribed clamping force.

Table 1
Dimensions of the elements of the modelled TE-module

Elements	**Size (mm)**
Ceramic, (Electric insulator), Al_2O_3	40x40x0.8
Copper, (Electric conductor), Cu	3.3x1.4x0.45
Solder layer, SnPb	1.4x1.4x0.15
Pellets, Bi_2Te_3	1.4x1.4x0.6

Table 2
Material properties of the elements of the modelled TE-module

Material property	**Ceramic (Electric insulator), Al_2O_3**	**Copper (Electric conductor), Cu**	**Solder layer, SnPb**
Coefficient of thermal expansion, CTE, 50-190°C (10^{-6} K^{-1})	4.89-6.03	16.7-17.3	27.0
Density, ρ (kg/m^3)	3970	8940	7260
Young's modulus, *E* (GPa)	380	115	44.5
Poisson's ratio, μ	0.26	0.31	0.33

Table 3
Properties of isotropic bismuth telluride used for simulations [14]

Material property	**p-type**	**n-type**
Coefficient of thermal expansion, CTE, 50-190°C (10^{-6} K^{-1})	16.8	16.8
Density, ρ (kg/m^3)	6858.7	7858.7
Young's modulus, *E* (GPa)	47	47
Poisson's ratio, μ	0.4	0.4

3.4 Results

3.4.1 Stresses and Displacements

Simulation results presented in Table 4, the maximum and minimum values of the von Mises stress are given for a temperature difference of 25°C and a compressive load of 1200 kPa.

Based on the results of simulations, it can be said that the maximum stress is not always found in the same pellet element, but always at the corner of a pellet, at the point of contact with the welding material as illustrated in Figure 4. By reducing the temperature interval, the maximum mechanical stress level also decreases.

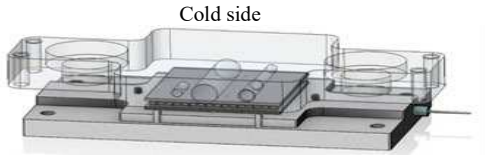


Figure 3
The assembled substructure

Table 4
The von Mises stress in the pellets (MPa) in the TE-module in the case of a maximum permissible load 350 N and the temperature difference of 25°C between the hot and cold side

Pellets height L (mm)	**von Mises stress (MPa)**	
	Min.	**Max.**
0.6	1.113	40.088
0.8	0.856	37.237
1	0.767	37.855
1.2	0.599	36.183
1.4	0.574	34.278
1.6	0.483	35.101
1.8	0.467	34.587
2	0.418	32.106
2.2	0.437	30.972
2.4	0.312	29.195

The ductile brazing alloy undergoes ductile deformation when the stress level exceeds the yield strength. This deformation reduces the stresses arising in the pellets. So, if Pb-Sn alloys were used instead of the more mechanically resistant Sn-Sb brazing alloy, the yield strength would be lower and the maximum stress level in the pellets would further decrease [14]. At the corners of the copper elements close to the ceramic surface, there are locally high stress regions, as large stress gradients develop in these areas, due to the difference between the coefficients of thermal expansion of the copper and the pellet. In addition, the tension decreases from the high stress area all the way to the base surface of the clamping element.

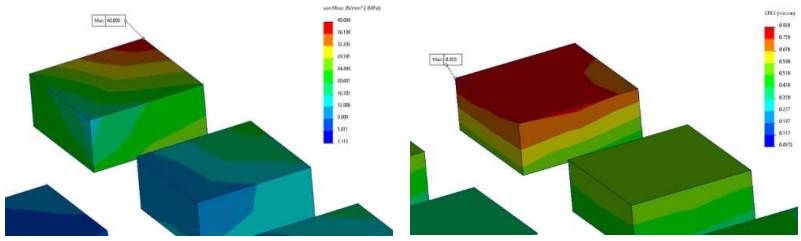


Figure 4
The von Mises stress (MPa) (left side) and displacement (micron) (right side) arising in the pellets (L=0.6 mm), in case of a temperature difference of 25°C between the hot and cold side in the TE-module

3.4.2 Summary of Results

This study examines the feasibility of improving the structural integrity of TE-module. For this purpose, FEA was carried out, where the deformation results a significant increase in the case of large temperature gradients, so it can be recommended that the thermoelectric module be installed with a damping spring, reducing the stresses arising in the elements of the general construction, which allows extending the life of the TE-module. With the use of the damping spring in the models, a significant reduction of stresses and total deformation was observed at the corners of the copper elements close to the ceramic surface, with lower von Mises stresses, shear stresses and smaller TE-module total deformation. In all cases, the maximum stresses occurred around the edge of the copper elements, near the ceramic plate. The higher stress levels around the corner zones were caused by stress concentrations. The mechanical stress distribution in the cross-sectional areas of the module varies uniformly. Based on the results, it is shown that with the variable pellet height in the design concepts, the maximum stress level and the total deflection can be reduced by changing the geometrical design and accordingly the mechanical performance/lifetime of the thermoelectric module can be improved.

4 Ranking of Design Concepts with Interval-based Target Value VIKOR Method

4.1 The Interval-based Target Value VIKOR Method

The interval-based target value VIKOR method is a possible solution to multi-attribute decision problems (MADM). To apply the VIKOR method, denote the number of criteria n and the number of materials that can be considered m.

Be it also $[x_i^L, x_i^U]$, the interval for the j^{th} criteria of the i^{th} material, $i = 1, \ldots, m$,

$j = 1, \ldots, n$

To make a decision we need a target value for what the ideal material would look like for us, so let's mark the target values for each characteristic $T_1, T_2, \ldots, T_n$, furthermore we also need a weighting of how important it is to be close to the target value for each characteristic, so let's mark the weights for each characteristic

$w_1, \ldots, w_n$, where:

$$w_j \geq 0, \; j=1, \ldots, \text{n and } \sum_{j=1}^{n} w_j = 1 \tag{1}$$

If a criterion is to be maximized or minimized (e.g., minimize cost), the maximum or minimum of the data for that criterion can be chosen as the target value. With the VIKOR algorithm, we determine for each material, one $[Q_i^L, Q_i^U]$, $i = 1, \ldots, m$ interval, which collectively indicates how far the given material falls from the target

value, then a ranking can be established by comparing intervals in pairs. Next, we normalize the data. For this, let's introduce the following:

$$x_j^{L\,min} \coloneqq \min\{x_{ij}^L : i = 1, \dots, m\} \tag{2}$$

$$x_j^{U\,max} \coloneqq \max\{x_{ij}^U : i = 1, \dots, m\} \tag{3}$$

notations where j=1, …, n. These indicate the minimum and maximum values available for each criterion among the materials. We will use this, in addition to the target T_j, to determine the range of the data and use it to normalize the deviation from the target value. Let's introduce next:

$$V_{ij}^L \coloneqq \frac{\left|x_{ij}^L - T_j\right|}{\max\{x_j^{U\,max},\, T_j\} - \min\{x_j^{L\,min},\, T_j\}} \in [0,\, 1] \tag{4}$$

$$V_{ij}^U \coloneqq \frac{\left|x_{ij}^U - T_j\right|}{\max\{x_j^{U\,max},\, T_j\} - \min\{x_j^{L\,min},\, T_j\}} \in [0,\, 1] \tag{5}$$

where i = 1, . . ., m, j = 1, . . ., n normalized quantities, the two quantities differ in that the target value for each criterion is compared with the lower or upper endpoint of the interval, so that characteristics moving on a larger scale do not excessively distort the results.) The following indicator numbers can then be calculated for each material:

$$S_i^L \coloneqq \sum_{j=1}^{n} w_j \;\; \min\{V_{ij}^L\,,\, V_{ij}^U\} \tag{6}$$

$$S_i^U \coloneqq \sum_{j=1}^{n} w_j \;\; \max\{V_{ij}^L\,,\, V_{ij}^U\} \tag{7}$$

$$R_i^L \coloneqq \max\{\min\{V_{ij}^L\,,\, V_{ij}^U\} : j = 1,\; \dots,\, n\} \tag{8}$$

$$R_i^U \coloneqq \max\{\max\{V_{ij}^L\,,\, V_{ij}^U\} : j = 1,\; \dots,\, n\} \tag{9}$$

$$\left[\frac{S_i^L - S^-}{S^+ - S^-}, \frac{S_i^U - S^-}{S^+ - S^-}\right] \tag{10}$$

$$S^+ = max\{S_i^U :\; i = 1,\; \dots,\, m\} \tag{11}$$

$$S^- = \min\{S_i^L :\; i = 1,\; \dots,\, m\} \tag{12}$$

interval, and

$$\left[\frac{R_i^L - R^-}{R^+ - R^-}, \frac{R_i^U - R^-}{R^+ - R^-}\right] \tag{13}$$

$$R^+ = \max\{R_i^U \coloneqq 1,\; \dots,\, m\} \tag{14}$$

$$R^- = \min\{R_i^L \coloneqq 1,\; \dots,\, m\} \tag{15}$$

it expresses individual regret. The two intervals can be combined according to how important we consider the individual indicators to be $v \in [0, 1]$ the weight of the majority of criteria (where $v = 0.5$ expresses a compromise solution). Then:

$$[Q_i^L, Q_i^U] = \left[v\frac{S_i^L - S^-}{S^+ - S^-} + (1-v)\left(\frac{R_i^L - R^-}{R^+ - R^-}\right),\ v\frac{S_i^U - S^-}{S^+ - S^-} + (1-v)\left(\frac{R_i^U - R^-}{R^+ - R^-}\right)\right] \quad (16)$$

indicates how far the material is from the target value. The minimum of these intervals is chosen to rank the materials. When comparing the intervals, we enter the last free parameter in the algorithm, the parameter α is the optimism level of the decision maker. The optimistic decision maker is characterized by larger α values, while the rational decision maker is characterized by $\alpha = 0.5$. The value of v lies in the range of 0-1 and these strategies can be compromised by $\nu = 0.5$ according to the suggestion of literature [16].

4.2 Ranking of Design Concepts

With the VIKOR selection method, we can specify an interval for each criterion of the design concepts, which describes the values between which, the given criteria of the given concept moves. This case study contains interval data, including language terms and target criteria. Here, we use an 11-point scale (Table 5) in order to better understand and display the quality criteria, as well as to convert the linguistic expressions into appropriate numbers.

Table 6 shows the weighting of the criteria. Determination of the most favorable values (target criteria) for each criterion:

Criterion 1: Pellet surface, (mm^2)

Criterion 2: Weight of pellets, (g)

Criterion 3: Displacement of pellets, (mm)

Criterion 4: Pellet volume, (mm^3)

Criterion 5: von Mises stress arising in pellets, (MPa)

Criterion 6: Pellet height, (mm)

Criterion 7: Total deformation energy arising in the pellets.

For this case, for all evaluated criterion, the lowest the value, the better. The goal is to select the most suitable concept based on the criteria, and to set up a ranking among the concepts, which one is the most appropriate based on the criteria examined. (Table 6) Table 7 presents the values of the criteria used for the evaluation and obtained during the simulations. Table 8 presents the relative importance of the criteria and the target values. Table 9 presents the ranking based on the criteria. Results presented in Table 9 show that the concept created with 1 mm high pellets received the highest ranking, and the design concepts implemented with 0.6 mm and 1.2 mm high pellets were the next best designs, i.e., these design

concepts were close to the ideal solution based on the VIKOR method. We found that for the best-performing design concept in the TE-module, the cross-sections of the pellets were closer to the optimal ratio, the pellet height increased and the contact resistance decreased.

Table 5
Criteria value in 11-point scale format

Quality value of the material selection factor	
Exceptionally low	0.045
Extremely low	0.135
Very low	0.255
Low	0.335
Below average	0.410
Average	0.500
Above average	0.590
High	0.665
Very high	0.745
Extremely high	0.865
Exceptionally high	0.955

Overall, in the initial stage of the construction process, the selection evaluation supported by the interval-based target value VIKOR method of the proposals for changing the construction geometry, in this case the size of the pellets, in the design concepts of the TE-module, accelerates and supports the early design process of constructions.

Table 6
Comparison and quantitative value of the investigated criteria

Criteria	**Total**	**Relative importance, w_j**
1 Pellet surface (mm^2)	6	0.071
2 Pellet weight (g)	9	0.107
3 Pellet displacement (mm)	12	0.142
4 Pellet volume (mm^3)	9	0.107
5 von Mises stress in pellets (MPa)	15	0.178
6 Pellet height L (mm)	17	0.202
7 Total deformation energy in pellets	16	0.190

Table 7
The values of the criteria used for the evaluation and obtained during the simulations

Pellet Height L (mm)	Pellet surface (mm^2)	Pellet volume (mm^3)	Pellet weight (g)	Total Deformation Energy in pellets Max.	von Mises stress in pellets Max. (MPa)	von Mises stress in pellets Min. (MPa)	Pellet Displ. Max. (mm)	Pellet Displ. Min. (mm)
0.6	7.28	1.18	0.01	1.5784E-06	40.088	1.113	0.838	0.0372
0.8	8.4	1.57	0.01	2.0629E-06	37.237	0.856	0.896	0.0374
1	9.52	1.96	0.01	1.2117E-06	37.855	0.767	0.948	0.0368
1.2	10.64	2.35	0.02	1.3458E-06	36.183	0.599	1	0.0361
1.4	11.76	2.74	0.02	1.3323E-06	34.278	0.574	1.04	0.0326
1.6	12.88	3.14	0.02	1.5202E-06	35.101	0.483	1.1	0.0307
1.8	14	3.53	0.02	1.6015E-06	34.587	0.467	1.15	0.0295
2	15.12	3.92	0.03	1.5710E-06	32.109	0.418	1.19	0.0252
2.2	16.24	4.31	0.03	2.4016E-06	30.972	0.437	1.22	0.0233
2.4	17.36	4.7	0.03	2.3845E-06	29.195	0.312	1.26	0.0219

Table 8
Determining the relative importance of criteria

Criteria	1 Pellet surface (mm^2)	2 Pellet weight (g)	3 Pellet displacement (mm)	4 Pellet volume (mm^3)	5 von Mises stress in pellets (MPa)	6 Pellet height L (mm)	7 Total deformation energy in pellets
Relative importance of criteria, w_j	0.071429	0.1071429	0.1428571	0.1071429	0.1785714	0.202381	0.190476
Target value, T_j	0.6	7.28	1.18	0.01	0.312	0.0219	1.2117
Max, $x_j^U max$	2.4	17.36	4.7	0.03	40.088	1.26	2.4016
Min, $x_j^L min$	0.6	7.28	1.18	0.01	0.312	0.0219	1.2117

Conclusions

Previous publications [10] [11] have concentrated on VIKOR decision methods. This study adds a systematic analysis to the foregoing, while at the same time, basing the decision maker's decision process, on two different supporting methods. First, a brief overview of the relevant design process for TE-module design concepts was described.

Table 9
Ranking of design concepts

Weighting for the strategy α	**0.5**	**Optimism level *v***	**0.5**				
Pellet Height L (mm)	**S_i^L**	**S_i^U**	**R_i^L**	**R_i^U**	**Q_i^L**	**Q_i^U**	**Rank**
0.6	0.064797	0.3706723	0.3081772	1	0.0552567	0.6730919	2
0.8	0.176903	0.4805804	0.7153542	0.9283236	0.380453	0.6892103	7
1	0.075817	0.3912661	0.2222222	0.9438606	0.0062357	0.6486562	1
1.2	0.185655	0.5029667	0.5	0.9018252	0.2469637	0.6848438	3
1.4	0.218479	0.534461	0.5	0.853932	0.2655384	0.6718777	4
1.6	0.283913	0.614117	0.5568182	0.8746229	0.3390932	0.7302556	5
1.8	0.332329	0.668666	0.6676136	0.9111542	0.4377167	0.7846088	6
2	0.415764	0.7484379	1	1	0.698609	0.8868662	9
2.2	0.584168	0.9168668	1	1	0.7939075	0.9821786	8
2.4	0.61631	0.9483594	1	1	0.8120963	1	10

This is followed by Section 2, which describes the distinguishing features of the steps of the adequacy assessment process. Section 3 of this work presents the research steps and the applied materials. The authors used interval-based target value VIKOR analysis, to analyze the case study results. The study illustrates, that individual assessments, by experts, are very beneficial in the evaluation process.

References

[1] Biswas K., He J., Blum I. D., Wu C. I., Hogan T. P., Seidman D. N., Dravid V. P., Kanatzidis M. G.: High-performance bulk thermoelectrics with all-scale hierarchical architectures. Nature. 2012, doi: 10.1038/nature11439

[2] Lee, C. C., Chiang, K. N.: Thermal Stress–Induced Interfacial Failure Modes of Advanced Electronic Devices. In: Hetnarski, R. B. (eds) Encyclopedia of

Thermal Stresses. Springer, Dordrecht. 2014, doi: 10.1007/978-94-007-2739-7_267

[3] E. Suhir: Thermal Stress Failures in Electronics and Photonics: Physics, Modelling, Prevention, Journal of Thermal Stresses, 36:6, 2013, pp. 537-563

[4] Zoui, M. A.; Bentouba, S.; Stocholm, J. G.; Bourouis, M.: A Review on Thermoelectric Generators: Progress and Applications. Energies 2020, pp. 3606

[5] A. S. Al-Merbati, B. S. Yilbas, A. Z. Sahin: Thermodynamics and thermal stress analysis of thermoelectric power generator: Influence of pin geometry on device performance, Applied Thermal Engineering, Volume 50, Issue 1, 2013, pp. 683-692

[6] U. Erturun, K. Mossi: A feasibility investigation on improving structural integrity of thermoelectric modules with varying geometry, in Proceedings of the ASME 2012 Conference on Smart Materials, Adaptive Structures and Intelligent Systems, September 2013, pp. 939-945

[7] K. N. Subramanian: Lead-Free Solders: Materials Reliability for Electronics, John Wiley & Sons, New York, NY, USA, 1st edition, 2012, pp. 1-520

[8] Gao J.-L., Du Q.-G., Zhang X.-D.: Jiang X.-Q.: Thermal stress analysis and structure parameter selection for a Bi_2Te_3-based thermoelectric module, Journal of Electronic Materials, Vol. 40, No. 5, 2011, pp. 884-888

[9] Prasad, B.: Concurrent Engineering Fundamentals, Vol I.: Integrated Product and Process Organization. doi: 10.13140/RG.2.1.2613.0005, 1996, pp. 1-132

[10] J. Albert, Á. Takács: The VIKOR Algorithm in Material Decision Support, DESIGN OF MACHINES AND STRUCTURES 12: 2022, pp. 5-13

[11] J. Albert, Á. Takács: Application aspects of the VIKOR algorithm in material selection decisions, GÉP 71, 7-8, 2020, pp. 65-68

[12] S. Opricovic: Multicriteria Optimization of Civil Engineering Systems. PhD Thesis, Faculty of Civil Engineering, Belgrade, 1998, 302 p.

[13] http://vijaydeep.in/wp-content/uploads/VG1.5Pg218-to-220.pdf

[14] TH. CLIN, S. TURENNE, D. VASILEVSKIY, R. A. MASUT: Numerical Simulation of the Thermomechanical Behavior of Extruded Bismuth Telluride Alloy Module, DOI: 10.1007/s11664-009-0756-9

[15] A. Jahan, K. L. Edwards, M. Bahraminasab: 4 - Multi-criteria decision-making for materials selection, Multi-criteria Decision Analysis for Supporting the Selection of Engineering Materials in Product Design (Second Edition), Butterworth-Heinemann, 2016, pp. 63-80

[16] Sayadi MK, Heydari M, Shahanaghi K. Extension of VIKOR method for decision making problem with interval numbers. Appl Math Model 2009; https://doi.org/10.1016/j.apm.2012.10.002

The Effect of Layer Thickness and Orientation of 3D Printed Workpieces, on The Micro- and Macrogeometric properties of Turned Parts

Gábor Kónya[1*], Péter Ficzere[2]

[1]Department of Innovative Vehicles and Materials, GAMF Faculty of Engineering and Computer Science, John von Neumann University, Izsáki út 10, H-6000 Kecskemét, Hungary; konya.gabor@nje.hu

[2]Department of Railway Vehicles and Vehicle System Analysis, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary; ficzere.peter@kjk.bme.hu

Abstract: 3D printing technologies have developed significantly over the last 30 years, having a major impact on all segments of today's industry. Additive Manufacturing (AM) offers the possibility to produce both prototype and finished parts, reducing product development time and costs while producing higher quality results. However, producing high precision and quality surfaces, such as threads, is still difficult with 3D printing technologies. To eliminate these problems, there are already machines that combine some form of the 3D printing process and of subtractive manufacturing technology. In this type of production, extra material is always left on the surface of the workpiece during the printing process, and functional surfaces are created by removing this extra material. In this scientific study, the authors have dealt with the effect of layer thickness and orientation of the 3D printed workpiece, for the micro- and macro-geometric properties, on the printed and the machined (turned) parts. The authors measured the surface roughness, the deviation from nominal size and determined the cylindricity, after printing and machining. During turning, the effects of printing orientation and chip formation process were investigated. The aim was to investigate the effect of the orientation and layer thickness of the printed objects, on the quality and cylindricity of the final turned parts.

Keywords: 3D printing; PLA; turning; cylindricity, surface roughness

1 Introduction

3D printing has revolutionized the automotive industry in several ways. By allowing for rapid prototyping, companies can quickly test and refine their designs, which results in faster time-to-market and reduced costs [1]. Additionally,

the ability to produce high-quality parts with complex geometries and internal structures has led to increased design freedom, enabling the creation of lighter and more efficient components. This, in turn, has made it possible to develop new vehicles that are lighter, more efficient, and have improved performance characteristics. Overall, the use of 3D printing technology in the automotive industry has resulted in a more flexible, efficient, and cost-effective manufacturing process [2] [3].

Utilizing Additive Manufacturing (AM), also known as 3D printing, which produces parts layer by layer based on models created with Computer Aided Design (CAD), it is possible to design geometries that would be difficult or even impossible to realize using traditional manufacturing methods [4]. These technologies allow greater design freedom, such as generative design, which allows conventional parts to be replaced by lighter parts with identical or improved strength characteristics. In addition, the option to print parts from multiple materials has opened up new avenues for innovation, allowing manufacturers to produce composite parts [5] that offer improved performance and durability [6] [7]. In conclusion, 3D printing has had a profound impact on the automotive industry, enabling the development of lighter, safer, and more environmentally friendly vehicles that offer improved performance and efficiency [8].

There are numerous papers comparing 3D printing processes, but the most comprehensive study was carried out by Hanon et al. According to this study, FDM (Fused Deposition Modelling) was found to be the most favorable 3D printing process based on several factors, including accuracy, printable size, post-processing, number of raw materials, machine size, and price of the machine [9]. While FDM may not have the highest accuracy compared to other 3D printing processes, it can still be an excellent choice for some applications. For example, if precise and aesthetically outstanding finishes are not a requirement, and functional surfaces can be post-machined, FDM can provide an affordable and efficient solution. Based on the findings of Hanon's paper, FDM 3D printing technology was selected for this study because it is relatively cheap compared to other 3D printing processes and has a wide range of available materials, therefore, it is widely used by companies [10].

In FDM 3D printing, the thermoplastic polymer filament is melted inside the extruder and then extruded through a nozzle onto the build platform. The extruder head moves along the "x-y" axes to create a cross-sectional layer of the model, and the platform is lowered in the "z" direction after each layer is completed. This process is repeated until the entire model is built up layer by layer, as shown in Fig. 1 [11] [12].

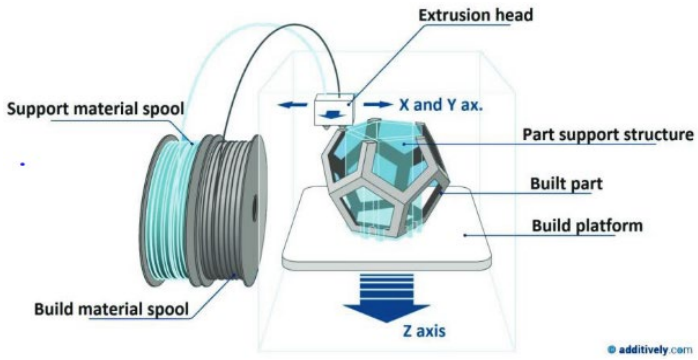


Figure 1
Printed method of FDM 3D printing

However, 3D printing processes are unable to print with the accuracy and surface quality that is a necessity in cases where high precision is required, such as a sealing surface. Even options that can be modified during pre-processing are not sufficient to improve these factors [13]. This is also typical for plastic and metallic parts, which often require post-machining [14]. Research work on PLA machining cannot be found in the literature. However, other engineering plastics, such as ABS or POM-C were turned [15] [16]. In both cases, the authors found that higher cutting speed and medium depth of cut are the more important when turning these materials.

This paper investigates the effects of layer thickness and orientation of the parts on the surface roughness, cylindricity and turning process. The aim is to determine what layer thickness and orientation is appropriate for printing products whose functional surfaces will be finished by turning.

2 Methodology

2.1 3D Printer and Printing Technology

The Prusa i3 is a popular and widely used 3D printer known for its reliability, affordability, and ease of use. Printing with PLA (Polylactic Acid) is a common choice for many 3D printing applications due to its good dimensional stability, low shrinkage, and ease of printing. PLA is a thermoplastic that is derived from renewable resources and is known for producing high-quality prints with a smooth surface finish [17]. The properties of PLA can vary depending on the specific manufacturer, but the ranges listed in Table 1 should give a good idea of the material's general characteristics.

Table 1
Property ranges for PLA materials [18] [19]

Properties	**PLA**
Tensile strength (MPa)	15.5-72.2
Tensile modulus (GPa)	2.020-3.550
Elongation at break (%)	0.5-9.2
Flexural strength (MPa)	52-115.1
Flexural modulus (GPa)	2.392-4.930
Printing temperature (°C)	190-220
Printing speed (mm/s)	40-90

By printing a total of 8 cylindrical test pieces with different orientations, you can gain a better understanding of the effects of orientation on the mechanical properties of the printed parts, therefore 4-4 Ø20X50 mm workpieces were printed with 0.05, 0.1, 0.2 and 0.4 mm layer thickness, as shown in Fig. 2.

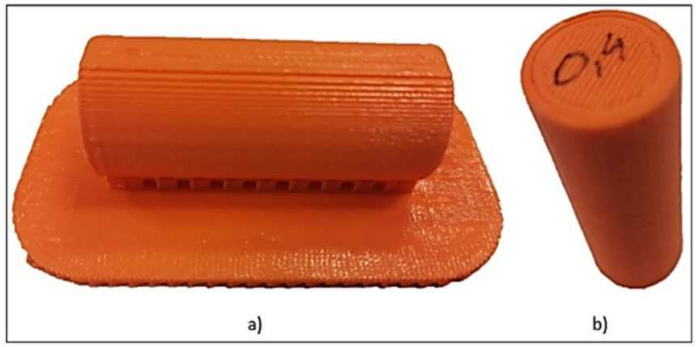


Figure 2
Printed workpieces: a) horizontal, b) vertical

The use of a CAD software like Solid Edge to model the cylinder and export an STL file is a common practice in 3D printing. The STL file format is widely used in 3D printing as it describes a 3D object as a series of triangles that make up the surface of the object. The angle subtended by the planes and tolerance specified during the export process can affect the overall accuracy and quality of the printed part. During the exporting, the angle subtended by the planes was 3° and the tolerance was 0.05 mm. Choosing the right structure is very important as it can influence not only the geometric dimensions but also the material properties [20]. Printing was done with a 30% density gyroid type fill, because it is known for its good mechanical properties and is often used in high-strength applications. [21].

When printing parts in a horizontal orientation, it is often necessary to use support material to ensure the printed part has adequate stability during the printing process. The support material helps to hold up the overhanging or cantilevered portions of the part, preventing them from collapsing or warping during printing.

The design of the support material is a crucial factor, as it can affect the overall quality and accuracy of the printed part. Therefore, at places with up to 40° inclination support material was utilized. This is shown with the gyroid infill in Fig. 3.

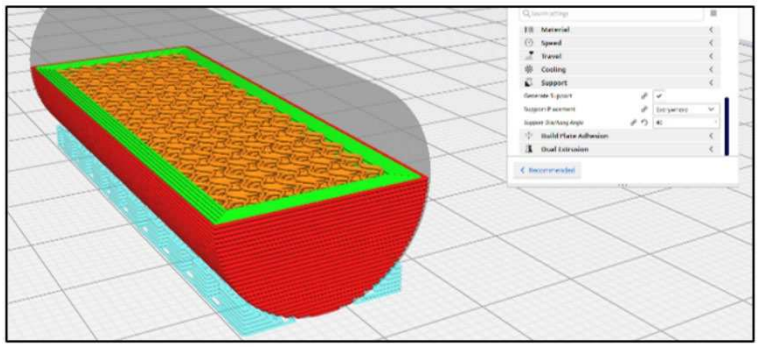


Figure 3
Production planning for the horizontally oriented workpieces
Note: horizontal orientation, 0.4 mm layer thickness

In order to carry out the turning experiments, the number of top and bottom layers of horizontal oriented pieces has been increased, otherwise the infill part would start earlier and the workpiece would not conform during machining. This can also cause errors, as the concentricity of the walls during printing can only be ensured layer by layer. For the unsupported parts, where the wall is supposed to be sufficiently steep, it was necessary to use active cooling because without it the layers became misaligned, as shown in Fig. 4.

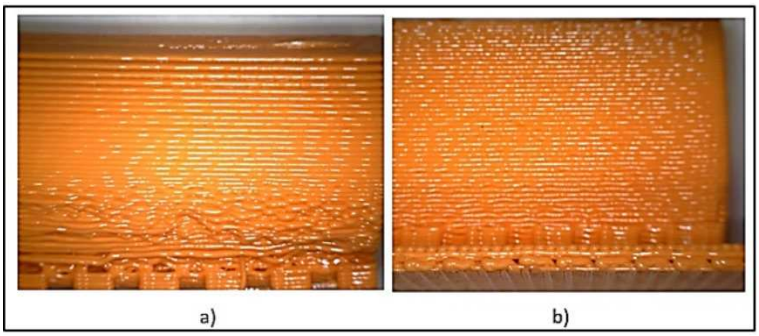


Figure 4
Printed part a) without active cooling, b) with active cooling
Note: 0.4 mm layer thickness, horizontal orientation

The printing parameters are listed in Table 2.

Table 2
Printing parameters

Properties	**PLA**
Layer thickness (mm)	0.05; 0.1; 0.2; 0.4
Wall thickness (mm)	2.5
Filling density (%)	30
Printing temperature (°C)	215
Printing speed (mm/s)	40
Fill printing speed (mm/s)	40
Active cooling	-

2.2 Cutting Parameters

After the surface roughness and cylindricity measurement of the 3D printed parts, each part was turned with the following parameters: cutting speed was v_c = 100 m/min, feed rate was v_f = 0.3 mm/rev. and the depth of cut was varied because the aim was to achieve a workpiece diameter of 18 mm in order to make the dimensional accuracy comparable. The cutting experiments were carried out on the NCT Euroturn 12 CNC lathe. A PCD (polycrystalline diamond) insert was used for the machining, as illustrated in Fig. 5.

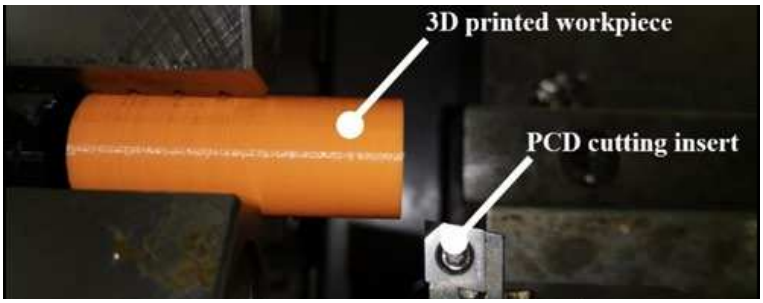


Figure 5
Experimental setup for turning

2.3 Surface Roughness Measurement

Measuring the surface roughness of the printed parts is an important step in evaluating the quality of the printed parts. A Mitutoyo Formtracer SV-C3100 tactile roughness tester was used for measuring the surface roughness of the workpieces according to MSZ EN IS 4287:2002 and the results were evaluated in Microsoft Excel. This information can help to determine which parameters have a notable influence on the average surface roughness (R_a).

2.4 Cylindricity Measurement

For cylindricity, each point on the real cylindrical surface must be located between two coaxial cylindrical surfaces with a radius difference of the specified tolerance [22], as shown in Fig. 6. Cylindricity was measured in a prism using a Mitutoyo 543-270B dial indicator with an accuracy of 0.01 mm, assessing the dimensional deviation from the nominal size. The Roundness Measurement System could not to be used because the dimensional deviation was too large. The measurement was taken from the end of the workpiece in three planes at 5, 10 and 15 mm. Four measurements were taken in each plane at 0°, 90°, 180° and 270°, as illustrated in Fig. 7. and the results were evaluated in Microsoft EXCEL.

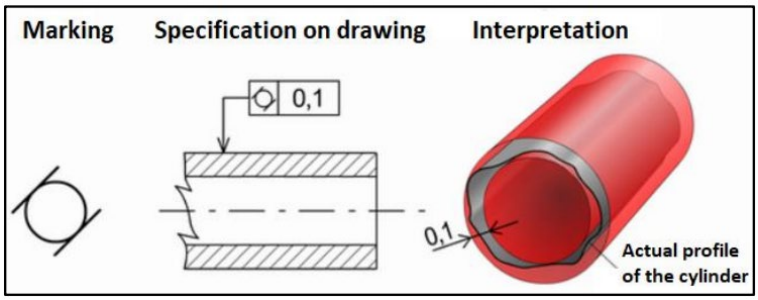


Figure 6
Interpretation of cylindricity, adapted from [22]

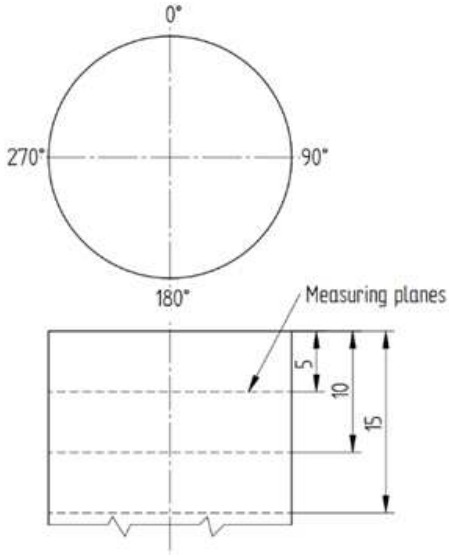


Figure 7
Principle of measurement

3 Results

3.1 Results of Surface Roughness Measurement

The roughness profiles measured after printing and turning for each layer thicknesses and orientations are shown in Fig. 8-15.

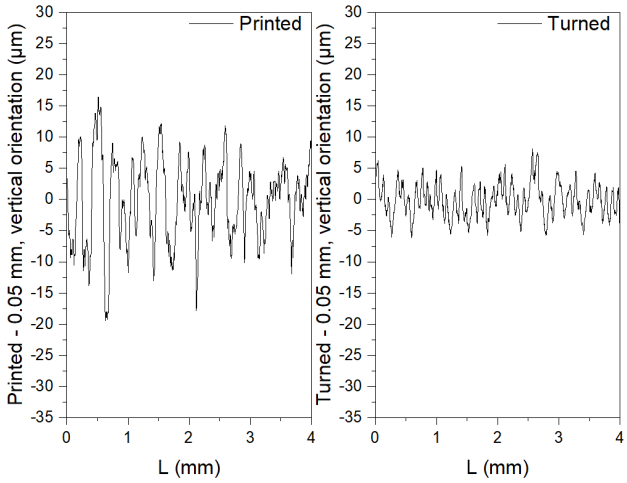


Figure 8
Roughness profiles after printing and turning

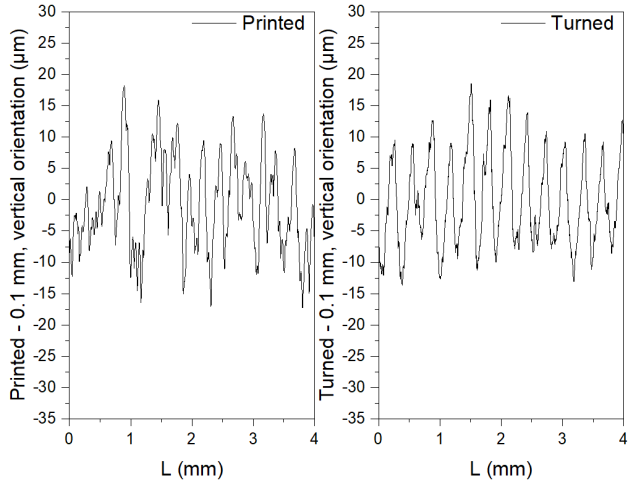


Figure 9
Roughness profiles after printing and turning

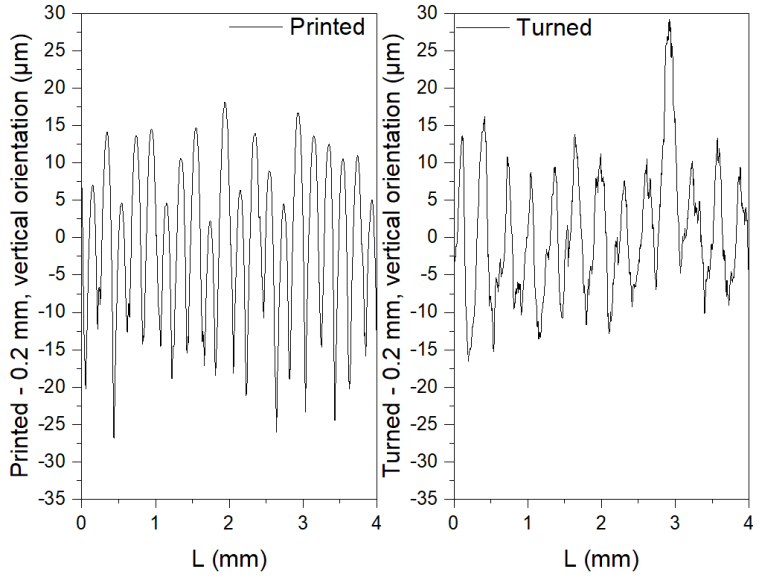


Figure 10
Roughness profiles after printing and turning

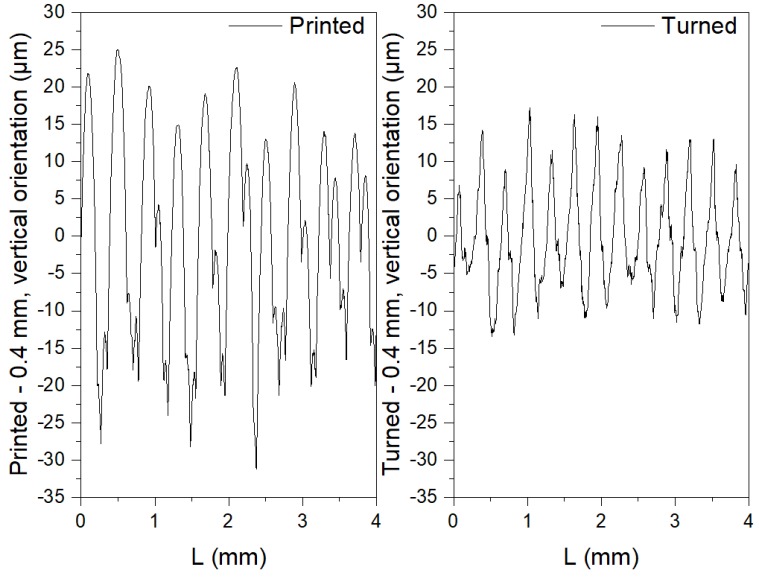


Figure 11
Roughness profiles after printing and turning

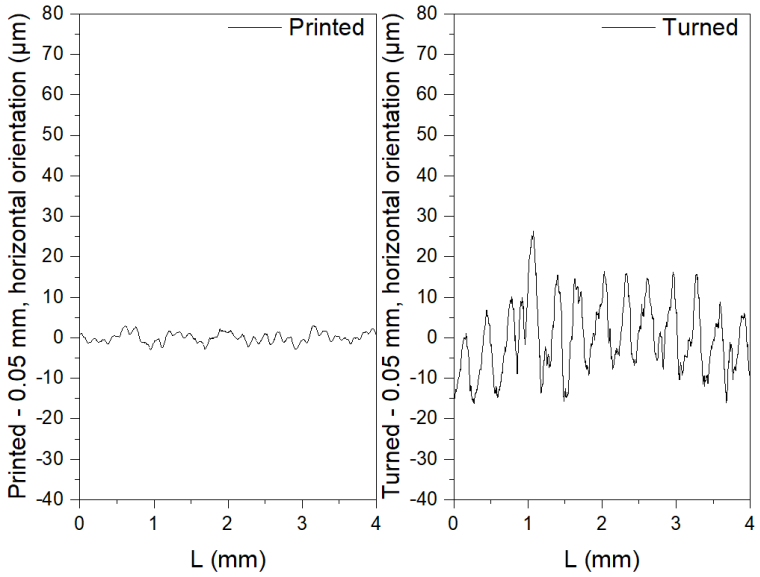


Figure 12
Roughness profiles after printing and turning

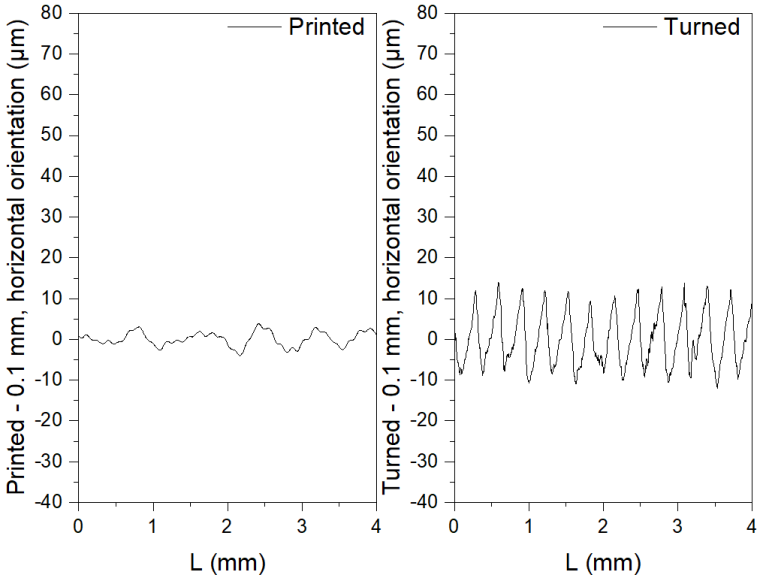


Figure 13
Roughness profiles after printing and turning

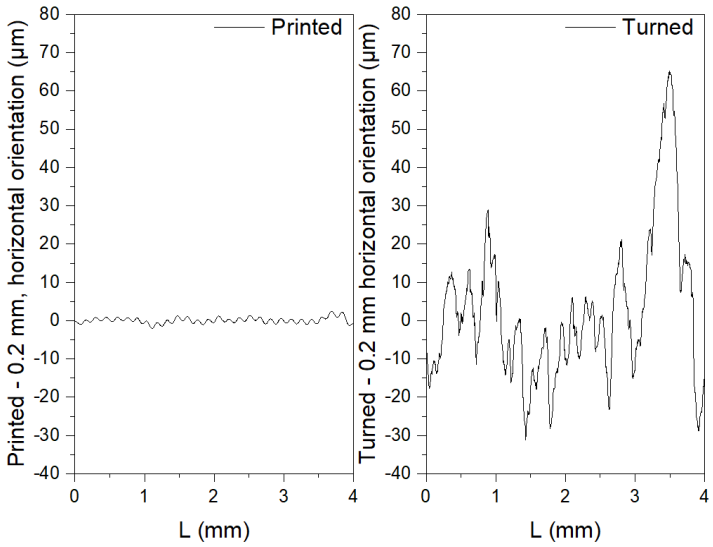


Figure 14
Roughness profiles after printing and turning

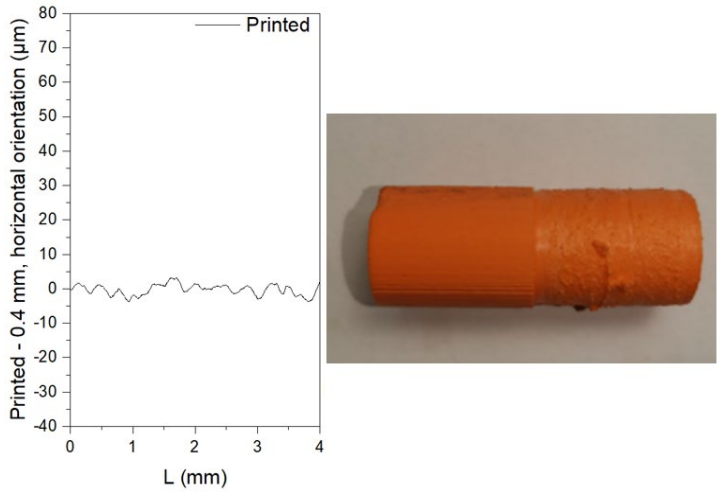


Figure 15
Roughness profiles after printing and turning

For each specimen, 3 roughness measurements were taken and then averaged, these are shown in Tables 3-6. As illustrated in Table 6 and Fig. 17, the surface roughness obtained after turning the 0.4 mm of layer thickness, horizontal oriented specimen has no data because it could not be measured as the surface was smeared during turning, as shown in Fig. 15.

Table 3

R_a values as a function of layer thickness for vertical orientation after printing

Layer thickness	R_{a1}	R_{a2}	R_{a3}	*avg.* R_a	Dispersion
0.05	7.981	7.780	9.017	8.259	0.664
0.1	11.064	9.051	7.612	9.242	1.734
0.2	12.803	11.630	10.786	11.740	1.013
0.4	24.552	23.981	22.202	23.578	1.226

Table 4

R_a values as a function of layer thickness for horizontal orientation after printing

Layer thickness	R_{a1}	R_{a2}	R_{a3}	*avg.* R_a	Dispersion
0.05	1.895	1.745	2.624	2.088	0.470
0.1	0.651	1.507	1.067	1.075	0.428
0.2	1.785	1.330	4.535	2.550	1.734
0.4	1.721	1.365	1.648	1.578	0.188

The average surface roughness (R_a) after printing and turning as a function of layer thickness in case of vertical orientation is shown in Fig. 16.

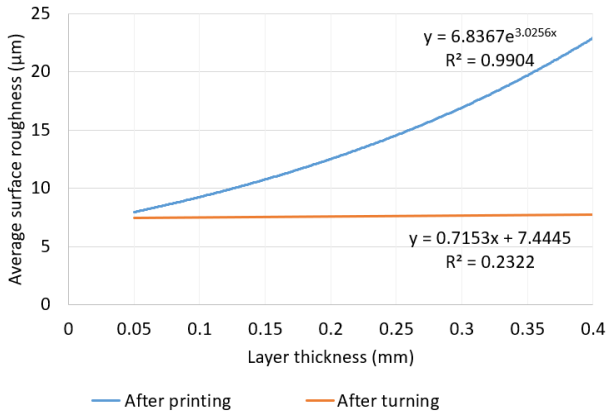


Figure 16

R_a as a function of the layer thickness after printing and turning in case of vertical orientation

Table 5

R_a values as a function of layer thickness for vertical orientation after turning

Layer thickness	R_{a1}	R_{a2}	R_{a3}	*avg.* R_a	Dispersion
0.05	7.605	7.138	6.968	7.237	0.330
0.1	8.282	7.587	7.286	7.718	0.511
0.2	7.119	7.686	8.324	7.710	0.603
0.4	7.571	8.223	7.154	7.649	0.539

Table 6

R_a values as a function of layer thickness for horizontal orientation after turning

Layer thickness	R_{a1}	R_{a2}	R_{a3}	*avg.* R_a	Dispersion
0.05	8.265	7.509	8.117	7.964	0.401
0.1	6.496	6.703	6.953	6.717	0.229
0.2	12.836	13.686	11.692	12.738	1.001
0.4	n/a.	n/a.	n/a.	n/a.	n/a.

The average surface roughness after printing and turning as a function of layer thickness in case of horizontal orientation is shown in Fig. 17.

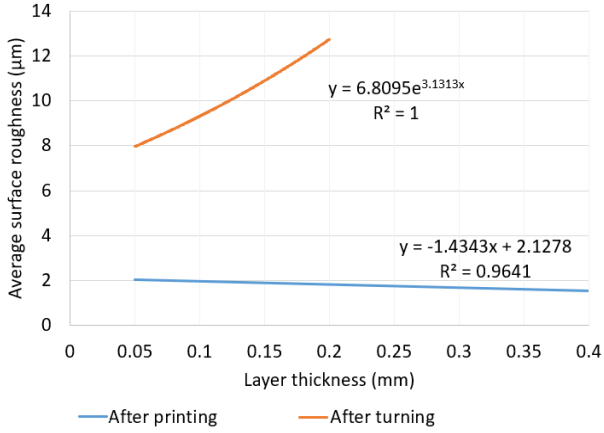


Figure 17

R_a as a function of the layer thickness after printing and turning in case of horizontal orientation

3.2 Results of Cylindricity Measurement

The deviation from nominal size and calculated tolerance as a function of layer thickness in case of vertical orientation after printing is shown in Table 7 and Fig. 18.

Table 7

Deviation from nominal size and calculated tolerance as a function of layer thickness for vertical orientation after printing

Layer thickness	Lower limit size (mm)	Upper limit size (mm)	Tolerance field width (mm)
0.05	-0.155	0.002	0.157
0.1	-0.091	0.06	0.151
0.2	-0.035	0.111	0.146
0.4	-0.3	-0.178	0.122

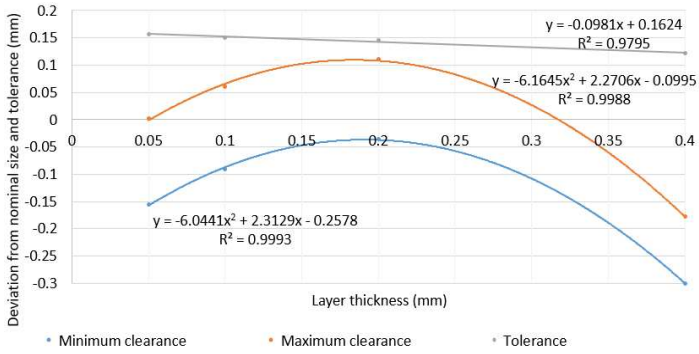


Figure 18

Deviation from nominal size and calculated tolerance as a function of layer thickness for vertical orientation after printing

The deviation from nominal size and calculated tolerance as a function of layer thickness in case of horizontal orientation after printing is shown in Table 8 and Fig. 19.

Table 8

Deviation from nominal size and calculated tolerance as a function of layer thickness for horizontal orientation after printing

Layer thickness	**Lower limit size (mm)**	**Upper limit size (mm)**	**Tolerance field width (mm)**
0.05	-0.097	0.296	0.393
0.1	0.012	0.352	0.34
0.2	0.095	0.36	0.265
0.4	0.343	0.688	0.345

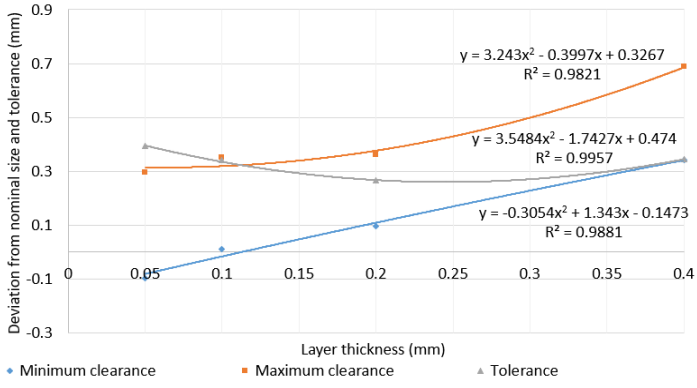


Figure 19

Deviation from nominal size and calculated tolerance as a function of layer thickness for horizontal orientation after printing

The deviation from nominal size and calculated tolerance as a function of layer thickness in case of vertical orientation after turning is shown in Table 9 and Fig. 20.

Table 9

Deviation from nominal size and calculated tolerance as a function of layer thickness for vertical orientation after turning

Layer thickness	Lower limit size (mm)	Upper limit size (mm)	Tolerance field width (mm)
0.05	-0.01	0.02	0.03
0.1	-0.01	0.03	0.04
0.2	-0.02	0.03	0.05
0.4	-0.01	0.04	0.05

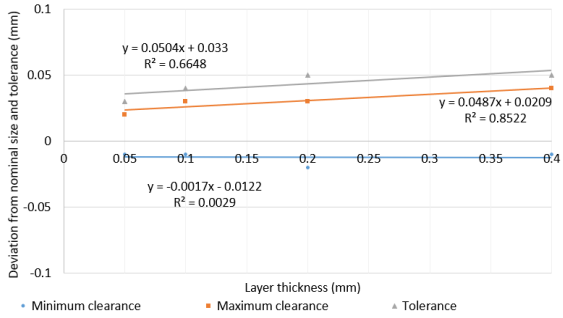


Figure 20

Deviation from nominal size and calculated tolerance as a function of layer thickness for vertical orientation after turning

The deviation from nominal size and calculated tolerance as a function of layer thickness in case of horizontal orientation after turning is shown in Table 10 and Fig. 21. As illustrated, the deviation from nominal size and calculated tolerance obtained after turning the 0.4 mm of layer thickness, horizontal oriented specimen has no data because it could not be measured as the surface was smeared during turning, as shown in Fig. 15.

Table 9

Deviation from nominal size and calculated tolerance as a function of layer thickness for horizontal orientation after turning

Layer thickness	Lower limit size (mm)	Upper limit size (mm)	Tolerance field width (mm)
0.05	-0.01	0.05	0.06
0.1	-0.03	0.04	0.07
0.2	-0.1	0.12	0.22
0.4	n/a.	n/a.	n/a.

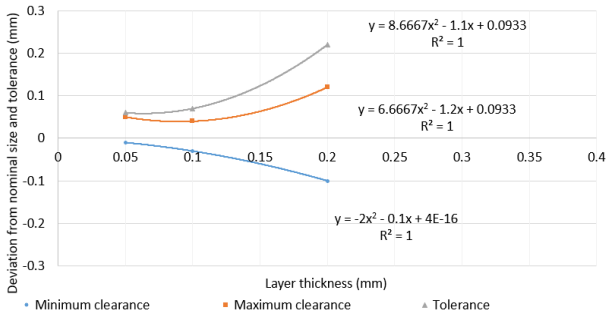


Figure 21

Deviation from nominal size and calculated tolerance as a function of layer thickness for horizonal orientation after turning

3.3 Chip Formation

The chip produced during the turning for both orientations are shown in Fig. 22.

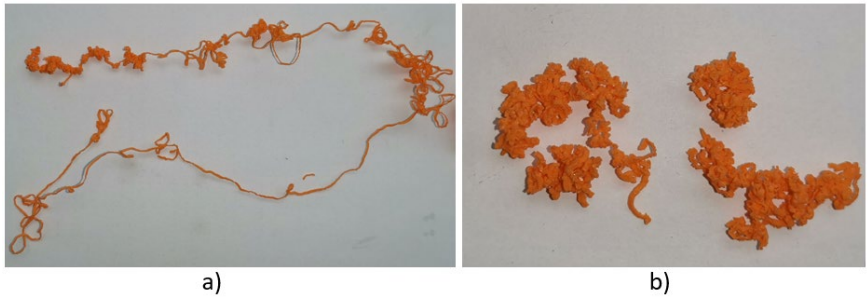


Figure 22

Flowing chips in case of a) vertical and fragmented chips b) horizontal orientation during turning

4 Analysis

4.1 Results of the Surface Roughness Measurement

The average surface roughness values after printing and turning and their plotting are shown in Table 3., Table 5. and Fig. 16 in case of vertical orientation. The surface roughness degradation increases with increasing layer thickness in case of printed surfaces. However, this deterioration is not proportional, as it is minimal when comparing 0.05 mm and 0.1 mm layer thicknesses, but the

machining time is halved. Table 5. Fig. 8-11. and Fig. 16 show that the surface roughness measured after printing has essentially no effect on the surface roughness after turning, with values ranging from 7.2 μm to 7.7 μm. Therefore, if the functional surface is to be produced by cutting technology, it is recommended to choose a layer thickness of 0.4 mm for printing, because it has no influence on the turned surface, but the printing time is nearly one eighth compared to the printing time required for printing with a layer thickness of 0.05 mm.

The average surface roughness values after printing and turning and their plotting are shown in Table 4, Table 6, Fig. 12-15 and Fig. 17 in case of horizontal orientation. In the case of printing, the surface roughness decreased minimally as the layer thickness increased, but this result is not significant as the measurement was taken parallel to the printing direction. However, the surface roughness increased exponentially as a function of layer thickness after turning in all cases. As shown in Fig. 15 and 17, after turning the specimen printed with a 0.4 mm layer thickness, the surface roughness could not be measured because the surface was smeared.

In the end, the results of the two orientations cannot be compared, because while in the vertical orientation the measurements were taken perpendicular to the printed layers, in the horizontal orientation the measurements were taken parallel to the layer orientation. From Fig. 16-17, it can be said that after turning, better results can be achieved in vertical orientation than in horizontal orientation, regardless of the layer thickness.

4.2 Results of Cylindricity Measurement

As illustrated in Table 7 and Fig. 18, the deviation from the nominal size increased in the negative direction with increasing layer thickness in case of vertical orientation workpieces after printing. It was also observed that as the layer thickness increased, the width of the tolerance decreased. As shown in Table 9 and Fig. 20, the layer thicknesses and the dimensional variation observed in them had no influence on the dimensional stability of the turned specimen. The target diameter of 18 mm was achieved with a tolerance of 0.03-0.05 mm due to the accuracy of the lathe.

In case of horizontal orientation, as shown in Table 8 and Fig. 19, the dimensional deviation from the nominal size increased with increasing layer thickness after printing. An optimum in dimensional stability and accuracy is observed at a layer thickness of 0.2 mm. As illustrated in Table 10 and Fig. 21, the dimensional tolerance and dimensional stability of the turned surface deteriorate with increasing layer thickness. The workpiece printed at 0.4 mm layer thickness was also not measurable because the surface was smeared during turning.

Finally, as show in Figs. 20-21, it can be determined that the accuracy and dimensional stability of the vertically printed specimen is much better than that of the horizontally printed specimens.

4.3 Chip Formation

Fig. 21 shows the resulting chips during turning. It can be observed that workpieces printed with a vertical orientation formed a flowing chip during machining, which is understandable since the direction of the printed filament is the same as the direction of the cutting speed. On the other hand, in the case of the specimens with a horizontal orientation, small, fragmented chips were formed, which is due to the direction of the printed filament being perpendicular to the direction of the cutting speed. Consequently, it is better to chip specimens in a horizontal orientation because the broken chips are easier to handle. However, chip breakage can also be improved in the vertical orientation, but further experiments are needed to investigate this.

Conclusions

As shown in Fig. 16, the roughness of the printed surface has no influence on the roughness of the turned surface, the result being almost constant as a function of the layer thickness in case of vertical orientation. Fig. 17, shows that the surface roughness of the turned surface increases as a function of layer thickness, and deteriorates to the extent that it was unmeasurable at a layer thickness of 0.4 mm. Comparing the surface roughness measured in terms of orientation, it was found that the surface roughness was better for all layer thicknesses in the vertical orientation.

Fig. 20 shows that the tolerance of the turned surfaces is nearly constant as a function of layer thickness in case of vertical orientation, so there is no effect on accuracy. On the other hand, in the horizontal orientation, the accuracy of the turned surface deteriorates significantly as a function of layer thickness, and was unmeasurable for a layer thickness of 0.4 mm. Comparing the dimensional accuracy measured in terms of orientation, it was found that it was better for all layer thicknesses in the vertical orientation.

If functional surfaces are to be finished by some cutting technology, it is advisable to choose the printing in vertical orientation with layer thickness of 0.4 mm. The surface roughness is almost constant as a function of layer thickness, and a dimensional accuracy of 0.05 is adequate for many engineering applications. Only the chip breakage is favorable for the horizontal orientation, but this can be improved by using a chip breaker or by modifying the process parameters. The latter, will require further investigation.

References

[1] P. Ficzere, "The Impact of the Positioning of Parts on the Variable Production Costs in the Case of Additive Manufacturing," *Periodica Polytechnica Transportation Engineering*, Vol. 50, No. 3, pp. 304-308, 2022

[2] P. Ficzere et al., "ECONOMICAL INVESTIGATION OF RAPID PROTOTYPING," *INTERNATIONAL JOURNAL FOR TRAFFIC AND TRANSPORT ENGINEERING*, Vol. 3, No. 3, pp. 344-350, 2013

[3] A. Takacs, "Safe In and Out of the Car," *In: Jármai, K., Cservenák, Á. (eds) Vehicle and Automotive Engineering 4. VAE 2022. Lecture Notes in Mechanical Engineering*, Miskolc, Hungary, pp. 63-70, 2023

[4] B. Ádám, Z. Weltsch, "Thermal and Mechanical Assessment of PLA-SEBS and PLA-SEBS-CNT Biopolymer Blends for 3D Printing," *Applied Sciences 2021*, Vol. 11, No. 13, p. 6218, 2021

[5] R. Velu et al., "3D printing technologies and composite materials for structural applications," *Green Composites for Automotive Applications*, pp. 171-196, 2019

[6] C. W. J. Lim et al., "An Overview of 3-D Printing in Manufacturing, Aerospace, and Automotive Industries," *IEEE Potentials*, Vol. 35, No. 4, pp. 18-22, 2016

[7] K. Tomasz et al., "Emission of particles and VOCs at 3D printing in automotive," *Smart Innovation, Systems and Technologies*, Vol. 155, pp. 485-494, 2019

[8] F. Garai et al, "Development of tubes filled with aluminium foams for lightweight vehicle manufacturing," *Materials Science and Engineering A*, Vol. 790, 2020

[9] S. M. M. Hanon et al., "Tribological Behaviour Comparison of ABS Polymer Manufactured Using Turning and 3D Printing," *International Journal of Engineering and Management Sciences*, Vol. 4, No. 1, pp. 46-57, 2019

[10] B. Ádám et al., "THE EFFECT OF DIFFERENT PRINTING PARAMETERS ON MECHANICAL AND THERMAL PROPERTIES OF PLA SPECIMENS,", *Gradus*, Vol. 7, No. 3, pp. 166-173, 2020

[11] K. Kun, "Reconstruction and development of a 3D printer using FDM technology," *Procedia Engineering*, Vol. 149, pp. 203-211, 2016

[12] K. Kandananond, "Surface Roughness Reduction in A Fused Filament Fabrication (FFF) Process using Central Composite Design Method," *Production Engineering Archives*, Vol. 28, No. 2, pp. 157-163, 2022

[13] A. W. Hashmi et al., "The Surface Quality Improvement Methods for FDM Printed Parts: A Review,", *Fused Deposition Modeling Based 3D Printing*, pp. 167-194, 2021

[14] B. Gadagi and R. Lekurwale, "A review on advances in 3D metal printing," *Materials Today: Proceedings*, Vol. 45, pp. 277-283, 2021

[15] S. R. S. Bharathi et al., "Multi objective optimization of CNC turning process parameters with Acrylonitrile Butadiene Styrene material," *Materials Today: Proceedings*, Vol. 27, pp. 2042-2047, 2020

[16] M. Trifunović et al., "Investigation of cutting and specific cutting energy in turning of POM-C using a PCD tool: Analysis and some optimization aspects," *Journal of Cleaner Production*, Vol. 303, p. 127043, 2021

[17] I. Fekete et al., "Highly toughened blends of poly(lactic acid) (PLA) and natural rubber (NR) for FDM-based 3D printing applications: The effect of composition and infill pattern," *Polymer Testing*, Vol. 99, p. 107205, 2021

[18] T. Yao et al., "A method to predict the ultimate tensile strength of 3D printing polylactic acid (PLA) materials with different printing orientations," *Composites Part B: Engineering*, Vol. 163, pp. 393-402, 2019

[19] A. Nugroho et al., "Effect of layer thickness on flexural properties of PLA (PolyLactid Acid) by 3D printing," *Journal of Physics: Conference Series*, Vol. 1130, No. 1, p. 012017, 2018

[20] A. Piros L. Trautmann, "Creating interior support structures with Lightweight Voronoi Scaffold," *International Journal on Interactive Design and Manufacturing*, Vol. 17, No. 1, pp. 93-101, 2023

[21] G. K. Maharjan et al., "Compressive Behaviour of 3D Printed Polymeric Gyroid Cellular Lattice Structure," *IOP Conference Series: Materials Science and Engineering*, Vol. 455, No. 1, p. 012047, 2018

[22] Sulinet knowledge base, "Tolerances and fits", [online] Available at: https://tudasbazis.sulinet.hu/hu/szakkepzes/mezogazdasag/muszaki-alapismeretek/alakturesek/alakturesek [Accessed: 14. 02. 2023.]

Quantifying Export Potential and Barriers of SMEs in V4

Jaroslav Belas

University of Information Technology and Management in Rzeszów, ul. Sucharskiego 2, 35- 225 Rzeszów, Poland. E-mail: jbelas@wsiz.edu.pl

Beata Gavurova

Faculty of Mining, Ecology, Process Control and Geotechnologies, Technical University of Kosice, 042 00 Kosice, Slovak Republic. E-mail: beata.gavurova@tuke.sk

Matus Kubak

Faculty of Economics, Technical University of Košice, Nemcovej 32, 040 01 Kosice, Slovak Republic. E-mail: matus.kubak@tuke.sk

Zuzana Rowland

Institute of Technology and Business in České Budějovice, Research department of economics and natural resources management, Okružní 517/10, 370 01 České Budějovice, Czech Republic, E-mail: rowland@mail.vstecb.cz

Abstract: The article aims at quantifying the export potential and barriers of SMEs in V4 using specific criteria. The research took place in SMEs in Visegrad countries from September 2019 to April 2020, including a questionnaire survey of 478 managers and business owners using logic regression. The results showed differences in export activities of SMEs regarding their legal form, province, size, type of management, legislative obstacles and tax policies. Language and cultural differences were not a barrier to SMEs' export activities. Micro enterprises' likelihood of export was 70% smaller than medium-sized firms. Small companies' prospect of trade abroad was 42% smaller than medium-scale organizations. The export potential grew with the size of an enterprise, indicating an increased export likelihood in limited liability and joint-stock companies with the highest chance in the transport and production sector. The type of management (an owner) was hugely impactful on the company's exporting activities. Our results give valuable

information about effective strategies for export marketing and national exporter development programmes.

Keywords: micro enterprises, small and medium-sized enterprises, exporting activities, export policies, export barriers, support for exporters

1 Introduction

Small and medium-sized enterprises (SMEs) are an engine of the European economy [1] because they play important roles in the creation of GDP, employment and the supply of goods and services. Varga also emphasises their ability to produce (supply) for export markets [2].

The driving forces behind globalization make SMEs indispensable for territorial development. Fierce market competition with large multinational and supra-national firms requires SMEs to generate sustainable competitive advantage. Large companies invest their funds and abilities in managers' know-how and high-quality export departments for managing export activities. The SMEs' lack of equal opportunities and resources calls for research in multinational trade and the multinationalization of SMEs.

We still need to unveil the SMEs' strategies when entering the multinational markets and mark their effectiveness while knowing that small enterprises' export processes are more complex than large companies [3] [4] [5]. Global thinking also creates ample opportunities for trading abroad, allowing governments to encourage multinationalization and export ventures [6]. Available scientific studies discriminate between export barriers and other companies' challenges - exporters. The former involves factors preventing firms from exporting, appealing for careful generalization of studied phenomena and consistency of the terms related to firms' multinationalization.

Research studies on SMEs' multinationalization are unclear, showing no content, theoretical or methodological consistency. We thereby need to integrate their outcomes, refer to previous studies on various business aspects and implement findings essential for corporate practice and policy. The topic deserves devising quality methodologies for territorial selection involving careful data collection, sampling equivalence and criteria of analysed dimensions. Future opportunities will also favour a network of research teams spanning through territories and cooperating with interested innovative agencies and associations. The development of new multinational databases would allow the creation of up-to-date qualitative/quantitative approaches, methods and sophisticated analytical instruments.

These circumstances prompted us to quantify the export potential and barriers in SMEs in V4 using specific criteria.

2 Literature Review and Theoretical Framework Development

Various studies have been tackling companies' multinationalization and export for the last five decades, discussing both issues and the impacts of globalization on the business sphere. Many inquiries are heterogeneous, dealing with partial problems of multinationalization or export in a country, region or company, comprising samples in multiple economic and political environments. Despite obvious shortcomings, the studies inspired us to explore how different managers and entrepreneurs perceive export activities. Many analyses also provide information on corporate and public policies and institutions examining the business environment.

Petrovito and Pozzolo inspected the relationship between credit constraints and exports of SMEs in 65 emerging and developing countries between 2003 and 2014, gathering intelligence on credit evaluation through firms' self-evaluation. The authors revealed a close link between severe, statistically and economically significant financial restrictions and the company's outlook for export, including the export's contribution to the overall sales (extensive and intensive margin). The impact on both export margins was enormous for small enterprises and firms operating in territories with a less developed financial system, inhibited economic freedom and poor-quality institutions [7]. Chaney warned about fixed input costs of companies entering foreign markets, disrupting their liquidity. In such a case, the export only rewards enterprises with smooth cash flow, including firms profiting from home sales and giant corporations [8].

Gabaix and Maggiori emphasized the export's susceptibility to variations in the rate of exchange, producing a theory of assessing exchange rates by capital flows in imperfect financial markets. The authors argue that foreign exchange is sensitive to the imbalance in money markets, insufficiently reducing economic shocks, which should be a cornerstone of traditional macroeconomic analysis [9]. Breckova aimed to better understand exporters' behaviour of Czech SMEs through a year's systematic monitoring of exporting models of SMEs. Government and regional authorities' failure to sufficiently inform small and medium-sized enterprises leaves untapped many export aids, which would otherwise come in handy to the firms concerned [10]. Civelek et al. compared international differences between SMEs' perceiving export barriers and firm-level characteristics in European Countries. The authors worked on the premise that all small and medium-sized enterprises have the same size, province or legal form, allowing them to inspect multiple export barriers given by the specific territorial layout [11].

Virglerova et al. quantified the impacts of multinationalization barriers on how SMEs see their future, collecting samples from countries of V4. Enterprises able to manage export risks will be more successful in the market irrespective of costs of multinationalization, tax policy, legislation and language or cultural differences

[12]. Some authors suggest that other geographical localities may witness a different situation given various sectoral policies [13] [14] [15]. Ayob and Freixanet analyzed SMEs' support for multinationalization and the effects of government programmes, revealing that national schemes for exporters' support are imperative in the discussed matter. The project involves multiple factors determining the firm's global marketing efficiency. The study advises effectively securing the exporter's aid and maintaining a positive attitude toward export and educational systems [16].

Wilkinson and Brouthers explored exporters' support in SMEs through export performance, outlining the role of the firm's resources, trade shows and programmes in this aiding process [17]. Tkacova et al., and Kocisova et al. argued that not all tools for the export scheme are the same effective, claiming that their efficiency depends on the corporate nature and business procedures [18] [19]. Leonidou inspected multiple export barriers, including systematic reviews of 39 obstacles from 32 empirical studies. The author splits internal (incorporating, informational, functional and marketing) from external (comprising, procedural, governmental, tasks and environmental) barriers, declaring that although export barriers emerge from different situations, they mainly hinge on the idiosyncratic managerial, organizational, and natural background of the firm [20] [21].

Tesfom and Lutz examined the export problems of SMEs in developing countries, compiling 40 studies published over 25 years. The authors classified export barriers into the company, product, industry, export market, and macro environment impediments, identifying similarities and differences in developed and developing countries [22]. Kahiyo warned about the lack of surveys exploring the relationship between the firm's multinationalization strategies and problems of trading abroad, revealing that successful multinationalization largely rests on export barriers. The author distinguishes rapid and gradual multinationalization. The former stems from a positive managerial orientation and a lack of confidence in the host market, while the latter arises from limited knowledge and the need for skills. The study also appeals to proper business education [23].

Leonidou et al. pointed out that although many SMEs show massive export potential, they lack the stimuli to use it. The authors suggested compelling reasons for SMEs to engage in export activities. The study involved 40 incentives for export: internal, external, reactive and proactive, depending on various factors, including time, space and industry. Apart from huge sales and profits, corporate growth and unique products, small companies want to expand their production capacity, be independent of the oversaturated domestic market and promptly respond to worldwide demand. The authors also include other motivating factors not dependent on export [21].

Altıntaş et al. claimed that specific barriers might significantly weaken export performance, analyzing 2,000 Turkish SMEs through a questionnaire. The results showed procedural impediments and fierce foreign market competition undermines the output. Eliminating these obstructions would lead to higher export efficiency [24].

Virglerova et al. found that companies able to manage export risks will be more successful in the market, advising not to underestimate export-unrelated stimuli, whose role might increase in significance in time [12]. Mataveli et al. explored the impact of four groups of causally conditioned barriers, including human capital, cultural, administrative and financial obstacles, on the product export barrier. Except for administrative obstructions, all impediments were hugely impactful [25].

Arteaga-Ortiz and Fernández-Ortiz argued that specific firms export more than others because their managers think of the barriers differently. The authors researched 2,590 SMEs, revealing that knowledge barriers and ignorance of export processes, potential exporting benefits and markets were highly impactful in avoiding foreign trade [26].

Narayanan assumed that we should not only understand the restraints SMEs are facing on their way to multinationalization but also find optimum approaches to success. The author reviewed new methods for eliminating the barriers based on Leonidou's Model of export barrier classification, concluding that modern techniques are highly effective for prosperous multinationalization [27].

Although companies are looking hard for a dynamic in-house solution to eliminate internal export barriers, overcoming external impediments requires support from the government and policymakers. Kahiya argued that most empirical studies explore the driving forces behind the export barriers. The author suggested 36 variables, including the firm's demography, nature of exporting companies, management, environmental, transport and multinational business factors [28]. The study formulates explicit theories (fund allocation, gradual multinationalization, network and institutional hypotheses) and implicit ones (rationalization), aligning all decisive criteria of export barriers. SMEs' exporting activities closely relate to sectoral policies integrating novelties and system development, methodologies and instruments [29] [30] [31]. Love and Roper pointed out interconnecting export and innovations, claiming that appropriate measures for supporting access to finance will stimulate the SMEs' cash flow and encourage investments in innovations and export development. The authors also suggested counselling and mentoring hereof [32].

Roy et al. perceived the imperativeness of SMEs in economic growth, arguing that SMEs face fierce international competition and must often move mountains to succeed in cut-throat global markets. The authors classify the obstacles into two groups: external barriers (governmental and economic, political-legal, procedural and currency and task and socio-cultural) and internal restraints (informational, managerial, financial and marketing). Both categories indicated a slight correlation. The analysis showed the hugest trade barrier for SMEs in the process of multinationalization of the procedural and currency impediment, followed by task and socio-cultural obstructions. The inquiry also blamed the managers' poor command of exchange rate variations [33].

Chandra et al. argue that SMEs' multinationalization processes in developing countries depend on different factors than in developed states, calling for a detailed inspection. The authors claim that future research will focus on understanding the needs of underserved markets [34]. Haddoud et al. confirm that SMEs in developing countries can be a source of active business firms and sustainable economic growth, analyzing exporting intentions of companies from developing countries [35].

Mendy and Rahman explore relations between people, institutions and SMEs' multinationalization in developing countries, mentioning barriers related to attractive employment. The authors suggest integrating cultural and other obstacles people must face into universal models for further studies of SMEs' multinationalization [36]. Kahiya and Dean argue that many export impediments arise from export stages, examining the process of export obstructions. The authors conclude that resource constraints, marketing, knowledge and experience and export procedure barriers are dependent on the export stage, assuming that differences are only perceptible in early to advanced stages of development [37].

Paul et al. devised other methods and theories, including qualitative analysis. Its comparative counterpart focused on multinationalization processes will always have limited access to public data, depending on the managers' willingness to impart the information. The authors prefer exploring SMEs' multinationalization outside companies, i.e. their regional origin. The SMEs' province involves multiple factors shaping their export activities [38].

Manolopoulos et al. analyzed the role of resources in SMEs' export, focusing on the institutional quality, deducing that formal, informal and regulatory institutions should control export behaviour. The authors surveyed 150 companies, concluding that formal and informal establishments hugely, yet differently, impact SMEs' export activities. The resource allocation for exporting ventures depends on the business reception of the domestic institutional context, which is imperative in export decision-making [39].

Paul, and Puig et al. apply the lack of capital, resources, global experience, negotiating skills, knowledge of foreign markets and governmental protection, insufficient information, poor choice of reliable partners and distributors and sluggish demand for SMEs' products to the main export impediments [40] [41].

Musteen et al. accentuate a human factor and familiarity with SMEs' multinationalization processes. The authors analyzed 169 SMEs from the Czech Republic, concluding that strong and diverse global networks rely on the extensive knowledge of SMEs' managing directors. Although the study did not explore possible links between the density of worldwide networks and business expertise, it confirms the firms' performance rewards a good command of foreign markets [42].

Terjesen et al. surveyed studies on business, contrasting their comparison potential with ambitions for policy-making and recommendations for professional

experience. The authors classified results into individuals, companies, industries and territories. They revealed immense business diversity throughout the regions and the firms' profound impact in interpreting the financial and export performance and economic growth. The survey calls for extensive research extending dominant theoretical perspectives (culture, resource allocation, economic advancement, human capital, transaction cost economy etc.) by management, global trade and business by integrating various theories [43].

Leonidou argues that perceiving export barriers is contingent on the company's size and suggests a long-term investigation of the firms' exporting behaviour, structure and conceptual problems and framing of new theoretical concepts [20]. Paul, Soriano and Dobon, and Paul et al. point to a genuine difference of opinion between experts, each of them bringing in new findings [40] [44] [38]. The cited studies have broken new ground in exploring further dimensions, relationships, aspects, factors and contextual relations, setting a conceptual framework for the continuous development of theories, methods and systems. The industry is long-term short of relevant studies on developing markets, multidimensional analyses of territorial exporting policies, global comparisons and works using analytical tools. Global inquiries require relevant international databases, quality research teams and institutional cooperation. The presented article fervently supports all efforts of the studies cited.

3 Data and Methodology

Data set was collected from September 2019 to April 2020 in V4 countries. A random sample of 8,250 SMEs in the Czech Republic, 10,100 SMEs in Slovakia, 7,680 SMEs in Poland and 8,750 SMEs in Hungary were selected. In order to use the random sampling method, information on SMEs was obtained from the CRIBIS database for firms operating in the Czech Republic and Slovakia and from the Central Statistical Office of Poland. Data collection in Hungary was provided by the project partner Óbuda University in Budapest. The random selection technique was ensured by the following set of steps: defining the research sample (firms with less than 250 employees); listing all firms and assigning a unique number to each firm using the "Randbetween" function in Microsoft Excel; sorting the dataset according to this unique number; sending an e-mail to the selected firms with a request to fill in the questionnaire. In the second phase of data collection, firms were contacted by telephone to arrange completion of the questionnaire. The questionnaire was filled by the manager or owner of the business.

The questionnaire consisted of over 60 questions focused on barriers to doing business, macroeconomic environment of the companies, risk assessment methods, export activities, marketing mix management, bankruptcy risks, corporate social responsibility, managerial attitudes, strategic management goals, innovative potential of companies etc.

The presented analysis aims to identify the determinants of export capacities of companies operating in the Visegrad business area. To do so, the following 5 questions were selected that concern the issue of export:

1) Does your company export products and services abroad?
2) Higher export costs are no obstacle to the export of our products.
3) Legislative differences are not an obstacle to the export of our products.
4) The differences in tax policy are not an obstacle to the export of our products.
5) Language and cultural differences are not an obstacle to the export of our products.

The respondents of the survey expressed their level of agreement with abovementioned statements on a five-point Likert scale with following scaling: 1: Strongly agree, 2: Agree, 3: Neither agree nor disagree, 4: Disagree, 5: Strongly disagree.

4 Analysis and Results

First, it needs to be said that out of 1,585 companies that participated in survey, only 30 % export their products abroad. Table 1 presents a brief overview of export activities of companies by the legal form and the company's size.

Table 1
Export activity

		Does your company export its products and services abroad?	
		Yes	No
Legal form	Sole trader	18.2%	81.8%
	Limited liability company	35.7%	64.3%
	Joint-stock company	46.3%	53.7%
	Another legal form	23.0%	77.0%
Company size	micro	19.7%	80.3%
	small	39.3%	60.7%
	medium	54.3%	45.7%
Business sectors	Manufacturing	61.3%	38.7%
	Retailing	29.1%	70.9%
	Construction	22.6%	77.4%
	Transportation	65.4%	34.6%

		Does your company export its products and services abroad?	
		Yes	No
Legal form	Sole trader	18.2%	81.8%
	Limited liability company	35.7%	64.3%
	Joint-stock company	46.3%	53.7%
	Another legal form	23.0%	77.0%
Company size	micro	19.7%	80.3%
	small	39.3%	60.7%
	medium	54.3%	45.7%
	Agriculture	29.9%	70.1%
	Tourism	31.8%	68.2%
	Services	14.3%	85.7%
	Another area	30.7%	69.3%

Source: Authors

In terms of legal form, it can be said that the smallest volume of export activities was recorded in the case of sole traders and other legal forms and that export activities are primary performed by limited liability companies and joint-stock companies. In terms of company's size, more than half of the medium-sized enterprises and more than 40 % of small enterprises export their goods and services abroad, compared to only 19.7 % of micro enterprises. In terms of business sector, more than 60% of manufacturing companies and transportation companies export their goods and services abroad.

Only those firms that answered positively to the first question were included in the next part of the research. The sample consisted of 478 SMEs. Specifically, there were 163 questionnaires responded in the Czech Republic, 109 in Hungary, 107 in the Slovak Republic, and 99 in Poland.

Next, the authors focus on national differences in perceived barriers, or obstacles to exporting goods and services, as seen in Figure 1. Then, the focus is on the statement "Higher export costs are no obstacle to the export of our products." It can be seen that the perception of higher export costs as an obstacle for export is the most pronounced in the Slovak Republic, followed by the Czech Republic and Hungary. In Poland, higher export costs are seen as an export barrier to a much lesser extent. As for the statement "Legislative differences are not an obstacle to the export of our products.", it can be concluded that legislative burden is seen as a barrier to export more often in Poland and Slovakia, and to a lesser extent in the Czech Republic and Hungary. Tax policy impact on the export activities of companies was addressed in the statement "The differences in the tax policy are not an obstacle to the export of our products." As follows from Figure 1, the least significant effect of tax policy on export is recorded in the Czech Republic and Hungary, while there is an evident stronger negative impact of the tax policy on export in Slovakia and Poland.

The smallest impact on export activities in the Visegrad countries can be found in the field of culture and language, which was addressed in "Language and cultural differences are not an obstacle to the export of our products." In the case of language and cultural differences, those are significantly less affecting export in all countries; however, its impact appears to be the smallest in the Czech Republic.

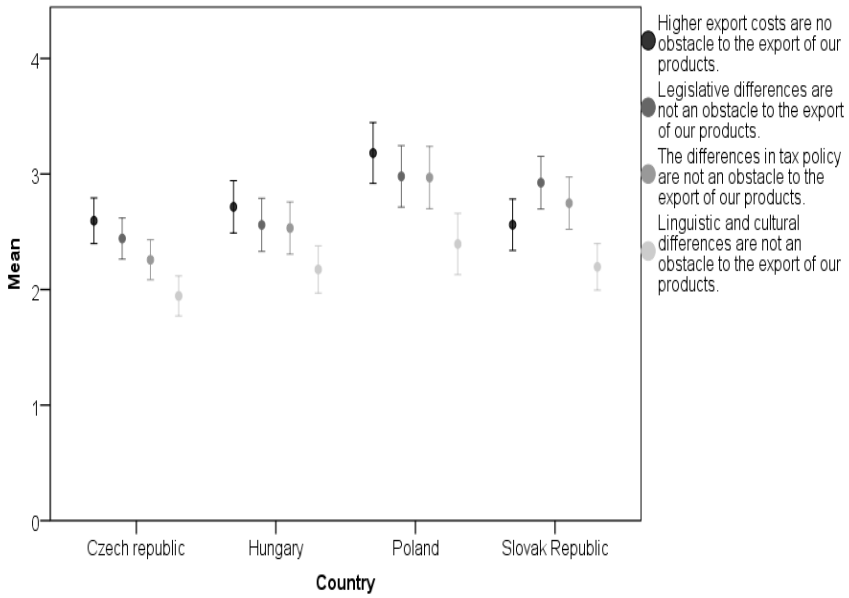


Figure 1
Perception of export barriers

Source: Authors

To be able to identify the characteristics of a company that underlie its potential to export, binary logistic regression is used, where the binary dependent variable takes the value of 0 if the company does not export its goods and services, and 1 if yes. At the beginning of the analysis, all considered characteristics that might possibly affect the company's export capacity, were integrated in the model, including the sector of activity, company's size, legal form, length of the business activities, gender, age, and level of education of manager/owner. Final regression model, where only statistically significant regressors are considered, has the following form:

$$ln\left(\frac{Pr(export)}{1-Pr(export)}\right)\left(\frac{Pr(export)}{1-Pr(export)}\right) = \beta_0 + \beta_1\, Company\ size_i + \beta_2 Legal\ form_i + \beta_3\, Sector_i + \beta_4\, Country_i + \beta_5\, Gender\ of\ the\ manager/owner_i$$

$$exportexport$$

not export its goods and services. According to the Likelihood ratio test, the model is well fitted and correctly predicts the observed phenomena. Regression analysis results are shown in Table 2.

Table 2
Binary logistic regression – propensity to export

Export [a]	B	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
				Lower Bound	Upper Bound
Intercept	-.918	.00			
Company size					
Micro	-1.178	.000	.308	.218	.435
Small	-.540	.006	.583	.397	.855
Medium	0[b]	.	.	.	.
Legal form					
Sole trader	.177	.538	1.193	.680	2.093
Limited liability company	.543	.042	1.721	1.021	2.902
Joint-stock company	.478	.165	1.613	.822	3.166
Another form of business	0[b]	.	.	.	.
Sector					
Manufacturing	1.090	.000	2.974	1.895	4.668
Retailing	-.026	.913	.975	.614	1.548
Construction	-.533	.052	.587	.343	1.004
Transportation	1.348	.000	3.849	1.906	7.773
Agriculture	.167	.595	1.182	.638	2.190
Tourism	.107	.785	1.112	.518	2.389
Services	-.773	.001	.462	.296	.720
Another area of doing business	0[b]	.	.	.	.
Country					
Czech Republic	.466	.006	1.593	1.145	2.217
Hungary	-.221	.262	.801	.545	1.180
Poland	.099	.612	1.104	.753	1.618
Slovak Republic	0[b]	.	.	.	.
Gender					
Male	.469	.001	1.599	1.212	2.109
Female	0[b]	.	.	.	.

a. The reference category is: No
b. This parameter is set to zero because it is redundant.

Source: Authors

First, it shall be stated that smaller enterprises are less likely to export than medium-sized enterprises. More specifically, micro-enterprises are almost 70% less likely to export than medium-sized enterprises, and small enterprises are 42% less likely to

export than medium-sized enterprises. It can thus be concluded that the probability of a company to be pro-export-oriented increases with the size.

As for the legal form of business as an explanatory variable, with another form of business used as the reference variable, the findings are as follows: the highest probability of export was identified in the case of limited liability companies, followed by joint-stock companies. The least probability of export is observed in the case of sole traders, although there is still a higher chance compared to the category "other forms of business".

As for the sectoral analysis in the context of exports, the reference value is another area of business. Compared to other areas of business, the transport sector is almost 4 times more likely to export its services, and the manufacturing sector is up to 3 times more likely to export its products. In the services and construction sector, exporting activities are about half less likely than in other areas of business. For other sectors, regression coefficients are not statistically significant.

Regarding the country of origin of a company, the estimated coefficient is statistically significant only in the Czech Republic, where a 59% higher probability of export is observed compared to Slovakia.

It is interesting that the companies run or owned by men show 60% higher probability of being pro-export oriented compared to companies owned or managed by women.

5 Discussion

The results indicated that only 30% of 1,585 investigated companies exported their products abroad. Analysing the firms' legal form, we found mainly limited liability and joint-stock companies engaged in export ventures. The company's size was also impactful on exporting, where more than 40% of small firms sold their goods and services abroad, leaving only 19.7% of foreign ventures for micro-enterprises. In terms of business sector, more than 60% of manufacturing companies and transportation companies export their goods and services abroad.

Along with our findings, Civilek et al. confirmed huge differences in enterprises' export activities regarding the legal form and size, revealing that diverse perception of barriers exists mainly among Czech and Slovak SMEs, which are older, smaller, and have limited liability [11].

We also found profound differences in the types of obstacles through countries. While high export costs were a grave impediment in Slovakia, followed by the Czech Republic and Hungary, Polish managers saw the biggest problem in legal restraints, pushing export expenses aside. The Czech Republic and Hungary felt disincentive effects of legislation to a lesser extent, followed by the tax policy,

whose impact was negligible in the countries concerned. Poland and Slovakia sensed the same negative influence of export tax policies and legislative impediments. None of the countries saw an obstacle in language and cultural differences, which relates to the outcomes of Virglerova et al., who proved no obstructions in different legislation, tax policy, culture or language of global activities of SMEs in V4 [12]. These findings clash with the verdict of Civilek et al., concluding that most SMEs encounter many obstacles in their exporting ventures. The authors argue that the main export barriers involve legal and tax burdens and language-cultural diversity, claiming that the country of SMEs' province is the main criterion of the different perceptions of export impediments [11]. Their analysis suggests that managers of SMEs of the same size, age, industry and legal form feel differently about exporting barriers in the country of their province. Belas et al., Dvorsky et al., Arteaga-Ortiz & Fernández-Ortiz and others arrived at a similar conclusion [45] [46] [26].

We also employed logistic regression to achieve the goals of the analytical part, obtaining different viewpoints of the SMEs' export activities regarding the size, legal form, industry, region and gender. Micro-companies had a 70% lower likelihood of exporting than medium-sized enterprises, while small organizations showed a 42% lower probability of foreign transactions than their medium-sized counterparts. We may conclude that the corporate growth scales with its export potential. Ključnikov et al., and Civelek et al. came to the same conclusion [47] [48] [11]. The legal form profoundly impacted the export undertakings of SMEs, indicating the best results in limited liability and joint-stock companies and the worst in sole traders. The industry saw the strongest export tendencies in the transport and manufacturing sector, while their halved numbers appeared in services and constructions, compared to the rest.

The country of the province also highly contributed to the pro-export orientation, indicating that Czech export undertakings surmounted Slovak efforts by 59%. Regarding gender, companies managed by males exported 60% more than companies directed by females. The results correspond with findings of other authors [49] [46].

Although the government support was subject to criticism, all SMEs in the surveyed countries can ask for aid for their export ventures. Breckova suggests that more than one-third of Czech exporters know about exporting support schemes used by managers participating in trade fairs and exhibitions. Contrary to the previous year, Czech entrepreneurs are slightly more aware of the government exporter's support schemes in the Czech Republic, including Single Contact Points or Anti-dumping Investigations, support for presentations abroad, information services, insurance, assistance services abroad etc. Business missions on behalf of the state administration proved ineffective [10]. The outcomes correlate with [19] [12].

Civilek et al. argue that managers and entrepreneurs of Slovak SMEs feel more relaxed about the surveyed barriers than their Czech counterparts [11].

They would appreciate better access to information and marketing services, export consultations, unlimited access to information on foreign business partners, personal support in the place of export etc. Ventures of state agencies, which should be principal business partners to SMEs, can also be helpful [18]. Export patterns applied by banks may also give valuable information for analyzing export activities and strategies. These theories comply with [48] [50] [51] [16] [52], and others.

SMEs' export and multinationalization arise from strategic choices anchored in a specific context. Studies compiled from surveys on SMEs from various sectors are only practical when helping understand the process and pace of SMEs' multinationalization. The articles illustrate differences in companies' behaviour and consequences. Our further research will explore destinations of SMEs' choice for multinationalization to examine local market challenges, institutional environment, socio-cultural diversity and local competitors [38] [47] [48]. The survey will expound upon export strategies of individual SMEs within a specific sector and identify consumer behaviour.

Despite plenty of studies on the multinationalization of SMEs and their potential competitiveness, there has not been a broad consensus on strategies guaranteeing success [15] [53]. Although most analyses confirm a close link between successful multinationalization and corporate performance, conflicting findings still prevent widespread agreement [18] [3]. Further research needs an effective model of successful corporate management, reflecting dimensions of multinationalization processes [54] [45] [15]. The design will require revising surveys of business and marketing strategies and exploring new determinants of performance and competitiveness [52] [51] [55].

Models developed decades ago ignore the current and future challenges of massive worldwide integration. Traditional methods of strategic management (SWOT, PESTEL, sensitivity analysis, scenario etc.) are obsolete and should give way to new corporate typologies observing effective strategies for SMEs to implement to be competitive in global markets [56] [57] [58]. Paul suggests using a SCOPE framework for a better orientation in the multinationalization process, helping low-technology SMEs withstand today's global pressures. This scheme allows SMEs to analyse their major problems and challenges, strengthening their long-term competitiveness [40]. We must delve into new issues and disputes SMEs face, evaluate applied theoretical models, strategies and tactics and compare the firms' growth with prevailing trends.

Conclusions

Despite many available studies and findings hereof, the issue of SMEs' export still stirs up avid interest, requiring up-to-date inquiries and surveys in today's global tendencies. Since secondary data are hard to obtain (institutional problems, managers' and owners' reluctance to provide them etc.), primary inputs and outputs remain the priority, effectively mapping SMEs' processes and transferring their results into corporate policies. The research rewards continuously exploring export

barriers in individual regions. The presented study aimed to quantify the SMEs' export potential and barriers in V4, unveiling differences in SMEs' export activities regarding the legal form and country of the province. The effect of legislative restraints and tax policies varied over the countries concerned, while language and culture did not constitute an impediment to SMEs' exporting ventures. Limited liability and joint-stock companies showed the strongest tendencies to export, with the highest likelihood in the transport and construction industry.

Collecting data on SMEs' export issues in individual sectors and analysing procedural, internal, and government barriers may direct firm and public policymakers to adopt effective export marketing strategies and national exporter development programmes. The findings can support and improve export stimulation projects since many countries develop schemes to encourage exporting products and services. The efficiency of these stimuli depends on adequate corporate segmentation. Perceiving firms' actual needs allow focusing on enterprises wanting to overcome export barriers using specific exporters' programmes. The creation of the projects should reflect these requirements and ensure enough export opportunities in the macroeconomic and microeconomic sectors. Including other research variables, e.g. network development monitoring, institutional effects, firms' adaptive abilities and innovative potential, will give rise to research teams creating new models and raising the bar to new scientific dimensions. These additional aspects will help better understand the processes of SMEs' multinationalization throughout the economic and political environment. Our outcomes also provide managers, policymakers and teaching professionals with valuable information and suggestions for further research.

This research has some limitations. Conclusions were drawn based on the attitudes of 478 entrepreneurs/managers from V4 countries. Although the research was conducted on a representative sample of respondents, the results cannot provide fundamental scientific insights, but they can enrich this research issue. This is also because the research was conducted under good economic conditions in all the countries studied. However, it can be assumed that the trends in this area are slightly different nowadays, when the economies of the world are affected by the war in Ukraine, high energy prices and high inflation rates. On the other hand, it can be assumed that SMEs will be able to cope with these changes relatively successfully.

Future scientific research will focus on exploring in more detail the directions outlined for the internationalisation of SMEs.

References

[1] Korcsmáros, E.; Renáta Machová, R.: Challenges of Burnout Prevention in Slovac SMEs– Focus on Optimal Employment. Acta Polytechnica Hungarica 2021, Vol. 18, No. 2, 87-104

[2] Varga, J.: Defining the Economic Role and Benefits of Micro, Small and Medium-sized Enterprises in the 21st Century with a Systematic Review of the Literature. Acta Polytechnica Hungarica, 2021, Vol. 18, No. 11, 209-228

[3] Zioło, M.; Kluza, K.; Kozuba, J.; Kelemen, M.; Niedzielski, P.; Zinczak, P.: Patterns of Interdependence between Financial Development, Fiscal Instruments, and Environmental Degradation in Developed and Converging EU Countries. Int. J. Environ. Res. Public Health 2020, 17, 4425

[4] Szabo, S.; Pilát, M.; Makó, S.; Korba, P.; Čičváková, M.; Kmec, Ľ.: Increasing the efficiency of aircraft ground handling—A case study. Aerospace, 2021, 9(1), 2

[5] Stehel, V.; Vochozka, M.: The Analysis of the Economical Value Added in Transport. Nase more, 2016, 63(3, SI), 185-188

[6] Kelemen, M.; Jevčák, J.: Security Management Education and Training of Critical Infrastructure Sectors' Experts, 2018 XIII International Scientific Conference - New Trends in Aviation Development (NTAD), 2018, pp. 68-71

[7] Pietrovito, F.; Pozzolo, A. F.: Credit constraints and exports of SMEs in emerging and developing countries. Small Business Economics, 2021, 56(1), 311-332

[8] Chaney, T.: Liquidity constrained exporters. Journal of Economic Dynamics and Control, 2016, 72, 141-154

[9] Gabaix, X.; Maggiori, M.: International liquidity and exchange rate dynamics. The Quarterly Journal of Economics, 2015, 130(3), 1369-1420

[10] Breckova, P.: Export Patterns of Small and Medium Sized Enterprises. European Research Studies Journal, 2018, 21(1), 43-51

[11] Civelek, M.; Polách, J.; Švihlíková, I.; Paták, M.: International Differences in the Perceptions of Export Obstacles By SMEs in the Same Firm-Level Characteristics: Evidence from European Countries. Folia Oeconomica Stetinensia, 2022, 22(1), 18-45

[12] Virglerova, Z.; Kliestik, T.; Rowland, Z.; Rozsa, Z.: Barriers to Internationalisation of SMEs in Visegrad Group Countries. Transformations in Business & Economics, 2020, 19(3)

[13] Petruf, M.; Korba, P.; Kolesár, J.: Roles of logistics in air transportation. Naše more: znanstveni časopis za more i pomorstvo, 2015, 62(3 Special Issue), 215-218

[14] Kužma, D.; Korba, P.; Hovanec, M.; & Dulina, Ľ.: The use of CAX systems as a tool for modeling construction element in the aviation industry. Naše more: znanstveni časopis za more i pomorstvo, 2016, 63(3 Special Issue), 134-139

[15] Gavurova, B.; Belas, J.; Bilan, Y.; Horak, J.: Study of legislative and administrative obstacles to SMEs business in the Czech Republic and Slovakia. Oeconomia Copernicana, 2020, 11(4), 689-719

[16] Ayob, A. H.; Freixanet, J.: Insights into public export promotion programs in an emerging economy: The case of Malaysian SMEs. Evaluation and program planning, 2014, 46, 38-46

[17] Wilkinson, T., & Brouthers, L. E. (2006) Trade promotion and SME export performance. *International Business Review*, 15(3), 233-252

[18] Tkacova, A.; Gavurova, B.; Danko, J.; Cepel, M.: The importance of evaluation of economic determinants in public procurement processes in Slovakia in 2010?2016. Oeconomia Copernicana, 2017, 8(3), 367-382

[19] Kocisova, K.; Gavurova, B.; Behun, M.: The evaluation of stability of Czech and Slovak banks. Oeconomia Copernicana, 2018, 9(2), 205-223

[20] Leonidou, L. C.: An analysis of the barriers hindering small business export development. Journal of small business management, 2004, 42(3), pp. 279-302

[21] Leonidou, L. C.; Katsikeas, C. S.; Palihawadana, D.; Spyropoulou, S.: An analytical review of the factors stimulating smaller firms to export: Implications for policy-makers. International Marketing Review, 2007, Vol. 24, No. 6, pp. 735-770

[22] Tesfom, G.; Lutz, C.: A classification of export marketing problems of small and medium sized manufacturing firms in developing countries. International journal of emerging markets, 2006, Vol. 1, No. 3, pp. 262-281

[23] Kahiya, E. T.: Export barriers and path to internationalization: A comparison of conventional enterprises and international new ventures. Journal of International Entrepreneurship, 2013, 11(1), 3-29

[24] Altıntaş, M. H.; Tokol, T.; Harcar, T.: The effects of export barriers on perceived export performance: An empirical research on SMEs in Turkey. EuroMed Journal of business, 2007, Vol. 2, No. 1, pp. 36-56

[25] Mataveli, M.; Ayala, J. C.; Gil, A. J.; Roldan, J. L.: An analysis of export barriers for firms in Brazil. European Research on Management and Business Economics, 2022, 28(3), 100200

[26] Arteaga-Ortiz, J.; Fernández-Ortiz, R.: Why don't we use the same export barrier measurement scale? An empirical analysis in small and medium-sized enterprises. Journal of Small Business Management, 2010, 48(3), 395-420

[27] Narayanan, V.: Export Barriers for Small and Medium-sized Enterprises: A Literature Review based on Leonidou's Model. Entrepreneurial Business and Economics Review, 2015, 3(2), 105-123

[28] Kahiya, E. T.: Five decades of research on export barriers: Review and future directions. International Business Review, 2018, 27(6), 1172-1188

[29] Antosko, M.; Korba, P.; Sabo, J.: One runway airport separations. Informatics, geoinformatics and remote sensing, SGEM. International Multidisciplinary Scientific GeoConference-SGEM, 2015, 241-248

[30] Džunda, M.; Dzurovčin, P.; Kaľavský, P.; Korba, P.; Cséfalvay, Z.; Hovanec, M.: The UWB radar application in the aviation security systems. Applied Sciences, 2021, 11(10), 4556

[31] Főző, L.; Andoga, R.; Schreiner, M.; Beneda, K.; Hovanec, M.; Korba, P.: Simulation aspects of adaptive control design for small turbojet engines. In 2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES), 2019, 000101-000106

[32] Love, J. H.; Roper, S.: SME innovation, exporting and growth: a review of existing evidence. International Small Business Journal, 2015, 33(1), 28-48

[33] Roy, A.; Sekhar, C.; Vyas, V.: Barriers to internationalization: A study of small and medium enterprises in India. Journal of International Entrepreneurship, 2016, 14(4), 513-538

[34] Chandra, A.; Paul, J.; Chavan, M.: Internationalization barriers of SMEs from developing countries: a review and research agenda. International Journal of Entrepreneurial Behavior & Research, 2020, 26(6), 1281-1310

[35] Haddoud, M. Y.; Jones, P.; Newbery, R.: Export intention in developing countries: configuration approach to managerial success factors. Journal of Small Business Management Plymouth, 2020, pp. 1-29

[36] Mendy, J.; Rahman, M.: Application of human resource management's universal model: an examination of people versus institutions as barriers of internationalization for SMEs in a small developing country. Thunderbird International Business Review, 2019, 61(2), 363-374

[37] Kahiya, E. T.; Dean, D. L.: Export stages and export barriers: revisiting traditional export development. Thunderbird International Business Review, 2016, 58(1), 75-89

[38] Paul, J.; Parthasarathy, S.; Gupta, P.: Exporting challenges of SMEs: a review and future research agenda. Journal of World Business, 2017, 52 (3), 327-342

[39] Manolopoulos, D.; Chatzopoulou, E.; Kottaridi, C.: Resources, home institutional context and SMEs' exporting: direct relationships and contingency effects. International Business Review, 2018, 27(5), 993-1006

[40] Paul, J.: SCOPE framework for SMEs: A new theoretical lens for success and internationalization. European Management Journal, 2020, 38(2), 219-230

[41] Puig, F.; Gonzalez-Loureiro, M.; Ghauri, P. N.: Running faster and jumping higher? Survival and growth in international manufacturing new ventures. International Small Business Journal, 2018, 36(7), 829-850

[42] Musteen, M.; Datta, D. K.; Butts, M. M.: Do international networks and foreign market knowledge facilitate SME internationalization? Evidence from the Czech Republic. Entrepreneurship: Theory and Practice, 2014, 38(4), 749-774

[43] Terjesen, S.; Hessels, J.; Li, D.: Comparative international entrepreneurship: a review and research agenda. Journal of Management, 2016, 42(1), 299-344

[44] Soriano, D. R.; Dobon, S. R.: Linking globalization of entrepreneurship in small organizations. Small Bus Econ 32, 2009, 233-239

[45] Belas, J.; Strnad, Z.; Gavurova, B.; Cepel, M.: Business Environment Quality Factors Research - Sme Management´s Platform. Polish Journal of Management Studies, 2019, Vol. 20, No. 1, pp. 64-77

[46] Dvorsky, J.; Kozubikova, L.; Ključnikov, A.; Ivanova, E.: Owners vs. Managers. Disparities of Attitudes on the Business Risk in SME Segment. Amfiteatru Economic, 2022, 24(59), 174-193

[47] Ključnikov A.; Belás J.; Kozubíková L.; Paseková P.: The Entreprenurial Perception of SME Business Environment Quality in the Czech Republic. Journal of Competitiveness, 2016, 8(1), 66-78

[48] Ključnikov, A.; Civelek, M.; Klimeš, C.; Farana, R.: Export risk perceptions of SMEs in selected Visegrad countries. Equilibrium. Quarterly Journal of Economics and Economic Policy, 2022, 17(1), 173-190

[49] Sobeková Majková, M.: The Relationship between the Risk of a Change of the Interest Rate and the Age of Entrepreneurs among Slovak SMEs. Journal of Competitiveness, 2016, 8(3), 125-138

[50] Kliuchnikava, Y.: The Impact of the Pandemic on Attitude to Innovations of SMEs in the Czech Republic. International Journal of Entrepreneurial Knowledge, 2022, 10(1), 34-45

[51] Belás, J.; Sopková, G.: A Model of Entrepreneurial Orientation. Transformation in Business & Economics, 2016, 15(2B(38B), 630-645

[52] Belás, J.; Bilan, Y.; Ključnikov, A.; Vincúrová, Z.; Macháček, J.: Actual problems of business risk in segment SME. Case study from Slovakia. International Journal of Entrepreneurial Knowledge, 2015, 3(1), 46-56

[53] Machova, V.; Vochozka, M.: Analysis of business companies based on artificial neural networks. In J. Horak (Ed.) Innovative Economic Symposium 2018 - Milestones and Trends of World Economy (IES2018), 2019, 61, VŠTE

[54] Sopko, J.; Kočišová, K.: Key Indicators and Determinants in the Context of the Financial Aspects of Health Systems in Selected Countries. Adiktologie, 2019, 19(4), 189-202

[55] Turisová, R.; Pačaiová, H.; Kotianová, Z.; Nagyová, A.; Hovanec, M.; Korba, P.: Evaluation of eMaintenance Application Based on the New Version of the EFQM Model. Sustainability 2021, 13, 3682

[56] Briestenský, R.; Ključnikov, A.: Identification of the Key Factors for Successful Hospital Management in Slovakia. Adiktologie, 2019, 19(4), 203-211

[57] Melnikova, L.; Cibereova, J.; Korba, P.: Building a training airport for pilots. Informatics, geoinformatics and remote sensing conference proceedings, SGEM 2016. 16th International Multidisciplinary Scientific Geoconference (SGEM 2016), pp. 109-116

[58] Bartos, V.; Vochozka, M.; Janikova, J.: Fair value in squeeze-out of large mining companies. Acta Montanistica Slovaca, 2021, 26(4), 712-731