Possible Implications for Land-Use Planning Mechanisms when Considering the Results of Monitoring and Modelling Air Pollution by Industry and Transport on the Example of Kazakhstan Cities

Olga Shvets[1], György Györök[2]

[1] D. Serikbayev East-Kazakhstan Technical University, Serikbayev st., 19, 070003, Ust-Kamenogorsk, Kazakhstan, oshvets@ektu.kz

[2] Alba Regia Technical Faculty of Engineering, Óbuda University, Budai út 45, H-8000 Székesfehérvár, Hungary, gyorok.gyorgy@amk.uni-obuda.hu

Abstract: For many Central Asian cities, the era of Soviet rule was – if not even reason for their existence – at least determining their historical and contemporary economic alignment. Even today in many cities of Central Asia such as Pavlodar, Dushanbe, Bishkek or Tashkent the secondary sector forms the economic backbone. That is also the case in Ust-Kamenogorsk, Kazakhstan. Due to declining public transport systems, growing motorization and comparatively old car fleets emissions from the transport sector became the second important source of air pollution inside the cities. Together with the emissions from the secondary sector, transport emissions can – under certain weather conditions – lead to health threatening living conditions in some parts of a city. Although filter technologies in the industry, an update of the car fleet and further investments in public transport might alleviate some of these effects they are for manifold reasons often not very easy to implement. Therefore, this paper tries to assess already existing and potential possibilities to consider potential emission hot-spots already during the process of planning urban extensions, specifically new living quarters. As a result, this investigation draws upon experiences that have been made in cities worldwide but specifically in Ust-Kamenogorsk, Kazakhstan.

Keywords: modelling; air pollution; environmental monitoring

1 Interrelations of Land Use and Transport

As land is one of central to the lives of people, we need to be working to secure land tenure for us, while implementing a plan for the sustainable development and environment protection [1]. Industrialization, increased intensity of land use and

technological changes has had adverse environmental consequences. Land use is a dynamic process [2, 3]. Land resources are constantly undergoing quantitative and qualitative changes associated with the impact of various natural and man-made processes. Effective land use management means organized land use adapted to economic, social, and, of course, environmental conditions [2, 3]. There are two concepts in the practice of land use: rational and optimal land use. Optimal land use and development assume the use of the maximum functional potential of the land, which is possible in specific environmental conditions, at minimum cost, without negative consequences. Practice shows that the implementation of this approach is unlikely, since any intrusion into the environment, even if it leaves no visible traces, leads to a change in the natural situation.

Different land-use participants can bring conflicts between the goals of urban development and life, ecological protection and production (Figure 1). McHarg was the first who proposed the ecological evaluation based on GIS in 1960s [4]. He developed the "pastry mode" of land suitability assessment [4]. A landscape evaluation model on behalf of conservation planning ideas was developed by Steinitz in 1990s [5] and by Aljoufie M. in 2020s [6].

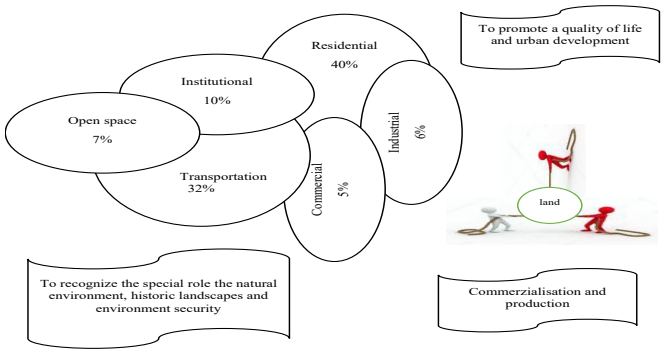


Figure 1
Urban land use types

The problem of information and software support of atmospheric air monitoring in large industrial cities is one of the most urgent for today. The main polluters are the enterprises of non-ferrous and ferrous metallurgy, motor transport, heat and power plants. It is necessary to attract modern automated tools and systems that allow to analyze the received data quickly, predict adverse environmental situations and visualize the obtained and predicted data on the map using GIS technologies to solve this problem. According to the development goals of land use management, urban land use types can be divided into residential land, industrial land, transportation, and so on (Figure 1).

The main goal of this study is to improve environmental safety by developing information and software support for estimation the current condition and forecasting the development of an environmental situation that meets the need to

automate the collection and processing of reliable, scientifically grounded and consistent information about the condition of the environment, depending on the organization of traffic.

It is necessary to solve the following tasks to achieve this goal:

- to estimate the current condition of information technology use for solving environmental monitoring problems;
- to formalize the natural-man-made system and its conditions from the position of describing complex systems;
- to develop a method for identifying the conditions of ecological sustainability of the transport system;
- to develop a model for the classification of conditions for the environmental sustainability of the transport system to identify the critical - the condition of an emergency situation;
- to build an information model of the subject area, containing a description of objects and subjects of monitoring;
- to provide visualization of evaluation results using GIS, as well as issuing recommendations for a particular condition,
- to prepare recommendations of land use taking into account environmental safety (public transport traffic organization, road markings, installation of additional road signs, etc.).

2 External (Societal) Effects of Transport

In recent years, the development of transport infrastructure in large cities has led to the fact that the process of movement began to occupy a significant part of the time [18, 5]. People usually choose the route, type of transport, thus, the features of pollutants. Movement can occur under different weather conditions, which affects the perception of the body, however, Thus, passengers can choose the route depending on the busy streets, depending on how they move inside the vehicle and, therefore, adjust microclimatic conditions. Thus, a set of durations affects all levels of exposure. The state of a person is based on the mutual influence of external, internal and social factors. Factors operating in the country can distinguish three main groups. External factors determine the complexity of the route. Social status meansthe place of a person in the society. Internal factors determine the body's response to harmful effects. Human exposure to harmful substances can be determined using direct and indirect methods. Direct methods require detailed personal data. Space methods are less labor intensive and are enclosed in measurements in microenvironment. All people in different microenvironment can communicate with people. In such cases, time and behavior

can be used for microenvironment. The main migrations in large urban agglomerations are “pendulum migrations”, associated with daily movements to and from work, from suburbs to cities and back, etc.

The social impact of transport and the distributional effects across various segments of society have traditionally been viewed as secondary or even tertiary concerns relative to the economic and environmental impact [7, 8, 9].

Social impact: presence of Infrastructure, presence of parked vehicles, presence of transport facilities, services and activities (accessibility) such as transport facilities; availability and physical access; level of service, etc. Social factors characterize the composition of the population and the characteristics of their life. Environmental factors determine the individual state, elapsed time and qualitative levels of the impact of the polluted environment on the population. Accounting for technical factors provided information allowing to quantify harmful levels of exposure. Table 1 presents indicators that take into account the above factors.

Table 1

Indicators for assessing the harmful effects on the population. Source: Black, W. [10]

Factors	Indicators
social	status (student, working, retired), movement group (gender, age) preferences (sports (physical education), occupation (working conditions)
ecological	condition (complaints of heart, discomfort of movement) bad habits (smoking) exposure time (distance, elapsed time) exposure level (street type, distance, smoking of fellow travellers)
technical	mode of transportation (type of transport) (by car, bus, tram, on foot) car emission characteristic (type, year of manufacture, mileage) admission (ventilation mode, change of ventilation mode, type of movement (on foot or car), frequency of movement, duration of movement)

For example, Hamburg and Göteborg. Hamburg (Germany). Population: 3.4 million inhabitants. HVV land area: 8.7 km^2. Passengers of public transport (2009): 656 million. The city of Hamburg approved an ambitious plan (1998-2021) to reduce emissions of pollutants: more than 300 measures relate to the growth of public transport and the increase of its modal share.

The city of Hamburg was awarded the title "European Green Capital 2011" (Figure 2).

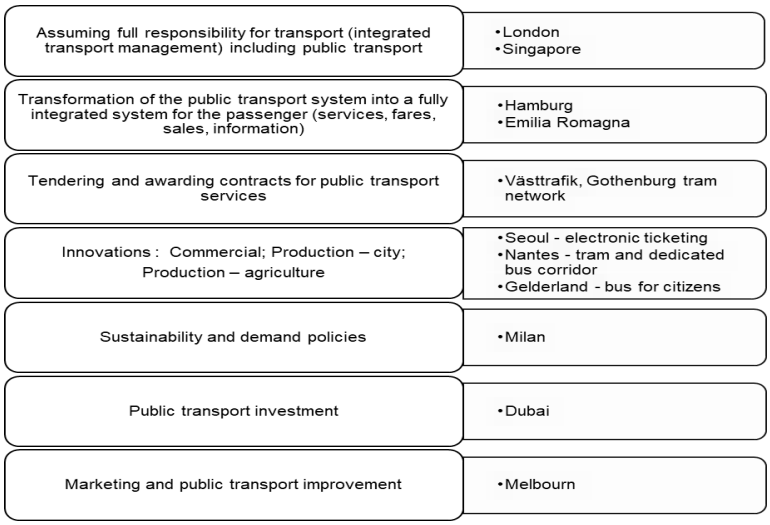


Figure 2
World successful practice in transport planning, Source: Statistical data

Managing Authority: HVV- Authority dei trasporti di Amburgo. Founded in 1965, the HVV is responsible for coordinating and fully integrating the dense urban public transport network (bus, metro, S-Bahn, regional railway and ferry lines. Hamburg includes ten municipalities (one of them is the city of Hamburg), HVV coordinates services 31 carriers (contracts are partially awarded through a public tender and partially - directly).

3 Monitoring and Modelling Air Pollution by Transport

Considering the issue of the formalization of the state of an industrial city atmosphere for effective monitoring, it is possible to take as a basis the “main ecological paradigm”, i.e. the concept of an ecosystem [11, 12], hence the recognition of the need for a systems approach [11, 12, 13]. The description of any complex system consists of three components: morphological, functional and informational [11, 12, 13].

3.1 Morphological Description of the State of the Atmosphere of an Industrial City

The morphological description gives a complete picture of the composition of the system, which is described by the four finite sets:

$$S_M = \{ \Sigma, Vr, \sigma, K \}, \quad (1)$$

where $\Sigma = \{ \Sigma_i \}$ – set of elements and their properties; $Vr = \{ Vr_i \}$ – set of relations; σ – structure; K – composition. The set of elements is a set of subsystems into which the morphological description no longer penetrates. The elemental composition contains the same type and different types of elements of different nature.

3.2 Informational and Functional Description of the Natural and Man-Made System

For each element of the private subsystem and the entire system as a whole, the functionality is defined by a set of morphological description X parameters (including external influences), a numerical functional n_z evaluating the quality of the system, and some mathematical operator of a deterministic or stochastic transformation ψ determining the relationship between the input X state and the state n_z output:

$$n_z = \Psi (X) \quad (2)$$

As can be seen from the above scheme of principles of complicating behavior, the response function n_z of the subsystem of the upper level depends on the functions describing the internal processes of the subordinate subsystems [1, 3].

It is customary to distinguish from the general theory of modeling physical systems five groups of parameters from the point of view of the method of their use in models:

– input parameters - $V = (v_1, v_2, \ldots, v_k)$, the values of which can be measured, but there is no possibility of impact on them (as applied to the model of the atmospheric air monitoring ecosystem, for example, meteorological conditions);

– control parameters - $U = (u_1, u_2, \ldots, u_r)$, with which it is possible to have a direct impact in accordance with certain requirements, which allows to control the system (these include a number of measures to establish the optimal operation modes of enterprises);

– disturbing (stochastic) effects - $\xi = (\xi_1, \xi_2, \ldots, \xi_l)$, the values of which randomly change over time and which are inaccessible for measurement, creating dispersion of unaccounted conditions or background noise;

– state parameters - $\boldsymbol{X} = (\boldsymbol{x_1,x_2,\ldots,x_n})$ - a set of internal parameters, the instantaneous values of which are determined by the current mode of functioning of the ecosystem and, ultimately, are the result of the total impact of input, control and disturbing factors, and the mutual influence of other intrasystem components;

– output (target or resultant) parameters - $\boldsymbol{n_z} = (\boldsymbol{z_1, z_2,\ldots, z_m})$, $z_k \in Z$, $k=1,n_z$ - some specially selected state parameters (or some functions from them) that are the subject of study (modelling, optimization) and are used as a criterion for the "well-being" of the entire ecosystem.

Disturbing factors can have both external and internal nature. Assuming that ecosystem parameters are related by some functional relations, which in the synthesized model are expressed by a set of equations of different mathematical nature (algebraic, logical, differential, finite difference, matrix, statistical, etc.), the expression can be written as:

$$\boldsymbol{n_z} = \boldsymbol{\Psi}(\boldsymbol{X, U, V}) + \boldsymbol{\xi} \tag{3}$$

Any ecosystem is a dynamic object, therefore, the equation of the static model (3) must be supplemented by a set of points in time $\boldsymbol{T}$, for which the instantaneous values of variables are measured. Since ecosystems also belong to objects with distributed parameters, whose components can change not only in time, but also in space $\boldsymbol{S}$, the general equation of the ecosystem model takes the form:

$$\boldsymbol{n_z} = \boldsymbol{\Psi}(\boldsymbol{X, U, V, T, S}) + \boldsymbol{\xi} \tag{4}$$

3.3 Input Parameters

Climate and meteorological conditions can be attributed to input parameters - $\boldsymbol{V} = (\boldsymbol{v_1,v_2,\ldots,v_k})$: v_1 - time of a year, v_2 - time of a day, v_3 - wind speed, v_4 - wind direction, v_5 - the degree of vertical stability of the atmosphere, v_6 - air temperature, v_7 - relative humidity.

3.4 Control Parameters

Consider the control parameters - $\boldsymbol{U} = (\boldsymbol{u_1,u_2,\ldots,u_r})$ of the ecosystem, which include the choice of operating mode of enterprises that contribute to the formation of the atmospheric air composition. Emissions of harmful substances into the atmosphere during the day can be considered a set of stationary random processes, each of which has its own distribution function f(t) [1, 3]. The stationarity of a random process is a consequence of the homogeneity of the process in time. It is true for a stationary process:

$$w_\xi(x_1;\ldots;x_n;t_1+\tau;\ldots t_n+\tau) = w_\xi(x_1;\ldots;x_n;t_1;\ldots t_n) \tag{5}$$

3.5 Disturbing Effects

As a disturbing (stochastic) effect – $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ – on the state of the atmospheric air, we consider the influence of motor vehicles. Automotive exhaust gases are a mixture of about 200 chemical elements and compounds. At present, it is considered that the main harmful components of exhaust gases are carbon monoxide, hydrocarbons and nitrogen oxides. The effects on the human body components of exhaust gases are divided into:

- ξ_1 - toxic - carbon oxide, nitrogen oxides, sulfur oxides, hydrocarbons, aldehydes, lead compounds;
- ξ_2 - carcinogenic - benz (a) pyrene;
- ξ_3 - irritant - sulfur oxides, hydrocarbons.

When analyzing the concentrations of impurities from stationary sources, the highest concentrations of the detected impurities are observed in the area of highways with heavy traffic, including - on Nezavisimost Avenue and, especially, at the node at the Palace of Sports. Of the 40 analyzed samples of the ξ_1 group, 68% exceeded the maximum one-time maximum permissible concentration of carbon monoxide, 20% - by suspended particles, 5% - by nitrogen dioxide. Also, the impact of vehicles on air pollution is a disturbing parameter during peak hours to create an unfavorable environmental situation. Determination of the intensity of the motor transport flow on the city's transport network. The study of calculation the amount of pollutants emissions on the nodes and links, on the road network of the city from motor vehicles is based on data on the intensity of the traffic flow. Figure 3 shows a cartogram characterizing the intensity of traffic flows (in cars per hour) on the main streets of Ust-Kamenogorsk city in the East Kazakhstan region.

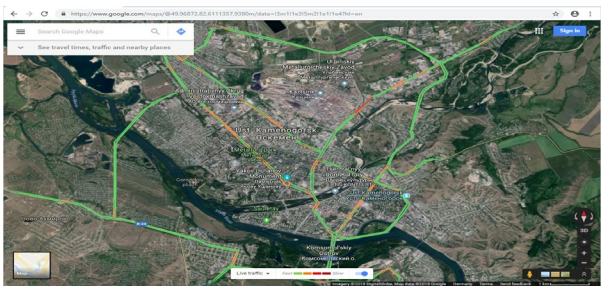


Figure 3
Cartogram of the average daily intensity of traffic flows in the city of Ust-Kamenogorsk, East Kazakhstan region. Source: Google map, Ust-Kamenogorsk

The intensity of the traffic flow (traffic) N_a is the number of vehicles passing through the cross-section of the road per unit of time. Year, month, day, hour and shorter periods of time (minutes, seconds) depending on the task of observation

and measuring instruments are taken as the estimated time period for determining the intensity of movement. It is possible to distinguish individual sections and zones on the road network of the city, in which traffic reaches its maximum size, as well as other sections where at the same time it is several times smaller. This spatial unevenness reflects the uneven placement of cargo and passenger-forming points and their places of attraction. The composition of the traffic flow is characterized by the ratio of different types of vehicles in it.

The composition of the traffic flow also affects the loading of streets and roads, which is primarily explained by the significant difference in the overall dimensions of the vehicles. For example, the length of passenger cars is 4-5 m, freight 6-8 m, the length of buses reaches 11 m, road trains - 24 m. Therefore, when driving in traffic, the concept of the dynamic size of a car is important, which depends mainly on the driver's reaction time and brake qualities of vehicles. The dynamic dimension L_∂ of a car in a dense traffic flow (Figure 4) means the road section that is minimally necessary for the safe movement of a car in a traffic flow at a given speed, the length of which includes the length of the car L_a and distance d, called the safety distance.

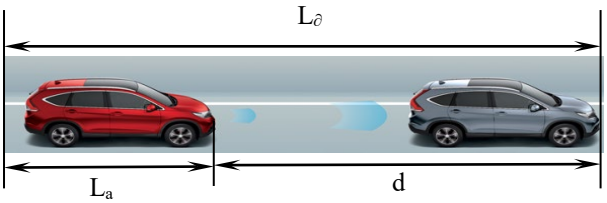


Figure 4

The dynamic dimension of the car in heavy traffic. Source: Pugachev, I. N. [14]

We can take the speed limit permitted in Traffic regulations of the Republic of Kazakhstan in populated areas (60 km/h) for estimation speed for a city highway. The baseline for determining the delay can be taken the normative speed of the message or the normative rate of movement for this type of road, if any. For example, if the estimated speed on the road is v_p = 60 km/h, which corresponds to a driving speed of 60 sec/km without delays, and the actual speed established by experienced testing v_f = 30 km/h (driving speed is 120 sec/km), then the loss of time each car in the stream is 60 sec/km. If the length of the considered section of the highway is, for example, 5 km, then the conditional delay of each car will be 5 minutes. Total loss of time for traffic flow:

$$T_\Delta = N_a t_\Delta T \tag{6}$$

where t_Δ is the average total delay of one car (seconds); T is the duration of observation (hour), N_a is the number of vehicles passing through the cross-section of the road per unit of time.

Traffic delays in real conditions can be divided into two main groups: on the nodes of roads and on links. Delays on nodes can be caused by maneuvering or slow-moving vehicles, pedestrian traffic, interference from standing vehicles, including during loading and unloading operations, as well as congestion associated with road glutting vehicles. Thus, we determined the characteristics of the state of traffic, the necessary data characterizing the traffic flow and the city's road network to simulate the spread of emissions of harmful substances from urban vehicles: traffic flow intensity, its composition by vehicle type, traffic density, traffic delays, length and the capacity of main streets and public roads. As the estimated speed for the city highway, we have adopted the allowed speed limit in populated areas: 60 km/h [14-16].

We propose to consider an assessment of the amount of energy and environmental impact of road traffic flows on the street and road network of Ust-Kamenogorsk, East Kazakhstan Region. It is assumed that the entire fleet of vehicles participates in the traffic flows, which is evenly distributed along the road network. Consider the entire city and the city's road network as a graph (crossroads (nodes) are vertexes, links are ribs). The graph edge has the following characteristics: the length, the number of lanes, the speed limit, the load and the presence of traffic lights. There are three categories of roads in one direction in Ust-Kamenogorsk: three-lane (Nezavisimosti Ave., Abay Ave, etc.), two-lane (Potanin St., Burov St., etc.), single-lane (Figure 5).

Figure 5

Map of Ust-Kamenogorsk city presented in the form of vertexes of the graph

It was analysed the state of the city's motor transport flows on the basis of data obtained from the Administrative Police Directorate of the Department of Internal Affairs of the East Kazakhstan region on August 1, 2018 to simulate the process of air pollution by motor transport (Table 2).

Thus, the total number of public transport is 2269 units, the total number of cars in the city of $M_{a.n.}$ = 100,400 units. The total length of the street transport network in the lanes L_d = 820.354 km.

We assume that the composition of automobile park is homogeneous, consisting of cars with gasoline engines, the average length of one car is L_{car} = 5 m (Table 2).

Table 2
Data on registered motor vehicles in Ust-Kamenogorsk (August 01, 2018) Source: Statistical data

Vehicle type	Number of registered vehicles, units		Total
	individuals	legal entities	
Passenger cars	81947	7654	89601
Trucks	3509	5021	8530
Buses	959	1310	2269
Motorcycles	1928	38	1966
Trailers	3809	1510	5319

We assume that the composition of automobile park is homogeneous, consisting of cars with gasoline engines, the average length of one car is L_{car} = 5 m.

Determine that the maximum density of traffic flows on one lane is equal to *Pmax* = 140 auto per km, where l = 5 is the length of one passenger car, d = 2 is the distance between cars. As a justification of this assumption, we note a significant part of the cars that make up the automobile park of Ust-Kamenogorsk city.

We propose to consider the hypothetical possibility that all vehicles of the city move around the city during the day, filling the transport network. Then, the average density of all vehicles of the city on the transport network with a uniform distribution is equal to p = 122,386 auto per km.

It is possible to determine the average speed of vehicles in busy and congested areas of the city having a state function. Since the intensity of traffic flows depends on the time of day, week, month and time of year, the time dependence should be considered. According to the calculations of the intensity of motor traffic flows over time during the day at the node “Palace of Sport” of Ust-Kamenogorsk, we determined the number of motor vehicles in the flows by field observation [16, 17].

The calculations carried out by field observation of vehicles at the main nodes of the city showed that the maximum number of motor vehicles exchanges fell on the “rush hour” in the period from 6:00 pm to 7:00 pm: 5,888 vehicles/hour. According to the “Methodological recommendation on the assessment of road capacity”, developed by the autonomous non-profit organization “Institute for Movement Safety Problems”, Moscow Automobile and Highway State Technical University (MADI), Irkutsk Technical University, Pacific University. "ROSDORNII", LLC "INEMDorTrans", it is recommended when calculating the throughput based on the value of the maximum practical throughput given in Table 3.

In addition, we analysed the data from a satellite interactive map using Google Maps online. On Google maps, the intensity of traffic flows is given in a four-color scheme from 06.00 to 22.00.

Table 3

Maximum practical throughput of traffic Source: Statistical data

Highways	Vehicles / hour
Two-lane	3600 in both directions
Three-lane	4000 in both directions
Four-lane without dividing strip	2100 on one lane
Four-lane with dividing strip	2200 on one lane
Six-lane without dividing strip	2200 on one lane
Six-lane with dividing strip	2300 on one lane
Highways with eight lanes	2300 on one lane

The user can determine by color the approximate number of vehicles. For example, green - traffic flow is free; orange - traffic flow with some interference; red - with slight delays; burgundy - delay. It was normalized four colors in the interval [0; 1] to process the results of data on the intensity of traffic flows on the map of Ust-Kamenogorsk. The example of results obtained on the dynamics of changes in the intensity of the traffic flow during the day on a map of Ust-Kamenogorsk for one day are shown in Figure 6 (a) color data b) approximate numerical data during the day).

Figure 6

Dynamics of changes in the intensity of the motor traffic during the day on the area Ushanov square, Ust-Kamenogorsk. Source: Based on statistical data and used Google map

It is possible to calculate the total amount of emissions per day and per year given the intensity of vehicles during the day.

Thus, knowing the hourly intensity of traffic and speed of traffic on each route and intersections, it was calculated the power of emission of harmful substances from road transport for the period of the most busy traffic. Thus, an assessment was made of the energy and environmental impact of traffic on the environment of Ust-Kamenogorsk.

The obtained emission functions of harmful substances from several sources and a disturbing parameter form the resulting function F_{rez}:

$$F_{rez}(t) = \sum_{1}^{n} f_i(t) \times \prod_{1}^{n} P_i(t) \tag{7}$$

It is necessary to take into account the probability P(t) of the release value at a certain point in time to find the predicted release value, then the release value of a particular substance from a particular source will be determined by the formula:

$$Q(t) = \int_0^t f(t) \times P(t)\, dt \tag{8}$$

The exhaustive probabilistic description of random processes is the assignment of the distribution laws for any set of samples: the distribution function

$$F_\xi(x_1;\ldots;x_n;t_1;\ldots t_n) = P[\xi(t_1) \le x_1;\ldots \xi(t_n) \le x_n] \tag{9}$$

or an appropriately determined probability density $w_\xi(x_1;\ldots;x_n;t_1;\ldots t_n)$.

Equations 7 and 8 are new and it is our contribution based on principles of mathematical modelling and equation 9 based on the distribution laws for any set of samples.

It is constructed a variation series and find for each hour of observations the number of intervals, step, minimum and maximum values, for example, at 1 a.m. The criterion χ^2 confirms the hypothesis about the normal distribution with the parameters a and σ^2, where $\alpha \in R$, $\sigma > 0$, $\xi \in N_{a,\ \sigma^2}$, ξ has the following distribution density:

$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad x \in R \tag{10}$$

The following numerical characteristics are used: mathematical expectation, variance, covariance function, correlation function for an approximate description [18, 19, 20]. In the above calculations, the expectation is calculated by the formula:

$$m_{1\xi}(t) = M[\xi(t)] = \int_{-\infty}^{\infty} x w_\xi(x;t) dx \tag{11}$$

The confidence interval for the mathematical expectation of the general population was calculated with a significance level of 0.05. It is necessary to know the correlation function of a random process [21] to verify the condition of ergodicity, which is the result of the transformation of the research process.

$$K_\xi(t_1;t_2) = M[\overset{\circ}{\xi}(t_1)\overset{\circ}{\xi}(t_2)] = B_\xi(t_1;t_2) - m_{1\xi}(t_1) m_{1\xi}(t_2) \tag{12}$$

3.6 Status Parameters

The parameters of the state - $\boldsymbol{X} = (\boldsymbol{x_1, x_2, \ldots, x_n})$ - are estimated by changing the concentrations of harmful emissions in the atmospheric air over time, and by such

characteristics as the depth and zone of actual contamination with a harmful substance. Let Conc = ($conc_{t1}$, …, $conc_{tk}$) - a set of concentrations of harmful substances, M_{usl} = (mu_{1t}, …, $mu_{вt}$) - a set of parameters characterizing the meteorological conditions at time t. In general, changes in concentration over time can be written as an autoregressive model as:

$$\hat{Y}_t = const + A_i \times Conc_t + B_j \times M_{usl} + \hat{E}_t \tag{13}$$

where Ai, i = 0, 1…p, k × k is the coefficient matrix depending on the impurity, Bj – k × l is the matrix of weather condition coefficients, Et is the disturbing parameter, const is the vector of constants. It is considered as weather conditions [16, 18, 19, 20, 22]: v - wind speed, vlag - relative humidity and ust - vertical atmospheric stability, measured by the coefficients of M.E. Berlyand model.

3.7 Output Parameters

Pollution parameters are measured during a certain time interval with a given discreteness and are fed to the input of the model, according to which the prediction is carried out. Checking the presence of process deviations from the model is performed by monitoring the deviations of the process indicators from the existing model values: $\Delta_{j=} y_{tj} - y_{tj}^{m}$, where y_{tj} are the coordinates of the vector Y_t at the point t, y_{tj}^{m} are the corresponding values calculated from the model. Any extraordinary event is preceded by any deviations from the normal course of a process. The nature of the development of the event and its consequences are determined by the destabilizing factors of different origin. This may be a natural, anthropogenic, social or other effect that disrupts the functioning of the system.

Ecological monitoring system consists from several blocks: emissions from enterprises, transport and climate conditions. It is very useful for detecting emergency situation in the city. It is necessary to model harmful emissions transfer from transport, in detail, for effective urban land use and trabsport planning.

3.8 Probabilistic-Statistical Modeling of the Harmful Impurities Transfer in the Atmosphere from Motor Vehicles

The atmosphere is a complex dynamic and turbulent system, in which various dynamic and physicochemical processes take place. Random chaotic velocity pulsations in all directions at all points of the flow are characteristic, giving almost all occurring processes a stochastic character for the turbulent flow of the atmosphere. Chaotic pulsation motions result in random, intensive mixing and specific turbulent diffusion, significantly exceeding the molecular, turbulent

viscosity of a gas, more uniform distribution than the laminar flow, averaging velocity distribution and its sharp drop in the near-wall area, a sharp increase in friction losses. The impurity particle can move along with the air flow or under the influence of external forces in the atmosphere, and through turbulent diffusion under the influence of turbulent pulsations of the atmosphere. Accordingly, the trajectory of the impurity particles can be considered a total random path: any of its coordinates at any time can be represented as the sum of deterministic and random components:

$$x(t) = \int_0^t u_x(t)dt + x'(t)\,, \tag{14}$$

where *x(t)* is the projection of deterministic velocity, m/s; *x'(t)* is a random process.

If we consider the motion of an impurity particle as a sequence of spasmodic displacements of length h over short periods of time Δt in one of six possible directions in the orthogonal coordinate system xyz, then the motion path will be three-dimensional broken line, and the direction of motion will be determined at each time point by the corresponding probabilities p_i.

The probabilities of each direction will be P_{+x}, P_{-x}, P_{+z} and P_{-z}. The sum is equal 1. (Figure 7).

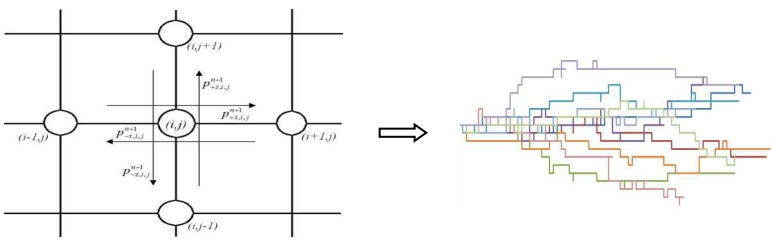


Figure 7

Particle transition schemes and a variant of calculating the possible trajectories of ten particles in a turbulent flow, (10 x 100 steps). Source: A result of modeling

Using the above-described probabilistic statistical model with known values of probabilities d, and a random number generator, respectively, the results of calculating options for possible trajectories of one and ten impurity particles in a turbulent flow are shown).

However, this approach makes it difficult to calculate when there are many sources and volumes of harmful substances. At any time, each particle can be in one of these grid nodes and each of these positions can be considered a possible state of a particle at a time with the corresponding probability *P(i,j,t)* (*i, j* – the number of grid nodes).

Presumably at each time point $t + \Delta t$, in the position (i,j), the probable direction of transition of the particle $P_{l,i,j}^{\ n+1}$ $(l=+x,-x,+z,-z)$ is determined, and accordingly, what flows in is added, and what follows is taken away (Figure 7).

According to Figure 8, the distribution of nitrogen dioxide from a complex node Ushanov Square, is considered with low wind as of 17-00 local time. The areola of impurity distribution are highlighted in different colors, taking into account the concentration.

The composition of attribute and spatial information is the implementation of a grid function on a plane with a constant step Dx horizontally and a constant step Dy vertically. The use of the grid function $q_{ij}=f(x_i,y_j)$ in interpolation procedures is generally accepted and is associated with a number of methodological advantages. It combines the convenience of storing the information received, the possibility of its statistical processing and mathematical transformation of the future map, as well as the ease of implementation of the procedures for constructing isolines (in the form of a broken line) and interval regions according to the values of interpolating functions in the grid nodes (Figure 8).

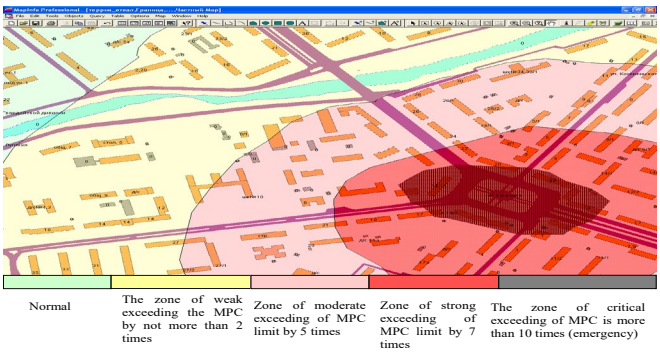


Figure 8

Distribution of nitrogen dioxide from Ushanov Square. Source: A result of modeling

When visualizing the simulation results for the construction of a grid in geographic information systems, Delaunay triangulation is used. After the application of isolines of impurity concentrations, it can be noted that there are areas on the map in which the excess concentration of impurities according to the criteria for identifying emergencies described in stage 2 is critical (emergency), despite the general "well-being" of the nature-man-made system state. Based on the results obtained, zones with a critical state of atmospheric air are selected at the boundaries of the critical impurity concentration. The topographic basis for building a map of Ust-Kamenogorsk city is the map obtained on the website of the Brief Research Group http://www.brif.kz/service/geoinformation_system.php. On a topographic map, the zone of possible contamination has the form of a circle, semicircle or sector with a wind speed forecast to be less than 0.5 m/s. 0.6-1 m/s and more than 1 m/s, respectively [16, 18, 19, 20].

After the calculations, a graph of the concentration distribution from the source is plotted, then the data obtained is visualized by plotting the concentration field in accordance with Figure 9.

In Figure 9, a carbon dioxide concentration field is plotted with an indication of the concentration excess values as compared to the MPC in fractions. The distribution of impurities is displayed on the map as concentration circles with centers in emission sources, which facilitates visual perception.

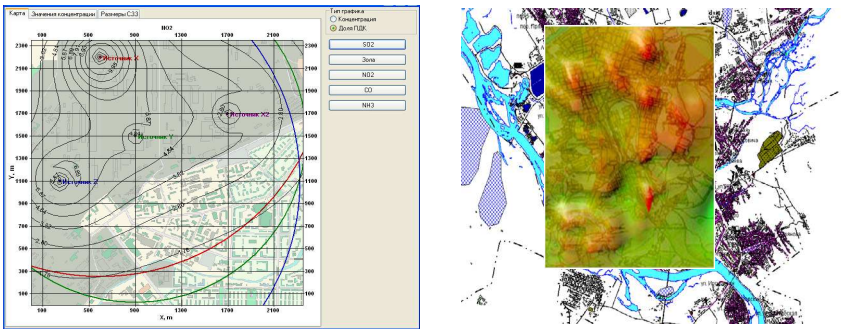


Figure 9

The “Concentration Field” window Three-dimensional map of the impurities distribution from multiple sources. Source: The result of visualization

You can also see the boundary isoline of the influence of each source on the formation of the field of impurity concentrations in the city as a whole [20]. When constructing a concentration field, it is possible to display in fractions of the MPC, which in some cases is more visual. The information presented in Figure 9 in the form of concentration isolines can be displayed by means of GIS technologies in the form of filled polygons, the color of which corresponds to the level of impurity concentration. In the places where impurity polygons overlap from several sources, the concentration level rises. The concentration reaches its highest value in the area of the source of pollution - for example, this is a complex node at peak hours. The negative impact on the public health is discussed in detail in the research [20]. With the help of GIS technology, it can be determined which buildings and institutions fall into an unfavorable zone at certain hours or are permanently in it. This information can be especially relevant for children's and health-improving institutions, such as kindergartens, schools, hospitals. GIS technologies make it possible to build three-dimensional maps of the distribution of impurities according to Figure 9.

After modelling and finding the places with harmful impact on people it is possible to prepare recommendations for effective land use including transport planning.

Common recommendations [22-26] are included in the strategy of sustainable development Ust-Kamenogorsk city of the Republic of Kazakhstan:

- Remove kindergartens, schools, hospitals and other official buildings in so-called good ecological zones.
- Reorganize and rebuild ecological-emergency areas.
- Decrease the negative impact of industry.
- Organize sustainable city transport system.

Kazakhstan residents, like most residents of modern cities, need more mobility and convenience from public transport in order to carry out daily door-to-door movement in the city with maximum comfort. Integration and an integrated approach should become the cornerstone of the transport life of the city, first of all, when a passenger uses different types of transport (inter-modality). As the research shows, the development of bicycle infrastructure can significantly change the transport habits of Kazakhstan residents in favor of increasing the share of sustainable transport and reducing the number of personal cars on the roads.

Conclusions

We can indicate following results of the work.

It was estimated the current condition of information technology use for solving environmental monitoring problems.

It was formalized the natural-man-made system and its conditions from the position of describing complex systems. We presented our own formalization on the bases of morphological, informational and functional description.

We proposed our own model for the classification of conditions for the environmental sustainability of the transport system to identify the critical - the condition of an emergency situation.

It was provided visualization of evaluation results using GIS (software was developed), as well as issuing recommendations of land use taking into account environmental safety for a particular condition.

Acknowledgement

This work was supported by “Sustainable Mobility Research in Central Asia (SUMRICA) project” and Technische Universitat Berlin.

References

[1] Scholz M., Hedmark A., Hartley W.: Recent advances in sustainable multifunctional land and urban management in Europe: a review. J. Environ. Plann. Manag., 55, 2012, pp. 833-854

[2] Li X. B.: Core of global environmental change research: frontier in land use and coverage change, Acta Geog. Sin. 51, 1996, pp. 553-558

[3] Vitousek P. M., Mooney H. A., Lubchenco J., Melillo J. M.: Human Domination of Earth's Ecosystems. In: Marzluff J. M. et al. (eds) Urban Ecology, 2008, Springer, Boston, MA, pp. 3-13, https://doi.org/10.1007/978-0-387-73412-5_1

[4] McHarg I.: Design with Nature, Natural History Press, New York, 1969

[5] Steinitz C.: A framework for theory and practice in landscape planning, GIS Eur., 1993, pp. 42-45

[6] Aljoufie M.: Toward an effective analysis of public transportation demand factors in car-dependent cities: case of Makkah City, Saudi Arabia, Urban, Planning and Transport Research, An Open Access Journal, Volume 9, 2021, Issue 1, pp. 258-276

[7] Geurs K.: Social Impacts of Transport: Literature Review and the State of the Practice of Transport Appraisal in the Netherlands and the United Kingdom Transport, Reviews 29(1), January 2009, DOI: 10.1080/01441640802130490

[8] Schiefelbusch M.: Rational planning for emotional mobility? The case of public transport development, Planning Theory, Technische Universität Berlin, 2010

[9] Theis T.: What is Sustainability? In T. Theis, & J. Tompkins, Sustainability: A Comprehensive Foundation. Houston: Connexions. Retrieved October 1, 2012, from http://cnx.org/content/m41188/1.7/

[10] Black W.: Sustainable Transportation: Problems and Solutions. New York: Guilford Press, 2010

[11] Bernstein J., Szuster B. W.: The new environmental paradigm scale: Reassessing the operationalization of contemporary environmentalism, The Journal of Environmental Education, Volume 50, 2019, Issue 2, pp. 73-83, https://doi.org/10.1080/00958964.2018.1512946

[12] Fedorov V. D.: Notes on the paradigm in general and the environmental paradigm in particular, Vestnik MSU, biological series, № 3, Moscow, 1977, pp. 8-22

[13] Khailov K. M.: "Life" and "Life on Earth": two scientifical paradigms, General biology journal, T. 59, № 2, Russia, 1998, pp. 137-151

[14] Pugachev I., Kulikov Y., Cheglov V.: Features of traffic organization and traffic safety in cities, Transportation Research Procedia, Volume 50, 2020, pp. 766-772, ISSN 2352-1465, https://doi.org/10.1016/j.trpro.2020.10.089

[15] Rajczyk P., Evtiukov S. New technological solutions in improving road safety, Transportation Research Procedia, 36, 2018, pp. 649-653, DOI: 10.1016/j.trpro.2018.12.133

[16] Repin S., Evtiukov S., Maksimov S. A method for quantitative assessment of vehicle reliability impact on road safety, Transportation Research Procedia, 36, 2018, pp. 661-668, DOI: 10.1016/j.trpro.2018.12.128

[17] Lukanin V. N. Buslaev A. P., Yashina M. V.: Road traffic and the environment, International journal of vehicle design, № 3, Interscience online Journal, 1998

[18] Berlyand, M. E.: Prediction and Regulation of Air Pollution, Springer, 1991, Volume 14, 312 p.

[19] Genrikhovich E.: Air Pollution Modeling and Its Application XII, Springer, 1998, 723 p., DOI: 10.1007/978-1-4757-9128-0_84

[20] Shvets O., Bugubayeva A., Rakhmetullina S., Wojcik W.: Development of intelligent information system of assessing the negative impact of industrial emissions on the public health, Journal of Theoretical and Applied Information Technology, Vol. 99 (15), August 2021, pp. 3737-3752, http://www.jatit.org/volumes/ninetynine15.php

[21] Matveev Yu. L.: Function and density distribution of pollutants and air temperature, Optics of the atmosphere and the ocean, T. 7, № 2, Russia, 1994, pp. 244-249

[22] Helmut E.: Landsberg. City climate, Hydrometeoizdat, Leningrad, 1983, 248 p.

[23] Low N.: Is Urban Transport Sustainable? In Making Urban Transport Sustainable, New York: Palgrave Macmillian, 2003, pp. 1-22

[24] Lee S., Lee M., Chun Y., Griffith D: Uncertainty in the effects of the modifiable areal unit problem under different levels of spatial autocorrelation: A simulation study, International Journal of Geographical Information Science, 33(6), pp. 1135-1154, https://doi.org/10.1080/13658816.2018.1542699

[25] Chen Y., Bouferguene A., Li H., Liu H., Shen Y., & Al-Hussein M.: Spatial gaps in urban public transport supply and demand from the perspective of sustainability, Journal of Cleaner Production, 195, pp. 1237-1248, https://doi.org/10.1016/j.jclepro.2018.06.021

[26] Chamseddine Z., Asmaa Ait Boubkr: Exploring the place of social impacts in urban transport planning: the case of Casablanca City, Urban, Planning and Transport Research, Volume 8, 2020, Issue 1, pp. 138-157, https://doi.org/10.1080/21650020.2020.1752793

Evaluation of Railway Vehicle Reliability Parameters

Ľubomír Ambriško, Janka Šaderová and Richard Antal

Institute of Logistics and Transport, Technical University of Košice,
Faculty of Mining, Ecology, Process Control and Geotechnologies
Letná 1/9, 042 00 Košice, Slovakia
lubomir.ambrisko@tuke.sk, janka.saderova@tuke.sk, richard.antal@tuke.sk

Abstract: The subject of this paper is an analysis of the maintenance of the railway vehicles used in rail passenger transport. The analysis used data on the failure rates of engineless as well as driving vehicles, and it was carried out using reliability parameter indicators. The following parameters were determined for the individual vehicles: failure intensity, failure-free operation probability, failure probability, mean time between failures, mean maintenance time, maintenance intensity, equipment maintainability and technical readiness within the period of the two pre-pandemic years. Higher failure rate was confirmed of the operated railway engineless vehicles.

Keywords: railway vehicles; reliability; maintenance

1 Introduction

Maintenance is one of the key tools for ensuring the reliable operation of objects. The operational reliability assurance process continues throughout the technical life of any operated object [1, 2]. An increasing emphasis has been placed on the prevention of operational problems caused by technical failures. The development in the field of technical diagnostics proceeds from a simple search for the causes of proven failures, through regular preventive diagnostic inspections, controls or revisions, to the continuous automatic monitoring of the object's technical condition [3]. Due to a growing interest in increasing efficiency, reducing downtimes and costs, while increasing the operational reliability and safety, maintenance has become an essential process. Briefly, maintenance is a tool that increases the productivity, quality, overall efficiency and optimisation of the available objects, and that provides assurance of their operability and safety [2].

Reliability is a complex component of an object's quality, which is expressed, from a qualitative point of view, through partial properties, such as failure-free operation, maintainability and durability. These properties are expressed by

individual indicators. A failure-free operation means that the required functions are continuously fulfilled in predefined modes and conditions. Conversely, a failure means the discontinuation of the object's operability. The calculation is, therefore, based on two mutually exclusive conditions, i.e. the failure-free operation and the downtime caused by failure, which randomly alternate during the object's technical life [4].

Railway transport is one of the main modes of transport throughout the world, as it is reliable, efficient, cost-effective and comfortable [5]. It requires minimal costs for fuel, it is very safe and, above all, it facilitates the transportation of large quantities of loads in an environment-friendly manner [6]. Over the years, the rail freight industry has witnessed a progressive increase in the quantities being transported, which has contributed to an increased number of trains in operation, with higher speeds and higher axle loads [7]. The basic unit of rail freight transport is the freight wagon. Depending on the type of transport, there are several types of freight wagons available, while the transport of fluids, especially those that are hazardous or combustible, is carried out using railway cisterns [8]. The maintenance work is focused not only on wagons but also on traction engine vehicles and railway tracks. The maintenance of the tracks isn't today fully implemented from different objective reasons, e.g. the absence of sufficient funding [9]. Furthermore, a reliable railway infrastructure is a crucial precondition for the growth and development of governments, companies and societies [10]. The primary purpose of maintenance is to keep the equipment operable and to prevent or minimise the outages caused by failures [11].

The basic function of the maintenance of railway vehicles is to maintain their safe and cost-efficient operation throughout their lifecycle. It is, therefore, necessary to take all necessary measures aimed at avoiding the failures that might lead to serious consequences, such as a train derailing [12]. Maintenance plays a crucial role in the reliability and availability of the rail transport [13], while achieving high-quality maintenance requires choosing the correct maintenance strategy. In recent years, a number of different maintenance approaches have been developed, each one of them representing a different generational approach based on the current level of progress [14]. Corrective maintenance is required after a failure, while condition-based maintenance (CBM) is the type of maintenance that is carried out proactively before a potential failure occurs. In addition, opportunistic maintenance is carried out in an opportunistic manner, by analysing the potential maintenance opportunities that result from the downtime of other equipment [15]. Preventive maintenance is effective in reducing the number of losses caused by degradation [16] and it has a remarkably good performance in reducing downtime and avoiding excessive maintenance work. Carrying out both preventive repairs and general repairs based on the maintenance staging also facilitates a more flexible and cost-efficient assignment of maintenance tasks when compared to the strategy focused on instant general repairs [17]. Within the rail transport field, the purpose of predictive maintenance planning is to maximise

the reliability and availability of the transport vehicles, which makes scheduling the expected maintenance actions a vital part of operating a railway [18].

The service life of a traction vehicle is presumed to be approximately 30 years; however, in most cases the service lives of the individual components are not identical, as the body and the bogie will usually endure for longer [19, 20]. The costs of operation and maintenance increase as the vehicle ages. One of the options is to then replace these vehicles with new ones, which is, of course, very costly; while the other option is modernising them, which is both more acceptable and cheaper [12].

Regular maintenance is necessary in order to keep railway vehicles and their components in operation, as well as to eliminate failures and repairs. If maintenance actions are omitted or inadequate, this may cause delays, cancelled lines, dangerous situations and even fatal injuries. Such negative events are very important, as they clearly affect the competitiveness and profitability of a railway company. However, the complexity of maintenance processes is continuously increasing for several reasons. First, articulated trains are gradually replacing the trains consisting of a locomotive and passenger cars. This has led to more complex requirements for the scheduled maintenance. Furthermore, purely mechanical systems are gradually being replaced with complex mechatronic devices and systems that combine mechanical, electronic and information technologies with very different properties. Therefore, it is necessary to choose an optimal maintenance strategy and to assess its effectiveness in terms of the reliability, availability and cost of the system's durability. The evaluation of a maintenance strategy requires modelling the effects of potential alternative strategies on the system's performance and on the life cycle costs. However, due to the greater complexity of these systems and the increased performance requirements regarding the system's reliability, availability and safety, the complexity of a maintenance strategy evaluation has also increased. One of the methods of coping with this increased complexity and with the requirement for a higher availability of railway vehicles is through modular maintenance. The concept of the modular maintenance involves vehicles that are designed so that the replaceable units in the line may be replaced as a whole module, and can then be revised or repaired separately in a workshop in order to increase the availability of the vehicle. Nonetheless, despite the growing use of devices for monitoring the system's condition and the subsequent availability of the related data, the majority of the maintenance tasks in the railway industry are still carried out using the traditional maintenance concept [21].

In this paper, the maintenance methods of railway means of transport are discussed and its purpose is to evaluate the reliability parameters of engineless and driving railway vehicles.

2 Experimental Part

In this work, data from the operation and failure rates of engineless and driving railway vehicles were analysed using indicators of reliability parameters. These parameters were applied in railway transport for the purpose of identifying the complex component of quality from a qualitative point of view [4] using sub-features such as failure-free, maintainability and readiness, which were expressed through individual relationships of monitored indicators.

2.1 Maintenance of Wagons and Locomotives

Maintenance includes all activities whose role and purpose are to maintain railway rolling stock in operable condition. Readiness, failure-free and maintainability are related to maintenance. If we apply the given terms to railway passenger transport, we get to the regulation [22], which determines how to proceed with the maintenance of wagons and locomotives.

Safety is important for any type of transport, but it cannot be achieved without prescribed maintenance. During the maintenance of passenger cars, it is necessary to follow the legislative procedures that specify the requirements. Technology and technological procedures for the maintenance of railway rolling stock are in accordance with STN and EN standards, railway technical standards, laws and decrees of the Ministry of Transport of the Slovak Republic, regulations and documents of the transport company, its measures, directives and regulations, maintenance regulations as well as manufacturers' documentation, or repairmen. Technical inspections are carried out in accordance with the provisions of the Decree of the Ministry of Transport, Posts and Telecommunications of the Slovak Republic on the traffic regulations of railways No. 351/2010 Coll. In accordance with the decree, the intervals of technical inspections are shown in Table 1.

Table 1

Intervals of technical inspections of railway rolling stocks [22]

No.	Type of railway rolling stocks	Interval of technical inspections (years)
1	Driving vehicles	0.5
2	Motor, electric wagons and units, control wagons	1
3	Four-axle passenger wagons for domestic transport and electric and motor unit loaded wagons with speeds above 120 $km.h^{-1}$	1.5
4	Four-axle passenger wagons for domestic transport and electric and motor unit loaded wagons with speeds of up to 120 $km.h^{-1}$	2
5	Two-axle and narrow-gauge passenger cars	2

2.2 The Reliability Parameters

The reliability parameters were expressed for 69 driving vehicles (electric locomotives of two different product lines and engine locomotives of the same product line) and for approximately 900 railway engineless vehicles (used in rail passenger transport) while applying the exponential distribution of failures, in particular the Weibull distribution. The reliability parameters [23, 24] were expresses as follows:

(i) **Failure intensity** λ is determined by the number of failures that occur per unit time:

$$\lambda = \frac{a}{T} \tag{1}$$

wherein:

a is the number of failures (or repairs) over the whole operation period and

T is the total operation time (hours, months, years).

For a group of n units, λ is expressed as follows:

$$\lambda = \frac{\sum a}{nT} \tag{2}$$

wherein:

T is the total operation time for the group of units, either as the average operation time or as the sum of operation times of the individual monitored units (hours).

(ii) A correlation between the **probability of failure-free operation** P_{ff} and the failure probability P_{Pf} is important for predicting the operability of the units:

$$P_{ff} = 1 - P_{Pf} \tag{3}$$

wherein:

$P_{ff} = e^{-\lambda T}$ is the je probability of failure-free operation over time T.

(iii) **Mean time between failures** is expressed as the reciprocal of the failure intensity during the normal operation, when the failure intensity is constant:

$$T_m = \frac{1}{\lambda} = \frac{T}{a} \tag{4}$$

(iv) **Mean maintenance time** $\varnothing$ is the ratio of the total time of maintenance actions and repairs to their quantity:

$$\phi = \frac{\sum t}{a} \tag{5}$$

wherein:

Σt is the sum of times of all maintenance actions and repairs and

a is the number of all maintenance actions and repairs.

(v) **Maintenance intensity** μ is the reciprocal of the mean maintenance time:

$$\mu = \frac{1}{\emptyset} \tag{6}$$

(vi) **Maintainability** M is the property of the units which facilitates prevention or fast elimination of the wearing process consequences through maintenance and repairs. It is probabilistic in nature and described by the following equation:

$$M = 1 - e^{-\mu t} \tag{7}$$

wherein:

$e^{-\mu t}$ is the probability that in time t no maintenance action will be carried out.

(vii) **Technical readiness of the unit**, expressed by the coefficient of technical readiness C_{tr}, is the ability of the unit to fulfill the required function at any moment in time:

$$C_{tr} = \frac{\text{Probable failure-free operation}}{\text{Total monitored period}} .100\ [\%] \tag{8}$$

3 Results and Discussion

The evaluation of the reliability of railway rolling stock from the point of view of quality, and from the point of view of quantity, as well as the mathematical expression of the indicators of these properties is the subject of this work. The calculation of indicators is based on two mutually exclusive situations, i.e. from the state of fault-free operation and the state of fault downtime, which randomly alternate during operation [4]. The individual reliability parameters of railway engineless vehicles (cars) and driving vehicles (locomotives) were identified using equations (1) through (8) based on the data obtained during the use of these vehicles in years 2018 and 2019.

3.1 Car Reliability Parameters

Tables 2 and 3 contain the maximum and minimum average values of the individual parameters of cars in year 2018 and in year 2019.

Table 2
Car reliability parameters in 2018

2018	λ (days^{-1})	P_{ff}	P_{Pf}	T_m (days)	$Ø$ (days)	μ (days^{-1})	M	C_{tr} (%)
Max.	5.0	1.0	1.0	359.0	111.0	1.0	1.0	100.0
Average	0.032	0.144	0.856	56.2	3.18	0.572	0.856	23.8

Table 3
Car reliability parameters in 2019

2019	λ (days^{-1})	P_{ff}	P_{Pf}	T_m (days)	$Ø$ (days)	μ (days^{-1})	M	C_{tr} (%)
Max.	2.0	1.0	1.0	347.0	202.0	1.0	1.0	100.0
Average	0.059	0.083	0.917	24.8	4.85	0.526	0.917	11.4

Figures 1 through 12 show all the obtained reliability parameters for the group of 925 cars in year 2018 and for 892 cars in year 2019. Figure 1 shows a graphic representation of the failure intensity (λ) of all cars in year 2018. The values are expressed in days^{-1}. The highest failure intensity was observed for the car of a certain production line.

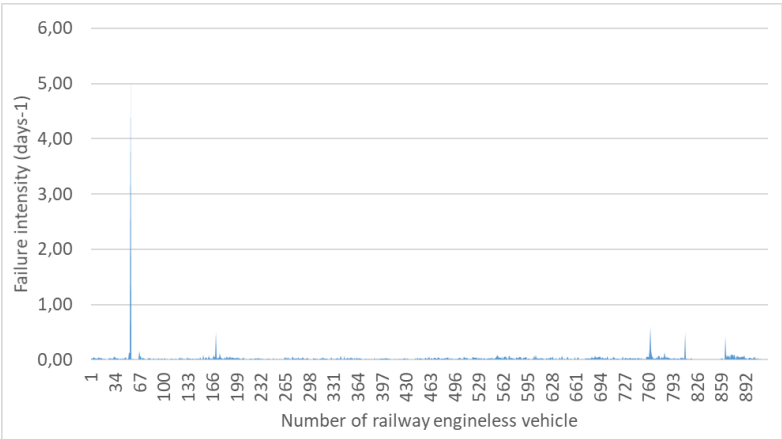


Figure 1
Failure intensity in 2018

The failure intensity (λ) of all cars in year 2019 is shown in Figure 2. The values are expressed in days^{-1}. The highest failure intensity was observed for the car of a product line different from the one with the highest value observed in 2018. The comparison clearly indicates that in 2019 the failure intensity was higher than in 2018.

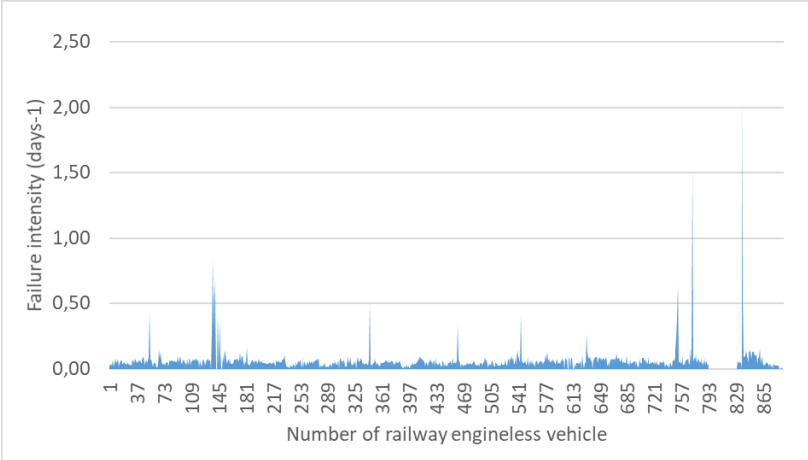


Figure 2
Failure intensity in 2019

Failure-free operation is a feature consisting in the continuous performance of the required function in prescribed modes and under specified conditions, which is quantified by the probability of failure-free operation. Figure 3 shows a graphic representation of the probability of failure-free operation (P_{ff}) in 2018 while Figure 4 shows the equivalent data for 2019. The average probability of failure-free operation was higher in 2018, corresponding to the observed lower failure intensity.

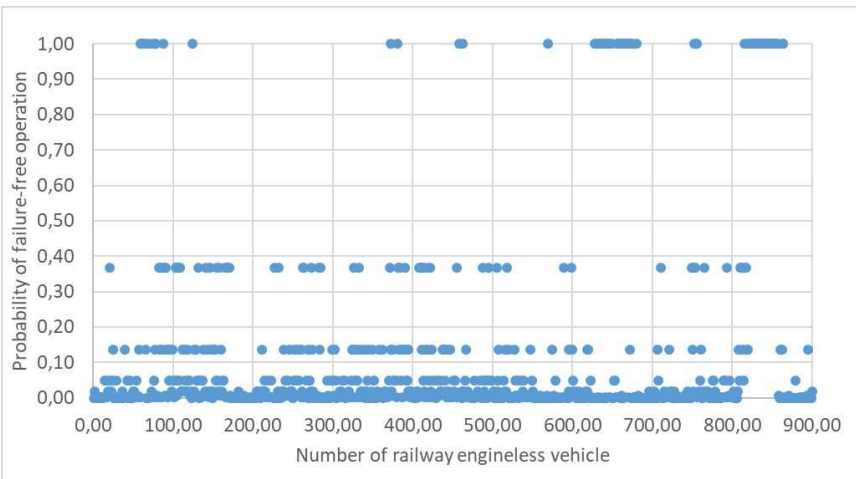


Figure 3
Probability of failure-free operation in 2018

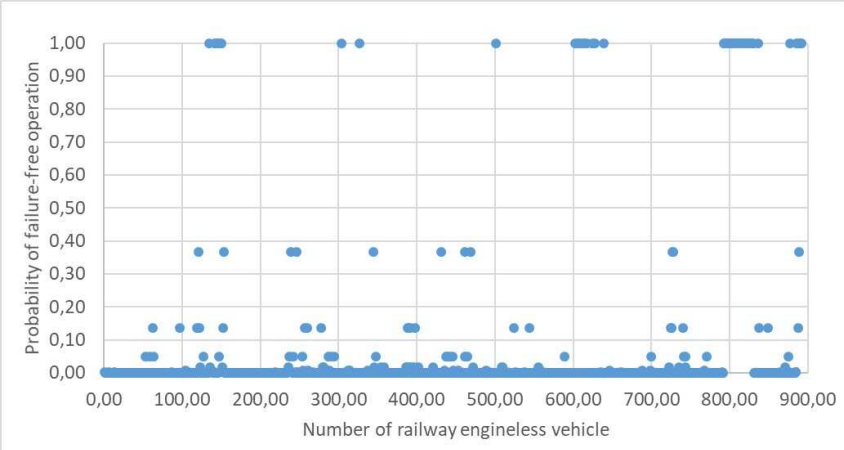


Figure 4
Probability of failure-free operation in 2019

Figure 5 shows the mean time between failures (T_m) in 2018, expressed in days. The highest observed value was 359 days; this is the probable period of the failure-free operation. Figure 6 shows the mean time between failures (T_m) in 2019 in days. The highest value (347 days) was observed for the car of the same product line as the one with the highest value observed in 2018.

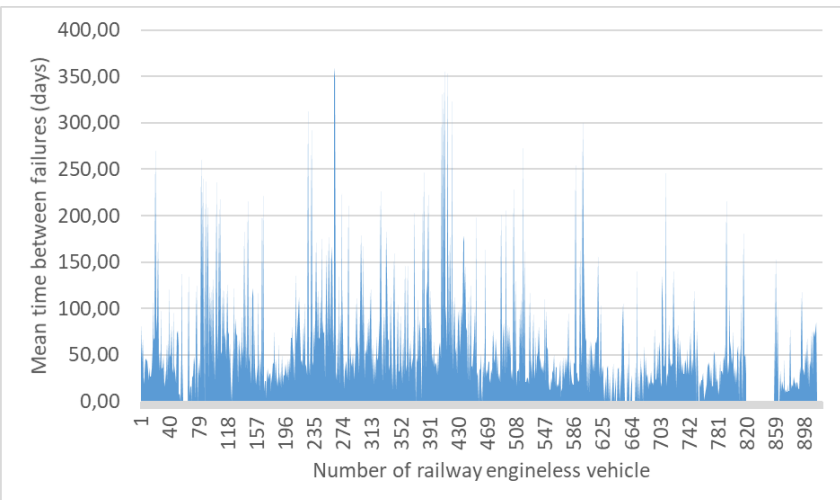


Figure 5
Mean time between failures in 2018

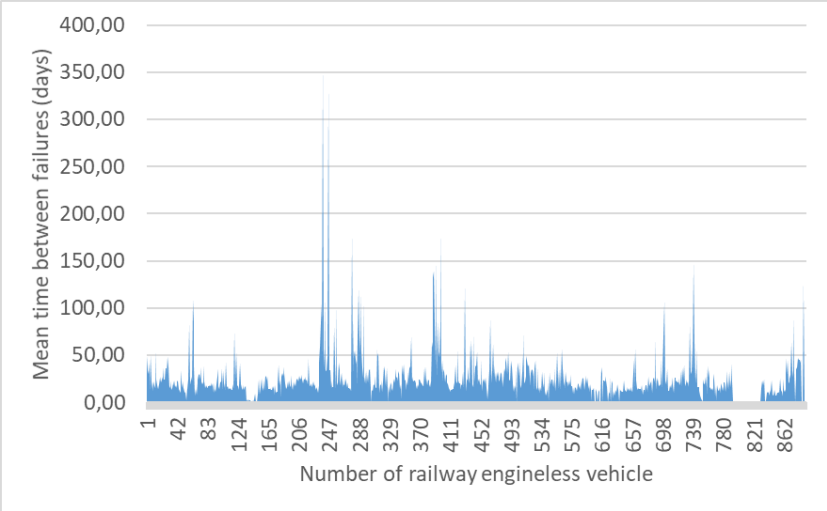


Figure 6
Mean time between failures in 2019

The exponential correlation between the mean maintenance time (φ) and maintenance intensity (μ) is presented in Figures 7 and 8 for years 2018 and 2019, accordingly. Longer mean maintenance time with lower maintenance intensity was observed in 2019.

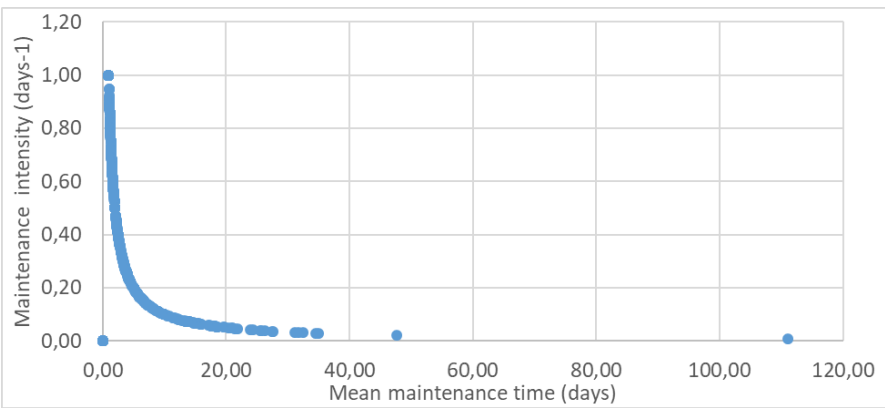


Figure 7
Exponential correlation between the mean maintenance time and maintenance intensity in 2018

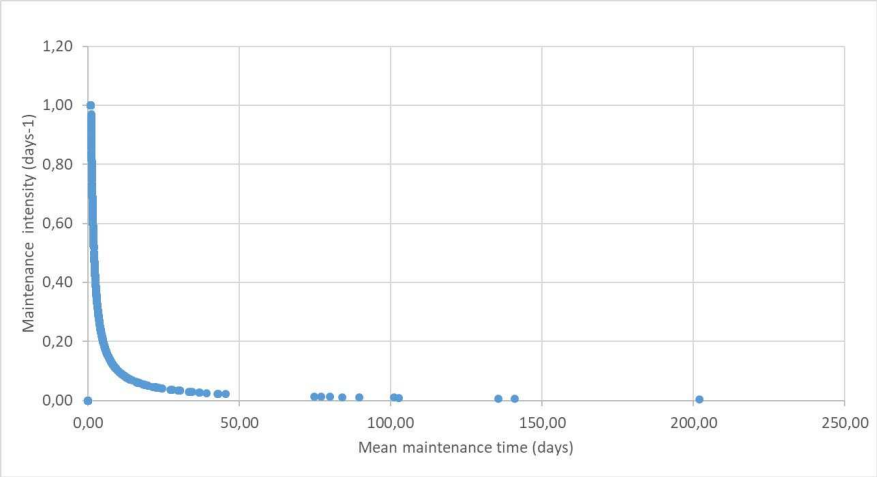


Figure 8
Exponential correlation between the mean maintenance time and maintenance intensity in 2019

Maintainability (*M*), i.e. the ability that consists in the ability to prevent breakdowns in the form of prescribed maintenance, in years 2018 and 2019 is shown in Figures 9 and 10. Higher average maintainability was observed in 2019.

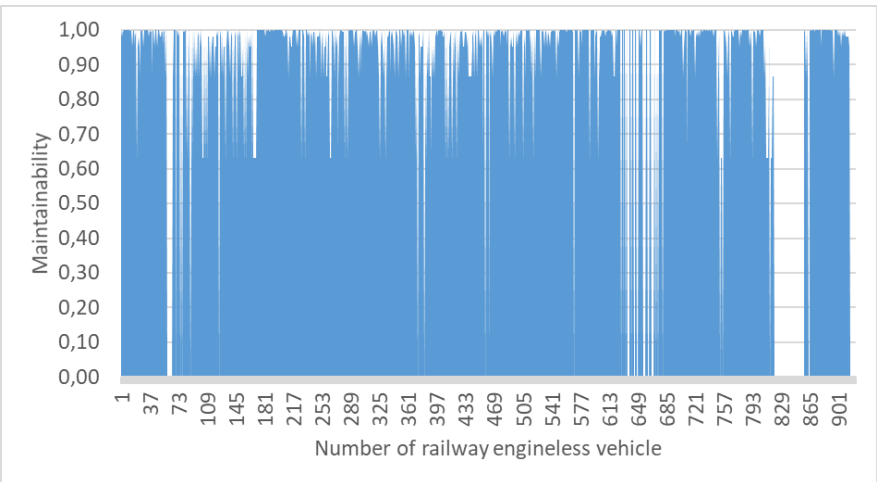


Figure 9
Maintainability in 2018

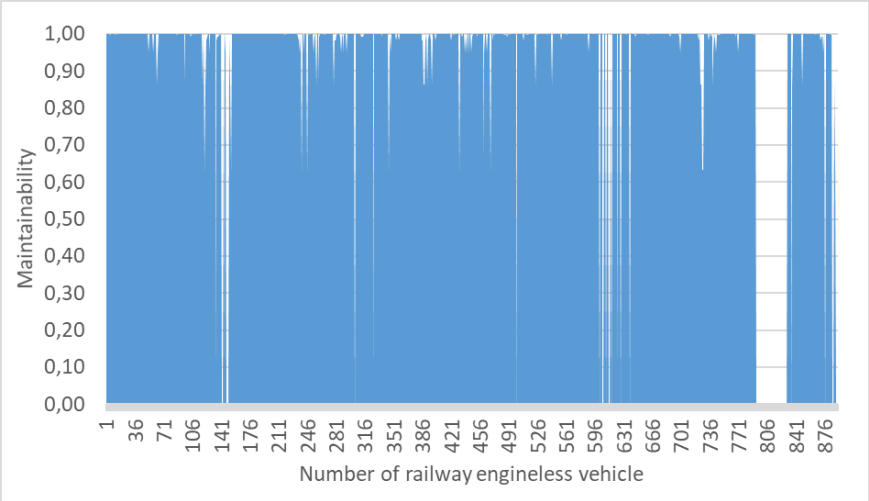


Figure 10
Maintainability in 2019

Figures 11 and 12 show the coefficient of technical readiness (C_{tr}) in years 2018 and 2019. A higher coefficient of technical readiness was confirmed for year 2018. The coefficient considers not only the time of trouble-free operation, but also the time required for repair and maintenance, but it does not allow assessing the size of continuous trouble-free operation.

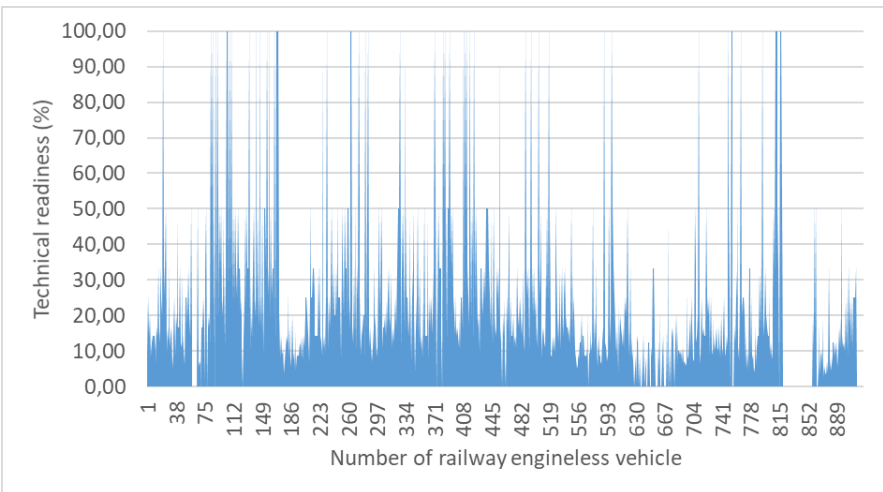


Figure 11
Coefficient of technical readiness in 2018

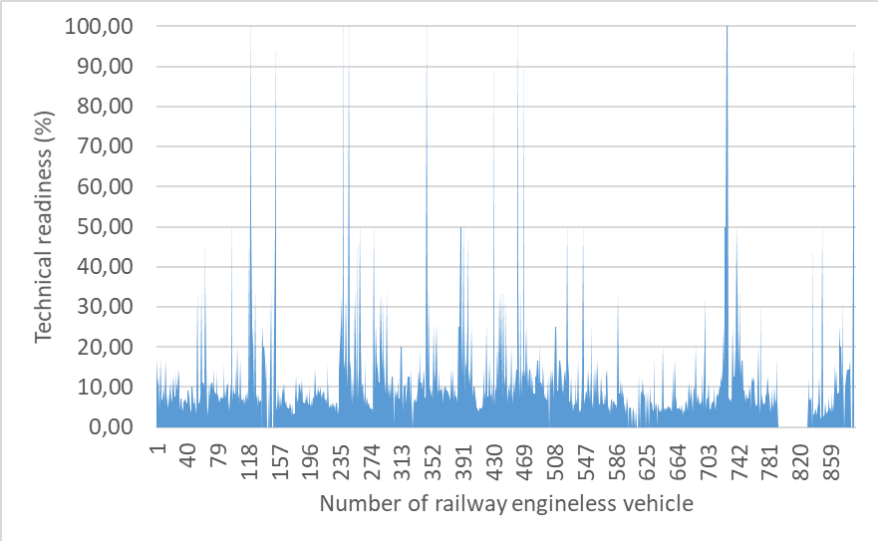


Figure 12
Coefficient of technical readiness in 2019

3.2 Locomotive Reliability Parameters

Tables 4 and 5 contain the average monthly values of indicators for locomotives in years 2018 and 2019 (bold figures are the maximum values). The calculations were made for 69 locomotives of three product lines.

Table 4
Locomotive reliability parameters in 2018

Average values in individual months (2018)								
Month	λ (h^{-1})	P_{ff}	P_{Pf}	T_m (h)	$Ø$ (h)	μ (h^{-1})	M	C_{tr} (%)
I.	0.212	0.877	0.123	0.788	6.042	0.941	0.867	3.886
II.	0.246	0.872	0.129	0.755	6.237	1.027	0.846	**4.825**
III.	0.372	0.861	0.139	0.628	6.120	1.203	**0.899**	2.234
IV.	0.270	0.862	0.138	0.730	7.134	1.060	0.860	4.124
V.	0.278	0.855	**0.145**	0.723	6.462	0.837	0.862	3.774
VI.	0.329	0.855	**0.145**	0.672	5.171	1.084	0.881	4.216
VII.	**0.396**	0.870	0.130	0.604	5.596	1.166	0.875	4.133
VIII.	0.258	0.869	0.131	0.742	6.239	0.964	0.861	4.226
IX.	0.265	0.858	0.142	0.735	6.063	1.263	0.886	4.037
X.	0.242	0.855	**0.145**	0.758	6.442	**1.301**	0.859	3.940

XI.	0.199	0.877	0.123	**0.802**	6.954	1.009	0.842	3.651
XII.	0.320	**0.879**	0.121	0.680	**7.690**	1.046	0.845	3.396
Yearly average	0.282	0.866	0.134	0.718	6.346	1.075	0.865	3.870

Table 5

Locomotive reliability parameters in 2019

Average values in individual months (2019)								
Month	λ (h^{-1})	P_{ff}	P_{Pf}	T_m (h)	$Ø$ (h)	μ (h^{-1})	M	C_{tr} (%)
I.	0.243	**0.896**	0.104	0.757	**7.466**	0.903	0.838	3.590
II.	**0.531**	**0.896**	0.104	0.469	6.775	0.918	0.827	4.398
III.	0.378	0.865	**0.135**	0.622	6.157	1.320	**0.874**	2.950
IV.	0.202	0.878	0.122	0.798	5.750	**1.374**	0.860	3.823
V.	0.235	0.869	0.131	0.765	4.973	1.101	0.870	4.292
VI.	0.219	0.885	0.115	0.781	5.197	1.118	0.849	**4.681**
VII.	0.202	0.881	0.119	0.798	6.703	1.023	0.855	4.037
VIII.	0.234	0.886	0.114	0.766	6.342	1.185	0.836	3.353
IX.	0.207	0.891	0.109	0.794	5.162	1.101	0.855	3.674
X.	0.200	0.878	0.122	0.800	6.035	1.296	0.852	3.137
XI.	0.199	0.888	0.112	**0.801**	5.682	1.351	0.861	3.120
XII.	0.203	0.879	0.121	0.797	5.138	1.157	0.873	2.910
Yearly average	0.254	0.883	0.117	0.746	5.948	1.154	0.854	3.664

In 2018, the failure intensity of the monitored locomotives reached a higher value than the value observed in 2019. A comparison of the individual calendar months didn't reveal any evident trend or a period that would indicate a higher failure rate caused by weather conditions. In 2019, the probability of the failure-free operation exhibited higher values. In the same year, the failure probability remained above 1.10 in all of the months, while in 2018 its values were higher. The mean time between failures and the mean maintenance time exhibited certain fluctuations during the monitored period, but the average values observed in the individual years did not exhibit any statistically significant differences. In 2018, the mean maintenance time was the highest in month XII, while in 2019 it was in month I. In both years, the mean time between failures reached the peak in month XI The intensity of locomotive maintenance was higher in 2019 than in 2018, while the minimum values correspond to the maximum values of the mean maintenance time. The maximum maintenance intensity was observed in month IV of 2019 and in month X of 2018. Higher average maintainability and a higher average coefficient of technical readiness were observed in 2018.

Conclusion

In recent years, the rail transport industry has witnessed increased quantities of transported loads. This has resulted in the deployment of more vehicles and in the necessity to maintain the vehicles' operability with a minimum occurrence of downtime. This paper presented the reliability parameters of railway engineless and driving vehicles observed in years 2018 and 2019, i.e. in the period before the COVID-19 pandemic. In the following year, i.e. in 2019, the observed values of these reliability parameters indicated higher failure rates of the operated railway engineless vehicles. This may have been caused by the fact that despite the availability of the data on the condition of rail cars, the prevailing type of maintenance was the conventional maintenance concept. A solution to this is to increase the proportion of preventive maintenance, which is efficient in reducing the losses caused by mechanical as well as chemical wear that leads to gradual degradation.

As for the engine railway vehicles, it is evident that the prevention of operational problems caused by technical failures resulted in a lower failure rate in 2019. Preventive diagnostic inspections have led to higher operational reliability of the traction vehicles.

Acknowledgment

Supported by the Scientific Grant Agency (VEGA) of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences, Grant No. VEGA 1/0429/18.

References

[1] Ambriško, Ľ.: Analysis of public transport vehicles maintenance costs, CLC 2018 Carpathian Logistics Congress, Ostrava, 2019, pp. 368-373

[2] Helebrant, F.: Technical Diagnostics and Reliability: IV. Operation and Maintenance of the Machine (in Czech). Ostrava: Technical University of Ostrava, 2015

[3] Grenčík, J. et al.: Maintenance Management II: Synergy of Theory and Practice (in Slovak). Košice: TU, 2020

[4] Marková, P.: Reliability, failure-free and maintainability of measuring devices (in Slovak). Automa 2008; 4(2), pp. 58-59

[5] Andrusca, M., Adam, M., Dragomir, A., Lunca, E.: Innovative integrated solution for monitoring and protection of power supply system from railway infrastructure. Sensors 2021; 21: 7858

[6] Shah, A. A., Bhatti, N. A., Dev, K., Chowdrry, B. S.: MUHAFIZ: IoT-based track recording vehicle for the damage analysis of the railway track. IEEE Internet of Things Journal 2021; 8, pp. 9397-9406

[7] Ribeiro, D., Bragança, C., Costa, C., Jorge, P., Silva, R., Arêde, A., Calçada, R.: Calibration of the numerical model of a freight railway vehicle based on experimental modal parameters. Structures 2022; 38, pp. 108-122

[8] Kuba, E., Moravčík, M., Pavelčík, V.: Examination of structural properties of a new railway wagon design for flammable substance transportation. Transportation Research Procedia 2021; 55, pp. 760-766

[9] Kurhan, M., Kurhan, D., Hmelevska, N.: Maintenance reliability of railway curves using their design parameters. Acta Polytechnica Hungarica 2022; 19(6), pp. 115-127

[10] Garmabaki, A. H. S., Thaduri, A., Famurewa, S., Kumar, U.: Adapting railway maintenance to climate change. Sustainability 2021; 13: 13856

[11] Hardt, F., Kotyrba, M., Volna, E., Jarusek, R.: Innovative approach to preventive maintenance of production equipment based on a modified TPM methodology for industry 4.0. Applied Science 2021; 11: 6953

[12] Zvolenský, P., Barta, D., Grenčík, J., Droździel, P., Kašiar, Ľ.: Improved method of processing the output parameters of the diesel locomotive engine for more efficient maintenance. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2021; 23(2), pp. 315-323

[13] Birolini, A.: Reliability Engineering: Theory and Practice. Heidelberg: Springer, 2007

[14] Davari, N., Veloso, B., Costa, G. d A., Pereira, P. M., Ribeiro, R. P., Gama, J.: A Survey on data-driven predictive maintenance for the railway industry. Sensors 2021; 21: 5739

[15] Binghai, Z., Yi, Q., Yuwang, L.: Proactive preventive maintenance policy for buffered serial production systems based on energy saving opportunistic windows. Journal of Cleaner Production 2020; 253: 19791

[16] Luo, M., Wu, S.: A mean-variance optimisation approach to collectively pricing warranty policies. International Journal of Production Economics 2018; 196, pp. 101-112

[17] Wu, T., Ma, X., Yang, L., Zhao, Y.: Proactive maintenance scheduling in consideration of imperfect repairs and production wait time. Journal of Manufacturing Systems 2019; 53, pp. 183-194

[18] Rokhforoz, P., Fink, O.: Hierarchical multi-agent predictive maintenance scheduling for trains using price-based approach. Computers & Industrial Engineering 2021; 159: 107475

[19] Stastniak, P., Moravcik, M., Baran, P., Smetanka, L.: Computer aided structural analysis of newly developed railway bogie frame. MATEC Web of Conferences 2018; 157: 02051

[20] Xiu, R., Spiryagin, M., Wu, Q., Yang, S., Liu, Y.: Fatigue life assessment methods for railway vehicle bogie frame. Engineering Failure Analysis 2020; 116, pp. 1-19

[21] Eisenberger D., Fink, O.: Assessment of maintance strategies for railway vehicles using Petri-nets. Transportation Research Procedia 2017; 27, pp. 205-214

[22] 2_19_OIK_v25: Maintenance of driving rolling stock, electric and motor wagons, units, connecting, embedded and control wagons (in Slovak), 2018

[23] Pačaiová H.: Maintenance management (in Slovak). Košice: TU, 2006

[24] Stastniak, P., Moravcik, M., Baran, P., Smetanka, L.: Computer aided structural analysis of newly developed railway bogie frame. MATEC Web of Conferences 2018; 157: 02051

Multi-Method Risk Assessment Process for Sustainable Business – A Compliance Research Follow-up Case Study

Ferenc Bognár, Petra Benedek

Department of Management and Business Economics, Budapest University of Technology and Economics Magyar tudósok körútja 2, H-1117 Budapest, Hungary; bognar.ferenc@gtk.bme.hu; benedek.petra@gtk.bme.hu

Abstract: Assessing and reducing compliance risks for competitiveness, business continuity, and organizational sustainability requires sound methodologies. This study aims to characterize a multi-method assessment process that can show the experts the complex assessment options. This paper presents three multidimensional assessment methods following a case study at a commercial bank. The decision-making patterns related to the experts and assessment methods can be identified. Based on the best fit of the patterns, a possible solution can be offered for designing actions to reduce the risk of noncompliance, providing guidance for improvement aligning with organizational objectives to increase business continuity and sustainability. The results of this study can serve as a methodological input for compliance risk assessment in banks and other organizations in highly regulated sectors.

Keywords: risk assessment; compliance risk; Risk Matrix; Failure Mode and Effect Analysis; FMEA; Partial Risk Map; PRISM

Introduction

Reputation, trust, reliability, and ethical behavior are as relevant issues as decades ago [1]. Partnerships and business cooperation are fundamentally based on mutual trust and commitment. Since business is embedded in the social structures (like the legal environment), companies need business models that provide good quality products to ensure competitiveness and are also sustainable from a social and environmental perspective [2]. Modern societies express expectations for sustainable development, which means becoming more sustainable with incremental development [3].

The United Nations' Sustainable Development Goals (SDG) promote sustainable economic growth, higher productivity, and technological innovation. As a result, companies increasingly reflect the SDGs in their business models and strategies,

integrating sustainability into core business [4]. As a result of intense legal pressure, preparing a sustainability report has become a legal obligation for large companies in the European Union [5]. This report makes sustainability a compliance issue. The European Commission has defined Key Performance Indicators with the legislation [6, 7].

The compliance function ensures legal compliance in a highly complex environment with various actors. Compliance means complying with and enforcing the legal regulations that apply to a company, including the owner's decisions. One of compliance management's main tasks is ensuring the coherence of external and internal regulations. In addition, building trust and confidence is the underlying premise of all compliance activities [8, 9].

As part of the organization's internal control system, corporate compliance is responsible for identifying and managing compliance risks at the organizational level. Therefore, the compliance department requires functional independence, strong senior management support, and a strong network of contacts. In addition, prioritizing compliance issues requires increased attention from management and compliance officers. This study facilitates the understanding of the risk assessment processes that contribute to decision-making and executive behavior.

Controls aim to ensure the orderly, economical, efficient, and effective management of assets and the proper fulfillment of reporting obligations. The "lines of defense" describe the internal control system developed based on EU legislation in the financial sector. New risks appearing because of technological advancement or other significant changes must be fully understood and regulated adequately by authorities [10]. In addition, the challenges of the past three years (such as rapid changes in customer expectations and pandemic measures) have forced many actors to rethink their business and risk models and contributed to the closer integration of risk management and operational management.

The banking sector is one of the most regulated industries, has a significant influence on the economy, and has considerable experience managing regulatory risks. Banks operate in a global environment with rapidly evolving national and international regulation and supervision (including measurement and quantification of operational risk) to ensure a well-functioning banking system and promote stability. Financial institutions have developed compliance frameworks, advanced approaches, and techniques. However, non-financial risks arising from operations (processes and systems) are similar to those outside the financial sector [11]. The changing regulation forces businesses to reassess risks, strategies and action plans to comply with legal requirements. Companies can better adapt to a changing environment by developing a practical approach to managing risks and strengthening overall flexibility.

This paper focuses on the importance of individual risk assessment, using more than one risk assessment method in a complex environment, and provides valuable insights into compliance risk assessment. The purpose of this paper is to present a

detailed, multi-method risk evaluation process. The paper is organized as follows. Section 1 presents the compliance risk assessment background of the study. Section 2 introduces the methodology and the empirical research, including the materials, while Section 3 presents the results. In Section 4, the results are discussed, highlighting managerial implications. Finally, the conclusions summarize this paper and propose directions for future research.

1 Compliance Risk Assessment

Compliance risks are various, conditional, and unique to each organization. Their assessment is a complex issue since experts can evaluate compliance risks on many terms from several perspectives. If risk assessment is carried out regularly, trends can provide information on strengths and areas to be developed.

As a compliance management system is growing past its initial phase, there is a shift of focus from mere compliance incident detection and reacting to incidents to preventing noncompliance and a proactive culture [12, 13].

Compliance risk assessment forms the basis for implementing compliance management systems and allocating appropriate resources and processes to address identified compliance risks [14].

The compliance risk assessment process has three well-distinguished steps: (1) identifying compliance obligations and related compliance risks, like permits, contractual relationships, corruption, fraud, and the industry or quality standards. Secondly, (2) analysis of the probability (likelihood of occurrence) and the consequences of noncompliance, like environmental damage, economic loss, damage to reputation, and administrative burdens. Finally, (3) risk evaluation includes ranking and prioritization of risks.

Risk identification is a systematic activity that reveals how compliance obligations manifest in activities, products, services, and other relevant aspects of operations (such as digital transformation). Identifying compliance risk involves identifying compliance resources and situations and creating a database of these. If there is outsourced activity (i.e., HR processes), the responsibility for compliance still lies with the outsourcing organization. Therefore, outsourced and third-party processes must also be examined for risk identification. In addition, changing circumstances (like pandemic or war situation) or organizational context (like mergers or entering new markets) triggers new risk identification.

Compliance risks are to be analyzed and evaluated regularly, at least yearly. Furthermore, even a single noncompliance event or a so-called "near miss" (or "close call"), when no negative consequence (like an interruption to operations) followed an incident, should trigger the reassessment process. Furthermore, the extent and level of compliance risk assessment should be proportionate to the

organization's context, size, and objectives. Appropriate techniques for the risk-based approach are listed in Annex B of IEC 31010:2019, containing 31 risk assessment techniques, including Failure Mode and Effect Analysis (FMEA) [15]. Finally, risk assessment is not a collection of static practices but should evolve considering external changes in the business environment [11].

Risk evaluation is the last step of the risk assessment process. The analysis is followed by ranking and prioritization of the compliance risks. FMEA, Risk Matrix and Partial Risk Map are examples of techniques that can serve as a basis for developing risk reduction strategies, action plans and allocating resources.

The primary motivation of this study is to characterize a multi-method assessment process, which can show experts the complex options for their compliance assessment work. Based on the proposed assessment process, decision-making patterns related to experts and assessment methods can be identified. Finally, based on the best fit of the patterns, a possible solution can be offered to design actions that reduce the risk of noncompliance. The research question of this paper is:

Which group assessment method (nominal group or focus group) gives the closest result to individual assessments in the compliance risk ranking process?

Two assumptions are examined in this paper.

Assumption 1. The aggregation of individual rankings shows closer results to individual rankings than discussion-based rankings. (A1)

Assumption 2. In the case of Risk Priority Number, the consensus level will be significantly higher than in the case of Risk Exposure and PRISM number. (A2)

2 Materials and Methods

Section 2.1 briefly introduces the applied risk assessment methodologies in the case study, just like Risk Matrix (RM), Failure Mode and Effects Analysis, and Partial Risk Map (PRISM). This section also briefly introduces the comparative analyses' statistical methods (Spearman's rho rank correlation and Kendall's W rank concordance coefficients). Finally, in Section 2.2, the risk assessment process flow and the characteristics of the data are presented.

2.1 Methods

Several risk assessment methods can be applied when the risk assessment process is based on multi-criteria. For example, according to the Basel Committee [16], the risk of noncompliance in the banking sector can be estimated through several criteria. Thus, applying multi-criteria risk assessment tools can be relevant in the ranking process of the case study. In this study, three significantly different risk assessment methods are applied systematically to analyze the presented cases.

These methods are the Risk Matrix [17], the Failure Mode and Effect Analysis [18], and the Partial Risk Map [19]. In the following, these methods and their applied parameters are interpreted briefly.

As visible in Figure 1, the applied risk assessment techniques have different evaluation structures. While RM is a one-time two-dimensional and FMEA is a one-time three-dimensional, PRISM is a three-time two-dimensional assessment methodology. While the assessment process of RM and FMEA applies only one assessment (based on 2 or 3 dimensions) [20, 21], PRISM applies three assessments (each based on two of the dimensions) [19]. The "severity of consequences" (*S*) dimension and the "probability of occurrence" (*O*) are typically involved in all three assessment techniques. In contrast, the "undetectability" (*D*) dimension is only applied by the FMEA [22] and PRISM methods [19]. The higher the value of *O*, *S*, and *D*, the higher the factor related to the risk of the incident. Please note that the lower the probability that a failure mode (faulty condition) is detected during regular operation, the higher the risk of non-detection.

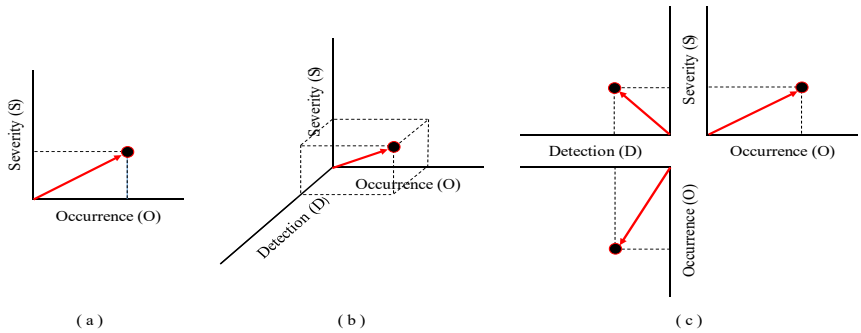


Figure 1

The structure of RM is shown by part (a), FMEA is shown by part (b), and PRISM is shown by part (c)

The role of aggregation functions is essential in the ranking result of the different methods. Based on [23], three typical aggregation functions are interpreted in these methods. The addition-based aggregation results linear, the multiplication-based aggregation results convex, while the sum-of-squares-based aggregation results in concave iso-risk lines for the ranking process. In the RM and PRISM methods, the addition and multiplication functions are used for the aggregation. In the case of FMEA, the multiplication function is often applied for aggregation [23, 24].

The purpose of applying the three methods is the same, to identify the possible or existing incidents by assessing the risks of the incidents. Once an incident is diagnosed as intolerably risky, risk reduction action is planned and launched to reduce the risk to a tolerable level [19]. The steps of evaluation and prioritization of risky incidents are different in the case of each method. Since these steps are not the focus of the study, a more profound introduction to different risk reduction processes of the different methods can be omitted. In the following, the definition of the incidents and the applied risk assessment functions are described.

Denote as m:=(o, s, d) an incident that has three risk characteristics: o probability of occurrence (O), s severity of consequences (S), and d undetectability (D). The characteristics have the following values, $o \in [1, 2, \cdots, i]$, $s \in [1, 2, \cdots, j]$ and $d \in [1, 2, \cdots, k]$. The scale lengths of o, s, and d are traditionally ten units in the case of FMEA, while four or five units in most cases of two-dimensional methods. Therefore, for all the incidents, the aggregate risk value can be calculated based on the values of the selected set of o, s, and d. The higher the value of the aggregation result, the higher the entire risk related to the incident.

As previously discussed, RM is a two-dimensional methodology built up to assess the risk of a particular process or object. The RM estimates the probability of occurrence (O) and the severity of consequences (S) of the specific process or object at the same time [20]. As previously noted, the addition aggregation function is applied in the case of RM and PRISM, while multiplication is used for the aggregation function of FMEA. Based on Equation 1, the Risk Exposure (RE) of a particular m incident can be calculated.

$$RE(m) = o + s \tag{1}$$

Since FMEA has one assessment vector and is a three-dimensional method, the indication of risk is compressed into an index called Risk Priority Number (RPN) [22]. Based on Equation 2, the RPN of a particular m incident can be calculated.

$$RPN(m) = o \times s \times d \tag{2}$$

The PRISM method applies a maximal value selection function over the three aggregation functions. The maximal value selection function can compress the risk level into the so-called PRISM number. Therefore, this method has three different partial risks related to the three two-dimensional assessments. Based on the PRISM number, the highest partial risks can be estimated [23]. Based on Equation 3, the PRISM number of a particular m incident can be calculated.

$$PRISM(m) = \max(o + s; o + d; s + d) \tag{3}$$

Based on the RE(m), RPN(m), and PRISM(m) values, the risk-based ranking of the incidents can be executed. Since the fundamental equations of the methods are different, the emerging rankings can also be different. Based on analyses of the differences between the rankings, information can be provided related to the topic of a research study. This paper uses the Spearman rank correlation analysis [25] to estimate the differences between paired rankings. Furthermore, the Kendall rank concordance analysis [26] is applied to estimate the differences between more than two rankings. Both analyses are often applied in focus-group-based qualitative assessments since the applied coefficients of the analyses can compare a minimal number (even just two) of records to each other [27]. A 5% significance level is applied in the analyses of both coefficients.

Spearman's rank correlation analysis is a statistical method usually applied to describe the strength and direction of a relationship between two variables. Spearman's rank correlation coefficient value ranges between -1 and 1. In the case of the same rankings, Spearman's rho equals 1. In the case of opposite rankings, Spearman's rho is -1. In the case of independent rankings, Spearman's rho is 0. In this study, Spearman's rank correlation coefficient was applied to analyze the pairwise similarities of the two rankings.

Kendall's rank concordance analysis is a non-parametric test. This statistic is often a prerequisite test before aggregating individual evaluations into group results. Based on the value of Kendall's W coefficient, the difference between at least two ranks can be characterized. The statistic is often used to compare different judgments to each other in social and economic sciences [26, 28]. The value of the coefficient is between 0 and 1. If the rankings are the same, the value of the coefficient is 1. If the rankings are the opposites, the coefficient is 0. The coefficient is generally used to interpret the agreement level of several judges. In the case of low coefficient values, the ranks are regarded as random, so different judgments cannot be aggregated.

2.2 Materials

In this subsection, materials are presented briefly, corresponding to the steps of the risk assessment process shown in Figure 2. Results from previous studies are also presented, as some serve as a relevant input for this study. The proposed risk assessment process consists of nine steps. From step 1 to step 4, the aim is on the details of group formation and assessment characteristics. Then, in step 5, the data validation process is performed. These five steps and their results are the subjects of previous studies [19, 27]. A brief description of the steps can be found in the following paragraphs.

In steps 1 and 2, case collection and expert selection are the focus. The assessment was launched at one of the top 5 largest Central and Eastern European commercial banks [19, 27]. Many possible compliance risks were described and presented by the bank. All the cases were related to the bank branch administration processes and focused on the characteristics of a possible non-compliant decision of the bank administrators. Any risk-taking behavior can open up additional opportunities for employee wrongdoing. Six risks were randomly selected for the process interpretation and results of the case study [19, 27]. The delegated compliance experts of the bank had at least ten years of experience in the field and were employed in the bank's headquarters. Two moderators also participated in the work of the focus group. The experts performed the assessment based on the *O*, *S*, and *D* dimensions, with four-four different factor values presented in [27].

In step 3, the focus group members had to assess the compliance cases individually [27]. Then in step 4, the group-decision assessment was delivered, including all the focus group members [19].

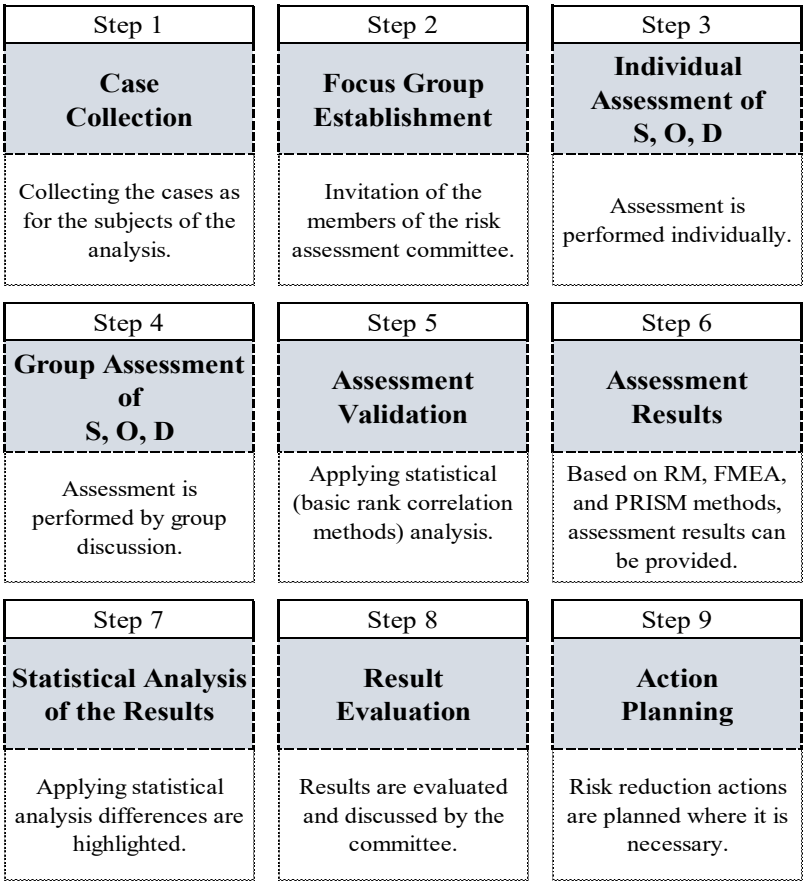


Figure 2
The designed risk assessment process

Table 1
The summary of individual and group assessments

	Individual assessments									Group assessments					
	Expert 1			Expert 2			Expert 3			Discussion-based			Aggregation of individual results		
Case	S	O	D	S	O	D	S	O	D	S	O	D	S	O	D
A	3	2	4	3	3	4	3	3	4	3	2	4	3	2,67	4
B	1	4	2	2	4	2	1	4	3	1	4	3	1,33	4	2,33
C	3	3	4	3	4	3	2	3	4	2	2	4	2,67	3,33	3,67
D	3	1	4	3	2	4	3	1	3	2	1	4	3	1,33	3,67
E	1	2	1	1	3	2	1	2	2	1	3	2	1	2,33	1,67
F	2	3	1	1	3	2	1	3	2	2	4	4	1,33	3	1,67

Table 1 includes the individual assessments, the discussion-based assessments, and the assessments based on aggregating the individual results.

As for step 5, the validation of the assessment results was executed. Since the subjective judgment rules of the experts can be significantly independent of the proposed ranking tables in [27], the assessment results had to be validated first. The offered validation process was described [27] based on an expert-by-expert comparison technique, and in the study, it was observed that the experts' rankings have no significant differences:

- the individual assessments of the experts could be aggregated (See the last three columns in Table 1);
- and the applied scales for assessing *S*, *O*, and *D* were interpreted as valid assessment scales.

Figure 3 shows the boundaries of the preceding studies and this article. The results related to the discussion-based PRISM method are described in detail in [19], while the aggregation-based FMEA results can be found in detail in [27].

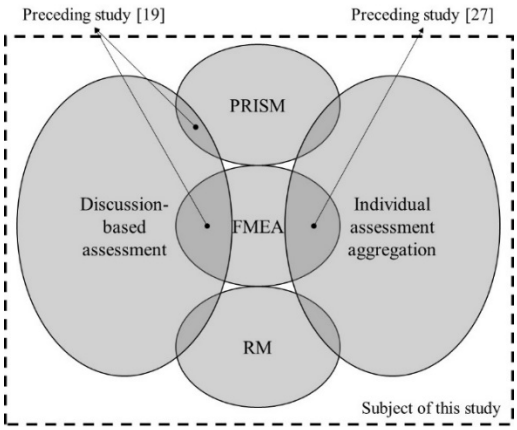


Figure 3
Boundaries of the preceding studies and this study

In step 6, the multi-criteria risk assessment of the compliance cases is performed by three risk assessment methods (RM, FMEA, PRISM) described in Section 2.1. Figure 4 shows the Risk Exposure-based, Figure 5 shows the RPN-based, while Figure 6 shows the PRISM-based assessments.

Different colors represent different Risk Exposure values, and based on Eq (1), the possible iso-risk lines in the Risk Matrix are linear (see Figure 4). The greener the matrix cell, the lower the Risk Exposure value, while the redder the matrix cell, the higher the Risk Exposure value. The iso-risk lines give the ranking of the assessments.

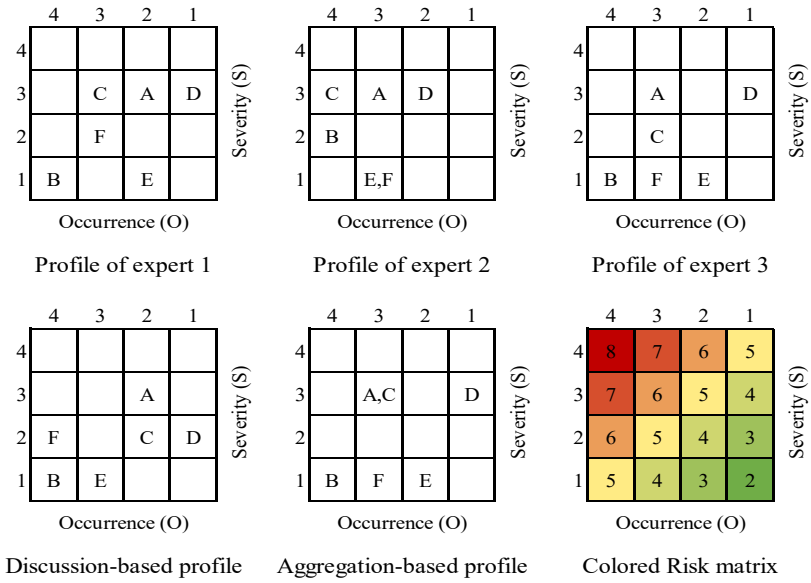


Figure 4
Risk Matrices with the assessed cases

The RPN values of the cases are visible in Figure 5. The greener the matrix cell, the lower the risk level, while the redder the matrix cell, the higher the risk level.

	Assessment of Expert 1	Assessment of Expert 2	Assessment of Expert 3	Discussion-based assessment	Aggregation-based assessment
A	24	36	36	24	32
B	8	16	12	12	12
C	36	36	24	16	33
D	12	24	9	8	15
E	2	6	4	6	4
F	6	6	6	32	7

Figure 5
The Risk Priority Number of the cases

The PRISM-based assessments are visible in Figure 6. Based on Eq (3), the possible iso-risk lines in the Partial Risk Map are linear. A case is consistently ranked by its highest PRISM number.

Table 2 shows the dense ranks of the cases by methods and experts, while Table 3 shows the dense ranks by methods and group assessments. Based on descending order, the higher the case rank, the higher the risk.

Based on Table 2 and Table 3, the paper's assumptions are examined, and the results of the analyses are presented and interpreted in Section 3.

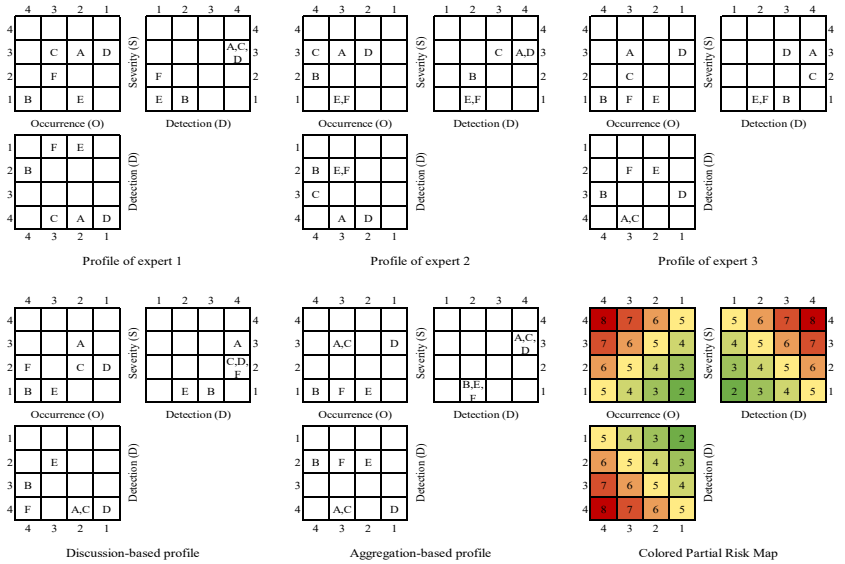


Figure 6
PRISM profiles with the assessed cases

Table 2
Individual ranking of the cases

	Risk Exposure			Risk Priority Number			PRISM number		
Case	Expert 1	Expert 2	Expert 3	Expert 1	Expert 2	Expert 3	Expert 1	Expert 2	Expert 3
A	2	2	1	2	1	1	1	1	1
B	2	2	2	4	3	3	2	2	1
C	1	1	2	1	1	2	1	1	1
D	3	3	3	3	2	4	1	1	2
E	4	4	4	6	4	6	4	3	4
F	2	4	3	5	4	5	3	3	3

Table 3
Discussion-based group assessment and aggregated ranking of the cases

	Discussion-based ranks			Aggregation-based ranks		
Case	Risk Exposure	Risk Priority Number	PRISM number	Risk Exposure	Risk Priority Number	PRISM number
A	2	2	2	1	2	1
B	2	4	2	2	4	2
C	3	3	3	1	1	1
D	4	5	3	3	3	1
E	3	6	4	4	6	4
F	1	1	1	3	5	3

3 Results

The first assumption's (A1) results are presented first. As step 7 of the risk assessment process, Table 4, Table 5, and Table 6 show the results of rank correlation analysis related to the three individual assessments and the two types of group assessments. Table 4 is related to the RM, Table 5 is related to the FMEA, and Table 6 is related to the PRISM method. The value of the Spearman's rho rank correlation coefficient is in the first matrix cell ("Sp. rho"), followed by the 2-tailed significance level ("Sig.").

Table 4

Correlations related to the RM-based comparisons

Group assessment	Individual assessment					
	Expert 1's rank		Expert 2's rank		Expert 3's rank	
	Sp. rho	Sig.	Sp. rho	Sig.	Sp. rho	Sig.
Discussion-based rank	0,375	0,464	-0,136	0,797	0,303	0,559
Aggregation-based rank	0,844	0,035	0,909	0,012	0,955	0,003

There are no significant correlations between individual ranks of the experts and discussion-based rank. On the other hand, between individual ranks of the experts and aggregation-based rank, there are only significant results with high-rank correlation coefficient values.

Table 5

Correlations related to the FMEA-based comparisons

Group assessment	Individual assessment					
	Expert 1's rank		Expert 2's rank		Expert 3's rank	
	Sp. rho	Sig.	Sp. rho	Sig.	Sp. rho	Sig.
Discussion-based rank	0,314	0,544	0,206	0,695	0,429	0,397
Aggregation-based rank	1,000	0,000	0,971	0,001	0,886	0,019

There are no significant correlations between individual ranks of the experts and discussion-based rank. Between individual ranks of the experts and aggregation-based rank, there are only significant results with high-rank correlation coefficient values.

Table 6

Correlations related to the PRISM-based comparisons

Group assessment	Individual assessment					
	Expert 1's rank		Expert 2's rank		Expert 3's rank	
	Sp. rho	Sig.	Sp. rho	Sig.	Sp. Rho	Sig.
Discussion-based rank	0,031	0,953	-0,127	0,810	0,281	0,589
Aggregation-based rank	1,000	0,000	0,984	0,000	0,742	0,091

There are no significant correlations between individual ranks of the experts and discussion-based rank. However, except for the case of Expert 3, the results are significant with high-rank correlation coefficient values between individual ranks of the experts and aggregation-based rank.

Secondly, the second assumption's (A2) results are presented. Kendall's rank concordance analysis is performed based on the expert rankings. The results of the analysis are summarized in Table 7.

Table 7
Kendall's rank concordance analysis results

	RM	FMEA	PRISM
Kendall's W	0,842	0,948	0,877
Significance level	0,027	0,014	0,022

Based on the results, the FMEA-based aggregation of the individual results has the highest Kendall's W value. However, in the case of all methods, the consensus level is significantly high.

4 Discussion and Managerial Implications

4.1 Discussion

According to ISO 37301:2021 [14], the systematic compliance risk assessment process significantly impacts business sustainability. Enterprise risks and risk management are critical factors in the 21st Century's complex business environment [29]. Based on [14, 29], the introduced compliance risk assessment process aims for a systematic and multi-methodological assessment of compliance risks. Analyzing the complex compliance risks requires a rigorous assessment process, including various indicators [17].

According to the case study at three banks operating in Central Eastern Europe, the qualitative approach for the research process is necessary [30]. Following the PDCA (Plan, Do, Check, Act) cycle, the evaluation phase of improvement projects is essential. Intense compliance supervision is fundamental in ensuring a sound banking system [31, 32]. Thus, in this follow-up case study, analyses are built on reliable data [27], and the assessment is simultaneously based on three different methods. However, the cost of regulatory compliance in the banking sector increases significantly [33]. The proposed multi-model risk assessment method produces sound results without sensibly increasing the evaluation cost because the assessment methods build on the same input.

Since the failure of bank boards and top experts in risk assessment can result in ineffective risk management [34], the proposed assessment process focuses specifically on the internal operations of the assessment board. Based on the proposed process, the results highlight some critical operation features that are useful to be discussed. The case study results highlight the importance of individual assessments, even in expert teams.

Since the ranking of the compliance risks is a multi-criteria decision-making process, the individual assessments can be helpful in the characterization and validation of the decision-making process. Individual risk assessment of complex phenomena leads to different rankings [27, 35]. Thus, differences between group assessments and individual assessments can be interpreted. Based on the results of this case study, there is a chance to have very different rankings if the assessment process is based on group discussion or individual assessment aggregation. Following the offered assessment process steps of this study can highlight the nature of the differences.

Significant differences were observed between discussion-based and aggregated rankings in this case study. The correlation between individual and aggregated rankings is higher than between individual and discussion-based rankings (see Assumption 1). The applied risk assessment methodology has no significant impact on the results of the assessments. Reflecting the necessity of sound bank systems [32] and effective risk management [34], focusing on the individual assessment level is essential. However, differences between rankings can be lower or higher in other banks, compliance fields, or other compliance cases. This study and its predecessor study [27] highlight the usefulness of individual assessments' operational analysis in the compliance assessment process.

Based on the results related to the consensus level of the individual assessments, no significant difference can be observed between the applied methods. The highest consensus level is observed in the case of FMEA, so the individual assessments were the most closely in the case of rankings by RPN (See Assumption 2). Instead of cases where only one method is applied [17], the research suggests applying more than one risk assessment method in the case of complex phenomena assessment. Thus, testing the validity of the individual assessment process step can be executed, and the inputs for the result evaluation will be more reliable.

A limitation of the study is that a limited number of experts considered the compliance cases. As for disseminating the multi-method risk assessment process, these numbers are enough, but the research can be performed again with an increased number of experts in the future. Furthermore, since risk factors are estimated based on previous observations, some uncertainty is associated with these estimations. Considering this uncertainty is a possible direction of extension to the proposed assessment process, like the measurement uncertainty consideration in forecasts and risk-based decisions [36, 37].

This study stops at the statistical analysis of the results (step 7). First, according to the change management process model in [38], it is essential to evaluate the assessment results, as shown in Step 8 and Step 9 in Figure 2. Step 8 refers to the discussion of the results by the expert group or committee, which should include identifying lessons learned in the evaluation process. Additionally, some knowledge or skill gaps may be defined at this point. Next, step 9 is about planning risk reduction action. Following the PRISM and FMEA methods, risk reduction can aim at either decreasing the severity of consequences (i.e., insurance or technology improvements), decreasing the probability of occurrence (i.e., with statistical control activities), or increasing the detectability of the failure, by new controls or alarm points or training. In the banking sector, following the traditional audit risk model, measures may aim at the incentive system or internal control system, like board structure or reporting. Finally, the risk of attitude or rationalization may be affected by training and education. [39, 40]

4.2 Managerial Implications

The senior management is responsible for determining the company's compliance policy, including the commitment to produce and sell only compliant products and to comply with applicable regulations and industry standards. In addition, top management seeks tangible evidence of returns on efforts to continue compliance improvement programs. Thus, based on the research results, operations research can help transform compliance management into a more reliable process.

A risk-based approach is the general best practice for compliance management, as described in Section 1. The three risk assessment techniques described in Section 2 contribute to a better understanding of techniques and a choice among them for better decision-making. The authors suggest supplementing the risk analysis of ISO 37301:2021 with the "undetection" dimension, as used in FMEA and PRISM methods [19].

The individual assessment of risks should be supplemented to the group-level assessment in the risk assessment process. At the end of the assessment process, more detailed risk mitigation actions can be planned. From a management perspective, following the designed compliance risk assessment process (See Figure 2) helps leaders and responsible risk owners to address compliance risks. The proposed process can be used as a comprehensive tool to select the necessary compliance measures optimally. The authors consider compliance a strong business driver, in line with quality and reliability management, that can lead to business sustainability [41].

Conclusions

This study provides insight into compliance risk assessment methodologies. Previous studies have focused on the discussion-based PRISM method [19] and the aggregation-based FMEA [27]. This follow-up case study complements the above

with statistical analysis to investigate the consensus level of the participants while the assessment is simultaneously based on three different methods.

First, a brief introduction to compliance management describes the organizational function responsible for fulfilling legal, regulatory, industrial, and other obligations. Then, Section 1 describes the three distinguished steps of the compliance risk assessment process.

The steps of the empirical research and the materials used are presented in Section 2. The primary data collection was at one of the largest commercial banks in Central and Eastern Europe in 2021. In a previous study [27], the authors applied the FMEA in compliance risk assessment in the financial sector for the first time. The results are shown in Section 3. The authors validated the interpretations from the case study through consensus analysis. In this paper, two assumptions have been examined.

Regarding the first assumption, the results highlight no significant correlations between individual rankings of the experts and discussion-based ranking. However, the results are significant and have high-rank correlation coefficient values between individual and aggregation-based rankings (except for the case of Expert 3 in the PRISM-based comparisons). This study highlights the usefulness of individual assessments of experts in the compliance risk assessment process.

As for the second assumption, the results show that the FMEA-based aggregation of the individual results has the highest consensus level. However, in the case of all methods, the consensus level is significantly high. Therefore, this research suggests applying more than one risk assessment method in the case of complex phenomena assessment, just like bank compliance issues. Thus, testing the validity of the individual assessment process step can be executed, and the inputs for the risk evaluation will be more reliable.

The results are discussed in Section 4. The results related to the discussion-based PRISM method are described in detail in [19], while the aggregation-based FMEA results can be found in detail in [27]. One key finding of this study is that individual assessments can help characterize and validate the decision-making process. No significant difference can be observed between the applied methods, while testing the validity of each step of the evaluation process is possible.

Future research will focus on further developments of the PRISM method. One possible developmental direction is to combine the PRISM methodology with MCDM methods to understand better the complexity of the risk assessment characteristics of the banking sector. The research related to PRISM integration with pairwise comparison and other MCDM techniques has already started as a possible methodological development direction. Currently, the first integration experiences related to AHP (Analytic Hierarchy Process), BWM (Best Worst Method), and TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) are being synthesized. Furthermore, since applying fuzzy sets provides excellent possibilities in the bank risk description under uncertainty [42-44], fuzzification is the upcoming developmental direction of the PRISM method.

The authors also plan to repeat the research with an increased number of experts and incidents to characterize more deeply the possible differences in the applied methods. Furthermore, future research will examine how the proposed compliance assessment process can be integrated with business risk assessments based on the PRISM approach. Finally, investigating the relationship between organizational resilience and compliance management is a topic with great potential.

References

[1] Ramos-González, M. D. M.; Rubio-Andrés, M.; Sastre-Castillo, M. Á.: Building Corporate Reputation through Sustainable Entrepreneurship: The Mediating Effect of Ethical Behavior, Sustainability 2017, 9, 1663, https://doi.org/10.3390/su9091663

[2] Eizenberg, E.; Jabareen, Y.: Social Sustainability: A New Conceptual Framework, Sustainability 2017, 9, 68, https://doi.org/10.3390/su9010068

[3] Wells P.: Sustainable business models and the automotive industry: A commentary, Centre for Automotive Industry Research, Cardiff Business School, Cardiff, UK, 2013

[4] Yamane, T.; Kaneko, S.: The sustainable development goals as new business norms: A survey experiment on stakeholder preferences, Ecological Economics 2022, 107236, https://doi.org/10.1016/j.ecolecon.2021.107236

[5] Surman, V.; Böcskei, E.: Examining Sustainability Attitudes in the case of Small and Medium-sized Enterprises, In Proceedings of FEB Zagreb 2022, 13th International Odyssey Conference on Economics and Business, Dubrovnik, Croatia, 1-4 June, 2022

[6] EU Regulation 2020/852. Available online: https://eur-lex.europa.eu/legal-content/HU/TXT/PDF/?uri=CELEX:32020R0852&from=EN (accessed on 30 April 2022)

[7] EU Regulation 2021/2178. Available online: https://eur-lex.europa.eu/legal-content/HU/TXT/PDF/?uri=CELEX:32021R2178&from=EN (accessed on 30 April 2022)

[8] Wong, C. M. L.; Jensen, O.: The paradox of trust: perceived risk and public compliance during the COVID-19 pandemic in Singapore, Journal of Risk Research 2020, 23, pp. 1021-1030, https://doi.org/10.1080/13669877.2020.1756386

[9] Stupak, I., Mansoor, M.; Smith, C. T.: Conceptual framework for increasing legitimacy and trust of sustainability governance, Energy, Sustainability and Society 2021, 11, https://doi.org/10.1186/s13705-021-00280-x

[10] Harkácsi, G. J.; Szegfű, L. P.: The Role of the Compliance Function in the Financial Sector in the Age of Digitalisation, Artificial Intelligence and Robotisation, Financial and Economic Review 2021, 20, pp. 152-170

[11] Poppensieker, T.; Schneider, S.; Thun, M.: Financial institutions and non-financial risk: Learning from the corporate approach, in: Non-Financial Risk Management: Emerging stronger after Covid-19, 1st Ed.; Kaiser, T., Risk Books: London, United Kingdom, 2021

[12] Hendra, R.: Comparative Review of the Latest Concept in Compliance Management & The Compliance Management Maturity Models, RSF Conference Series: Business, Management and Social Sciences 2021, 1, 116-124, https://doi.org/10.31098/bmss.v1i5.457

[13] Holter Antonsen, H.; Madsen, D. Ø.: Developing a Maturity Model for the Compliance Function of Investment Firms: A Preliminary Case Study from Norway, Adm. Sci. 2021, 11, 109. https://doi.org/10.3390/admsci11040109

[14] ISO: Compliance management systems - Requirements with guidance for use. ISO 37301:2021

[15] IEC: Risk management — Risk assessment techniques 31010:2019, 2019

[16] Basel Committee on Banking Supervision: Principles for Effective Risk Data Aggeration and Risk Reporting, 2013, http://www.bis.org/publ/bcbs222.pdf 22/08/2021

[17] Losiewicz-Dniestrzanska, E.: Monitoring of compliance risk in the bank, Procedia Economics and Finance 2015, 26: 800-805, https://doi.org/10.1016/S2212-5671(15)00846-134

[18] Mirghafori, S. H.; Takalo, S. K.; Dastranj, M.: Banking service quality management using fuzzy FMEA (a case study: Central Melli Bank of Rafsanjan), International Journal of Quality and Innovation 2016, 3, 1-16, https://doi.org/10.1504/IJQI.2016.079907

[19] Bognár, F.; Benedek, P.: A Novel Risk Assessment Methodology – A Case Study of the PRISM Methodology in a Compliance Management Sensitive Sector, Acta Polytechnica Hungarica 2021, 18, 89-108, https://doi.org/10.12700/APH.18.7.2021.7.5

[20] Qazi, A., Shamayleh, A., El-Sayegh, S., Formaneck, S.: Prioritizing risks in sustainable construction projects using a risk matrix-based Monte Carlo Simulation approach, Sustainable Cities and Society, 65, (2021) 102576, https://doi.org/10.1016/j.scs.2020.102576

[21] Wang, R.; Wang, J.: Risk Analysis of Out-drum Mixing Cement Solidification by HAZOP and Risk Matrix, Annals of Nuclear Energy, 2020, 147, 107679, https://doi.org/10.1016/j.anucene.2020.107679

[22] Liu, H. C.; Liu, L.; Liu, N.: Risk evaluation approaches in failure mode and effects analysis: A literature review, Expert Systems with Applications 2013, 40, 828-838, https://doi.org/10.1016/j.eswa.2012.08.010

[23] Bognár, F.; Hegedűs, Cs.: Description and Consequences on some Aggregation functions of PRISM (Partial Risk Map) Risk Assessment Method, Mathematics, 2022, 10, 676, https://doi.org/10.3390/math10050676

[24] Braglia, M. MAFMA: Multi-attribute failure mode analysis, International Journal of Quality and Reliability Management 2000, 17, 1017-1033, https://doi.org/10.1108/02656710010353885

[25] Spearman, C.: The Proof and Measurement of Association between Two Things, The American Journal of Psychology, 1904, 15, 72-101, https://doi.org/10.2307/1412159

[26] Kendall, M. G.: Rank Correlation Methods, Griffin: London, UK, 1970

[27] Bognár, F.; Benedek, P.: Case Study on a Potential Application of Failure Mode and Effects Analysis in Assessing Compliance Risks, Risks 2021, 9, 164, https://doi.org/10.3390/risks9090164

[28] Berényi, L.; Deutsch, N.: Corporate Social Responsibility and Business Philosophies among Hungarian Business Students, Sustainability 2021, 13, 9914, https://doi.org/10.3390/su13179914

[29] Strelicz, A.: Risks and threats in cyberspace – The key to success in digitalization, Journal of Physics Conference Series 2021, 1935, 012009, https://doi.org/10.1088/1742-6596/1935/1/012009

[30] Bauer, S.; Bernroider, E. W. N.; Chudzikowski, K.: Prevention is better than cure! Designing information security awareness programs to overcome users' noncompliance with information security policies in banks, Computers & Security 2017, 68, 145-159, https://doi.org/10.1016/j.cose.2017.04.009

[31] Demirgüç-Kunt, A.; Detragiache, A.; Tressel, T.: Banking on the principles: Compliance with Basel Core Principles and bank soundness, Journal of Financial Intermediation 2008, 17, 511-542, https://doi.org/10.1016/j.jfi.2007.10.003

[32] Demirgüç-Kunt, A.; Detragiache, A.: Basel Core Principles and bank soundness: Does compliance matter?, Journal of Financial Stability 2011, 7, 179-190, https://doi.org/10.1016/j.jfs.2010.03.003

[33] Kaminski, P.; Mikkelsen, D.; Poppensieker, T.; Robu, K.: Sustainable compliance: Seven steps toward effectiveness and efficiency, McKinsey & Company 2017, 1-9

[34] Agarwal, S.; Kamath, S.; Subramanian, K., Tantri, P.: Board conduct in banks, Journal of Banking & Finance, 2022, 106441, https://doi.org/10.1016/j.jbankfin.2022.106441

[35] Qin, J.; Xu, Y.; Pedrycz, W.: Failure mode and effects analysis (FMEA) for risk assessment based on interval type-2 fuzzy evidential reasoning method, Applied Soft Computing, 2020, 89, 106134, https://doi.org/10.1016/j.asoc.2020.106134

[36] Hegedűs, Cs.; Kosztyán Zs. T.: The consideration of measurement uncertainty in forecast and maintenance related decisions, Problems of Management in the 21st Century 2011, 1, pp. 46-69, https://doi.org/10.33225/pmc/11.01.46

[37] Kosztyán, Zs. T.; Hegedűs, Cs.: Computer-Aided Diagnostic Methods to Forecast Condition-Based Maintenance Tasks, Lecture Notes in Electrical Engineering 2013, 151, 367-380, https://doi.org/10.1007%2F978-1-4614-3558-7_30

[38] Horváth, Cs.; Mecsei, I.: Sustainability Inspections in Printing Industry, International Joint Conference on Environmental and Light Industry Technologies. 19-20 November 2015 Budapest, Hungary Obuda University pp. 19-24

[39] Cressey, D. R.: Other People's Money, 1973, Patterson Smith, Montclair

[40] Fortvingler, J.; Szívós, L.: Different Approaches to Fraud Risk Assessment and Their Implications on Audit Planning, Periodica Polytechnica Social and Management Sciences, 2016, 24(2), pp. 102-112, https://doi.org/10.3311/PPso.8436

[41] Aigner, D. J.; Hopkins, J.; Johansson, R.: Beyond Compliance: Sustainable Business Practices and the Bottom Line, American Journal of Agricultural Economics 2003, 85, pp. 1126-1139, https://www.jstor.org/stable/1244884

[42] Cao, C.; Zhang M.: Credit Risk Evaluation of Quantum Communications Listed Companies in China Based on Fermatean Fuzzy TOPSIS, Procedia Computer Science, 2022, 199, pp. 361-368, https://doi.org/10.1016/j.procs.2022.01.044

[43] Yuanyuan Zhou, Y.; Zheng, c.; Goh, M.: Statistics-based approach for large-scale group decision-making under incomplete Pythagorean fuzzy information with risk attitude, Knowledge-Based Systems, 2022, 235, 107654, https://doi.org/10.1016/j.knosys.2021.107654

[44] Bonet, I.; Peña, A.; Lochmuller, C.; Patiño, H. A.; Chiclana, F.; Góngora, M.: Applying fuzzy scenarios for the measurement of operational risk, Applied Soft Computing, 2021, 112, 107785, https://doi.org/10.1016/j.asoc.2021.107785

Dynamic Modelling and Experimental Analysis of Tankless Solar Heat Process System for Preheating Water in the Food Industry

Rajab Ghabour[1], Péter Korzenszky[2*]

[1]Mechanical engineering doctoral school, Hungarian University of Agriculture and Life Sciences, Pater Karoly u 1, 2100 Gödöllő, Hungary; Ghabour.Rajab@phd.uni-mate.hu

[2*]Institute of Technology, Hungarian University of Agriculture and Life Sciences, Pater Karoly u 1, 2100 Gödöllő, Hungary

Corresponding: Péter Korzenszky (Korzenszky.Peter.Emod@uni-mate.hu)

Abstract: Worldwide, industrial process heat accounts for over two-thirds of global industrial energy consumption. Wherein solar thermal systems are expected to be a reliable alternative solution. The temperature output of the flat-plate collector (FPC) is enough to serve specific industrial applications where low temperature is needed for heat processes, as in the food and beverage sector where nearly all the demand is needed heating and more than 40% is below 100°C. This work presents a tankless solar heat for industrial process (SHIP) system using FPC to produce low temperature for process heat of a food factory in the climate of central Europe. The modelling system is developed based on heat transfer and thermodynamics phenomena, and it is resolved mathematically and validated experimentally onsite. Moreover, this study considers the Budapest region (47.50° N, 19.04° E- Hungary) as a case study. The metrological data is extracted by MeteoSyn software. The objective function is to meet the set temperature requirements for preheating water for food factories to minimize the annual energy costs by saving more fuel. The heat exchanger secondary loop's six different temperature configurations were used as 60, 55, 50, 45, 40 and 35°C, respectively, and each measurement was tested for several summertime days. Results show that tankless SHIP can help preheat water for temperatures below 100°C. By comparing the different configurations, they are acierating that all set temperature combinations can be achieved at a lower mass flow rate of 60 litres per hour for the primary and secondary loops. The obtained outcomes of this work show that integrating solar thermal systems into industrial processes for low-to-medium temperatures is a very economical and promising solution to replace fossil fuels.

Keywords: tankless system; solar heat; flowrate; food processing

1 Introduction

Solar radiation can be utilized directly by photosynthesis or by converting the radiation into electricity. While indirectly heating a particular medium using low-to-medium temperature solar collectors for heating applications such as solar cooking, solar dryer, or solar water heating systems. Also, solar thermal energy for low-temperature applications such as domestic or industrial water heating is the most common around the globe. There are significant differences between the SHIP and domestic hot water (DHW) systems. Firstly, using the solar tank is a must for the DHW, while it is not an obligation in the SHIP systems. Secondly, the required temperature in the case of DHW is permanently fixed at around 60°C, while for SHIP, it varies between 40°C to 600°C. Finally, the DHW system can have a simple configuration and integration, while the SHIP has more sophisticated and complex to integrate into the heat process due to the sensitivity of each industrial case study. Also, integrating SHIP into the industrial process is times harder and more complicated than a simple DHW system, but since industrial demand accounts for a significant portion of the national and global demand, it will result in significant savings in fuel and natural gas. There are different types of solar collectors for water heating applications flat-plate collector (FPC), evacuated tube collector (ETC), and compound parabolic collector (CPC). Several researchers have designed and analyzed FPC using exergy analysis to obtain the optimal thermal energy conversion [1], [2]. At the same time, other researchers have studied the thermal performance of the energy storage tank in the solar heating system [3]. As a result, It is reported that using solar energy in the industrial sector can enormously reduce CO_2 equivalent emissions by 65-75% [4].

Globally, energy is essential in industrial processes and used for economic growth and continuous development. To assess the sustainability of any energy source, the availability, security, and environmental effect are critical parameters. Therefore, scientists and researchers focus on these concerns to achieve a sustainable and secure energy future. In the most recent literature, global energy consumption is expected to increase by 33% from 2010 to 2030, and the industrial sector will take a big part in this upsurge [5]. The average energy consumption in the industrial sector is higher than any other sector, estimated at 35% globally [6] and 30% in the southern Mediterranean European countries. 78% of this energy is needed for heat processes and 22% for electricity [7]. Nevertheless, it varies by economic activities, regional location, and modernization. Solar thermal energy would significantly contribute to industrial energy systems by tapping into this worldwide potential.

Potential studies were carried out for different regions and countries to use SHIP systems. A comprehensive study was performed for the USA, Spain, Portugal, Austria, Australia, Italy, Netherlands, Sweden, Cyprus, Tunisia, and Germany to analyze the industrial heat demand below 300°C [8]-[11]. At first, suitable sectors were selected to determine a theoretical potential. Therefore, high waste heat or

low heat demands were excluded. As a result, sectors of chemicals, food, beverages, and textiles are suitable for using solar heat. The low-to-medium temperature heat demand in those sectors is suitable for drying, cleaning, washing, and surface treatment. Then, the theoretical potential was divided into a short-therm (<100°C) and mid-therm (<250°C) because the solar collector's technology for heat process above 100°C is not readily available in the market, so it did not reach a mature level by the time of the study. As a result, the food and beverages sector is considered suitable for using solar heat in all potential studies. Within food and beverage factories, a significant share of the heat demand accounts for 58% and 42% for low and medium temperatures. It is noteworthy that more than 50% of the installed solar heating systems were used for washing or water heating processes. In 2006, more than 40% of it was used for processes in 60-100 °C, which is a very suitable range for commercial solar collectors [12].

Using a boiler as a process heating device to generate hot water or steam in the industrial sector is expected, which consumes significant energy. Fossil fuel is used in all major industries to produce hot water or steam for industrial processes. This consumption results in greenhouse gas emissions that cause climate change, where CO_2 emission alone involves 76% of global warming [13]. Since fossil fuels play a significant role in meeting the required energy for different heat processes, renewable energy sources for process heat generation are more suitable due to their sustainability. For example, solar thermal collectors directly convert the solar energy into heat compared with 3-4 times higher energy efficiency than the photovoltaic powered electrical heater. Also, storing electrical energy using batteries is very expensive and still under development, making the PV system less favourable in the heating energy market [14].

Nevertheless, the current status of implementing solar heat for SHIP is still at an incentive stage of commercialization due to several barriers: relatively high upfront costs, low awareness of policymakers, and the need to develop more suitable solar technologies for medium and high-temperature processes [15]. According to the most recent published reports, solar energy will afford 45% of the global energy requirements by 2050. Many strategies and projects have been launched to fulfill this target.

Milk, fruits and vegetable production have a considerable heat demand, and it is a promising sector for integration with solar heat systems. In contrast, the sugar industry has the highest energy demand, where production mainly occurs in winter and autumn. Furthermore, some subsectors that appear suitable for solar heat are mineral water, feed and malt, where standard processes such as pasteurizing liquid goods at 65-100 °C and cleaning production facilities and products in all subsectors at 60-90 °C.

In many industries, integrating solar systems does not require a storage tank. Because either the heat provided by the solar system is higher than the heat needed, or the solar system will be integrated into the processing system as preheating water before entering the conventional boiler or steam generator.

In other words, this system can be used when the heat process requires a continuous operation and a load, such as in food processes, where the load is always higher than the supplied heat from the solar system. In this case, the integration of solar thermal systems remains low to avoid high costs related to the storage system. In such a configuration, the hot water supplied from the solar heat system will be fed directly to the existing heat supply or the heat process, as shown in Figure 1.

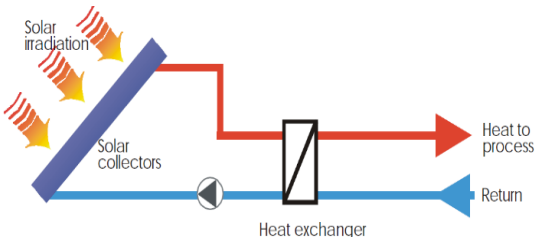


Figure 1
Tankless solar heat for industrial process system

Although many experimental and simulation studies on FPC and SHIP systems have been reported in the literature, detailed simulation for tankless solar heat systems in central Europe with real-time experiments is yet to be carried out. Many types of research were focused on SHIP systems in central Europe to highlight the massive energetic potential in the literature. In central Europe, acceptable solar radiation levels are available, which can meet the low to medium temperature requirements in several industrial sectors. Preheating water in those industries can account for significant fossil fuel and natural gas savings annually. In the V4 group, Hungary has the highest solar energy potential compared to the other bloc countries.

Nevertheless, no actual experiment was conducted to prove those hypotheses regarding SHIP. Also, no dynamic system simulation with T*sol is available to perform parametric analysis of different parameters of a tankless solar system for the industrial heat process considering parameters like water inlet temperature, mass flow rate, ambient temperature, outlet temperature, solar thermal heat gain, and so forth. The present work focuses on T*sol software to analyze a tankless solar heat for an industrial process system used to preheat water for the food industry in Budapest, Hungary, at different output temperatures. This research includes a comprehensive statistical study of the temperature profiles to meet the needed set-temperature values. Then, the simulation and the actual results were evaluated regarding the temperature and weather parameters. The well-engineered tankless SHIP can supply industrial heat processes per the required outlet temperature range.

2 Method

2.1 Hungarian Energy Status

All over the Europe Union (EU), solar water heater systems are given financial incentives to encourage the usage of renewable energy, except for Finland, Greece, and Denmark. While in Hungary, which is part of the "Visegrád group - V4 group", currently solar capacity has increased tenfold in the past three years (2017-2020) due to the favourable investment environment and the lower cost of solar technology [16]. Hungary also has ten solar power plants with more than 10 MWp, and five solar power plants under 10 MWp spread all over the country [17]. Moreover, 40% of the total consumed energy goes for heating and cooling purposes [18]. Since energy security is a worldwide concern, all nations have framed energy policies in contrast with sustainable growth.

Similarly, the EU framed two targets (by 2030 and 2050) to reduce greenhouse emissions by at least 55% compared to 1990 [19]. Reducing carbon emissions is a significant challenge in all EU and especially V4 countries due to the dependency on conventional energy sources. The EU aims to increase the renewable energy share to 32% by 2030, while Hungary plans for 21% by 2030, and the current status reports that the renewable energy share will be 12% in 2019 [20]. According to the available records, solar, biomass, biogas, hydro, and wind are the most significant resources in Hungary, as in Figure 2. Biomass was the primary renewable energy source for energy production, and it was reduced drastically in less than ten years from 67% to 48%.

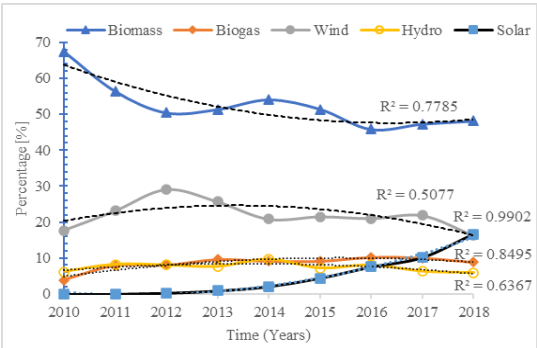


Figure 2
Regression line analysis of various renewable energy sources in Hungary using poly trendline (2010-2018)

On the other hand, bio-based sources have increased slightly, and wind and hydropower maintained their share. In contrast, solar energy has the most extensive share growth in the last decade, from almost 0% to 16% (coefficient of

determination R^2=0.9902). If the trendline and conditions remain the same, it can reach more than 30%. Consider that the new technological advancements can positively affect the long-term aims.

Several solar parks with small to high-power capacities were installed in the last decade throughout Hungary. According to the Hungarian solar Association, the growth parameters have doubled in the last few years. Comparing it with the neighbourhood countries is essential to unfold the potential capacity. For example, Germany has the highest number of solar panels, though Hungary has 50% more solar radiation than Germany throughout the year. Moreover, in Vienna, Austria, the annual sunshine hours are around 1850 hours, while in Budapest, Hungary, there are 2000 hours.

Nevertheless, Austria is the world-leading country in solar water heating systems regarding the thermal energy produced (kWh) per one thousand capita [21]. Based on the growth data, the regression analysis was depicted in Excel. The two lines represent the solar growth potential in the Hungarian region. The coefficient of determining the solar energy growth prediction has a linear value. While for determining more accurate results, two clusters were studied with more significant and lesser than 10 MWp solar park capacities. The graph confirms that the expected growth of solar parks is more significant than 10 MWp if further developments are achieved in this technology.

The regional map analysis is helpful for visualizing the region's current falling and irradiance scenario. The recorded global irradiance data is a powerful tool to predict solar potential [22] accurately. According to the regional maps, Hungary receives the highest region V4 in annual irradiation of 1200-1400 kWhm^{-2}. So, the annual energy generation prediction was the highest, with a range of 900-1050 kWhm^{-2}. While the optimally inclined collectors can harvest a sum of irradiation in the range of 1300-1600 kWhm^{-2}, and the energy production potential is 975-1200 kWhm$^{-2,}$ as in Figure 3.

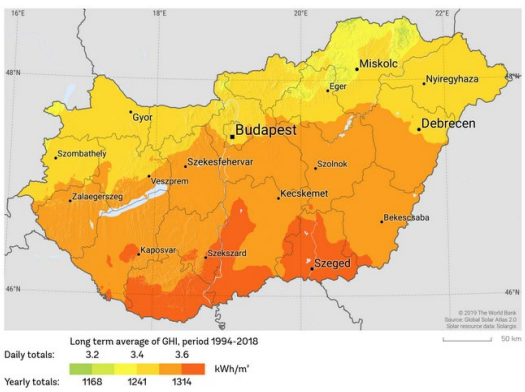


Figure 3

Global solar radiation in Hungary on a horizontal surface

2.2 Model Development and Simulation

A dynamic analysis tool is needed to accurately describe the system's response to the rapid environmental change in weather conditions to study the SHIP system's performance [23], [24]. T*sol software is a professional simulation program for designing solar thermal systems, including domestic hot water (DHW), swimming pool heating, district heating, and heat process systems. It calculates and simulates the thermal process in these systems by providing the components and tools of the solar systems and all relevant data [25]. This software enables the engineers to design the system optimally at a low cost and time. The calculations are based on the energy balance flows and provide predictions according to the hourly metrological data [26].

This section develops the comprehensive models of the tankless SHIP system by collecting the meteorological data for an entire year and comparing it with the average data records between 1995-and 2012 with a deeper focus on the summertime results.

2.2.1 SHIP Concept

The SHIP comprises a solar collector integrated with an external heat exchanger in a closed loop. In this external loop, a water and propylene glycol mixture runs in the piping system to convey the absorbed energy from the solar collector to the heat process using a heat exchanger. The concept assumes that the flow goes out of the solar collector at a specific temperature (denoted by T Pr In) which is the same input temperature of the primary side of the heat exchanger since the piping system is well-insulated. Moreover, the output temperature from the heat exchanger in the primary loop (denoted by T Pr Out) is the same temperature that flows in the collector for the same reason. On the secondary side of the heat exchanger, there is an open loop that warms up the network water (denoted by T Sec In) by exchanging the heat with the fluid on the primary side so it can be ready for the heat process at the required temperature (denoted by T Sec Out).

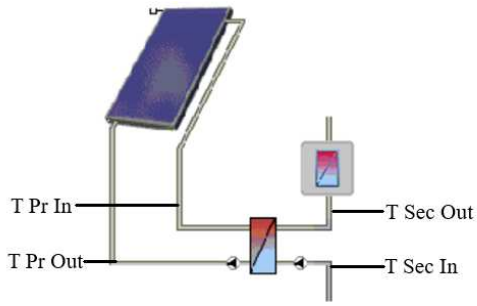


Figure 4
A tankless solar system with heat exchanger and FPC collector schematic

The schematic concept of the system is shown in Figure 4. The concept assumes no possibility of mixing between the fluid on the primary and the secondary side to avoid mixing the antifreeze liquid with the process heat, which can be risky for human health, mainly in food processing like pasteurizing.

2.2.2 Modelling in T*sol

The simulation is carried out for six different configurations of a food process industry to assess and optimize the working parameters. The following assumptions and inputs are considered:

- The weather data is used for the geographical location of Budapest (KMI weather station), Hungary.
- The solar fraction can reach up to cover the whole required energy during certain weather conditions. However, central Europe's annual feasible solar fraction is 40-60%. Therefore, an auxiliary heating system is considered in natural systems to cover the rest of the demand and ensure a continuous supply of the process heat.
- The efficiency of the solar thermal collector is dependent on the ambient temperature. Therefore, an average ambient temperature of 25 °C is the relevant temperature for the modelling analysis.
- The examined case study of the food process requires several temperatures for processing and cleaning.
- The temperature of the network water is taken from the data records

3 Experiment

The components used in the T*sol model configuration are (1) single FPC. (2) external heat exchanger transfers the absorbed energy from the primary loop to the secondary one. (3) heat exchanger. The water flows in the pipes using an active circulation pump in a closed loop in the primary loop. The flow rate of the primary loop is 60 litres per hour. The fluid is a mixture of glycol and water with an 18% volumetric percentage to avoid freezing and bursting during low-temperature. The intended experimental setup consists of several primary data logging and instrumentation components: (1) IMRe, and (2) data logger (ALMEMO® 2890-9) in a universal input data logger that connects nine sensors through multiple channels, (3) pyranometer, (4) mechanical flow meter, (5) temperature sensor (k-type). (6) Open Energy Monitor is monitoring, modelling, and assessment tool to monitor live data and records. (7) TECH controller to adjust the working schedule of the process heat. The measurement ranges and the accuracy of each instrument are shown in Table 1.

The experiment set up has been installed in the Hungarian University of agriculture and life sciences renewable energy laboratory, Gödöllő, Hungary. Experiments were carried out 24 hours a day during August and September 2021, when the weather was relatively stable, with rain interruption for a few days during the whole period. All the data have been recorded persistently by the data logger every minute.

The actual experiment (mainly the solar collector) was mounted on a metallic structure and directed at an angle of 33º facing the south to harvest the maximum radiations throughout the day. Knowing that this angle is the optimum angle for Budapest during the summer climate was recommended by the solar electricity handbook 2019.

Table 1
Measurement tools and measuring range

Instrument	Type	Measuring range	Accuracy
Data logger	ALMEMO® 2890-9	over 70 measuring ranges	-
Pyranometer	THEODOR FRIEDRICHS 6003.000 BG	0 - 1300 [Wm^{-2}]	< 10%
Mechanical flow meter	Zenner DE-07-MI001-PTB010 (R80HR40V) with impulse output	0.05 [l] to 10,000 [m^3]	Cold water ± 5 % hot water ± 3 %
Thermocouple (k-Type)	Programmable Resolution 1-Wire Digital ThermometerDS18B20	-55 to +125 [°C]	0.5 [±°C]
IoT device	IMRe	Relative to each sensor	

As in Figure 5, the IMRe system is an internet of things (IoT) device that collects the energy usage data of a specific case and uploads the sensor data directly to the internet [27]. A single user can collect data using one or more devices, which provides higher scalability and flexibility if required. The collected data can be viewed and evaluated using an internet browser. The monitoring system has two main parts, the local group, which contains IMRe hardware and sensors. The second group is the cloud, where a server runs Emoncms framework processes for all raw signals and then converts them into their corresponding physical values [28]. It should be mentioned that it is possible to build and generate additional calculated values in the server. All these data are stored in the cloud database and can be easily visualized to the end-user.

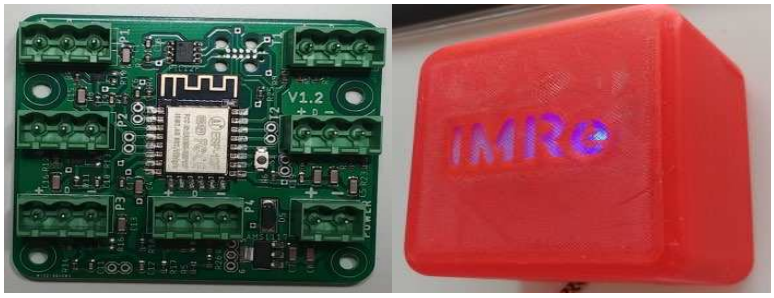


Figure 5
IMRe controlling system

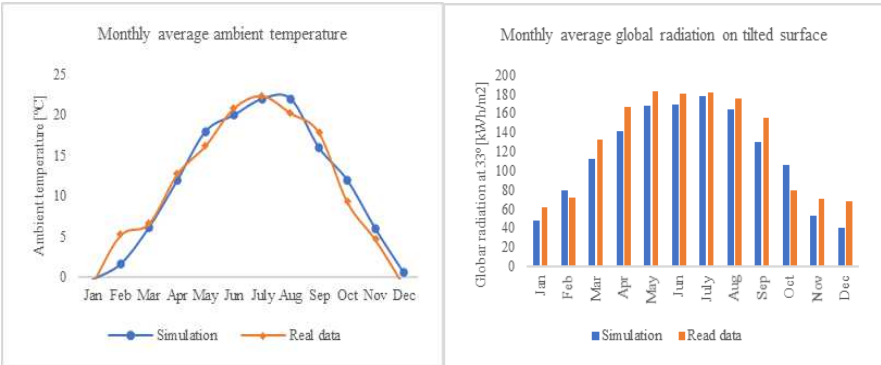
4 Results

The outcomes of this research, along with the corresponding physical explanations, are described in this paragraph. Data were collected for the continual flow of hot water at different process heat temperatures 35-60°C for preheating water and 60/60 litres per hour configurations of flow rates in the primary and secondary loop and actual solar radiation from the sun.

4.1 Hungarian Weather Profile

The Hungarian climate is continental, with cold, snowy and frigid winters and low humid hot summers. The average annual temperature is 9.70°C based on long-term statistics and 11.25°C based on the software outputs. In comparison, temperature extremes are 42°C in summer and -29°C in winter. The minimum and maximum variation of the annual ambient temperatures obtained from the T*sol simulation compared to the actual data shown in Figure 6 (a). In the same plot, the solar radiation's annual variation is also portrayed to show the interactive pattern between the two parameters, as in Figure 6 (b). It is evident from the graph that the ambient temperature varies with the solar radiation, and the ambient temperature profile varies between -1 and 23 °C in average monthly rates.

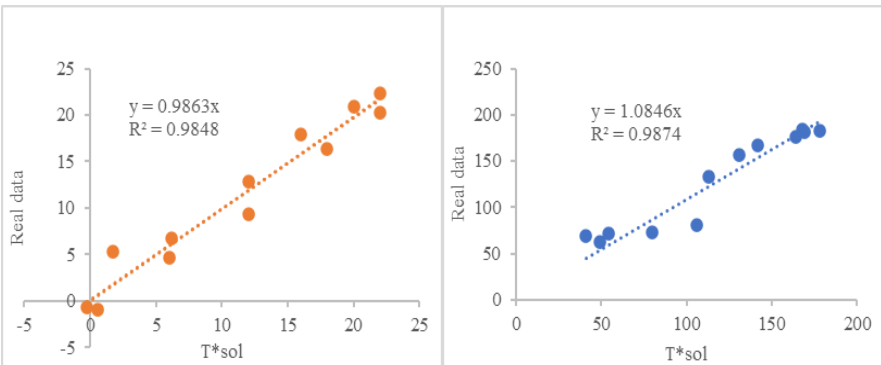
The comparison is conducted between the T*sol weather data file, TMY, an average recorder between 1995 and 2012, and the actual experiment data at the site location. As a result, the average differences are more than one hundred $kWhm^{-2,}$ and the error is 10% between the two yields, as in Figure 6 (b). While for the ambient air temperature profiles, the average difference is less than 0.1°C, approximately 0.8% error.



(a) *(b)*

Figure 6
Monthly average a) ambient temperature b) solar radiation

A comparison analysis was conducted to understand this difference. It was found that there is a linear correlation using (0,0) set intercept between the two ambient temperature profiles and the determination coefficient R^2= 0.9848. Similarly, for the global radiation profiles, it was found that the two profiles have a linear correlation with a determination coefficient R^2= 0.9874, as in Figure 7.



(a) *(b)*

Figure 7
Coefficients of determination for linear weather data a) ambient temperature b) solar radiation

4.2 Model Validation

The authenticity of the dynamic simulation model using T*sol software of the tankless SHIP has been examined experimentally by real-time measurement for three months at 60/60 litres per hour for both loops (primary and secondary).

Figure 8 shows the black dashed line, representing the heat process set temperature. It is pretty apparent from the graph that the orange line (representing the secondary output temperature) is always nearby the set temperature. Hence, the simulation model is appropriate to carry on further to predict system performance.

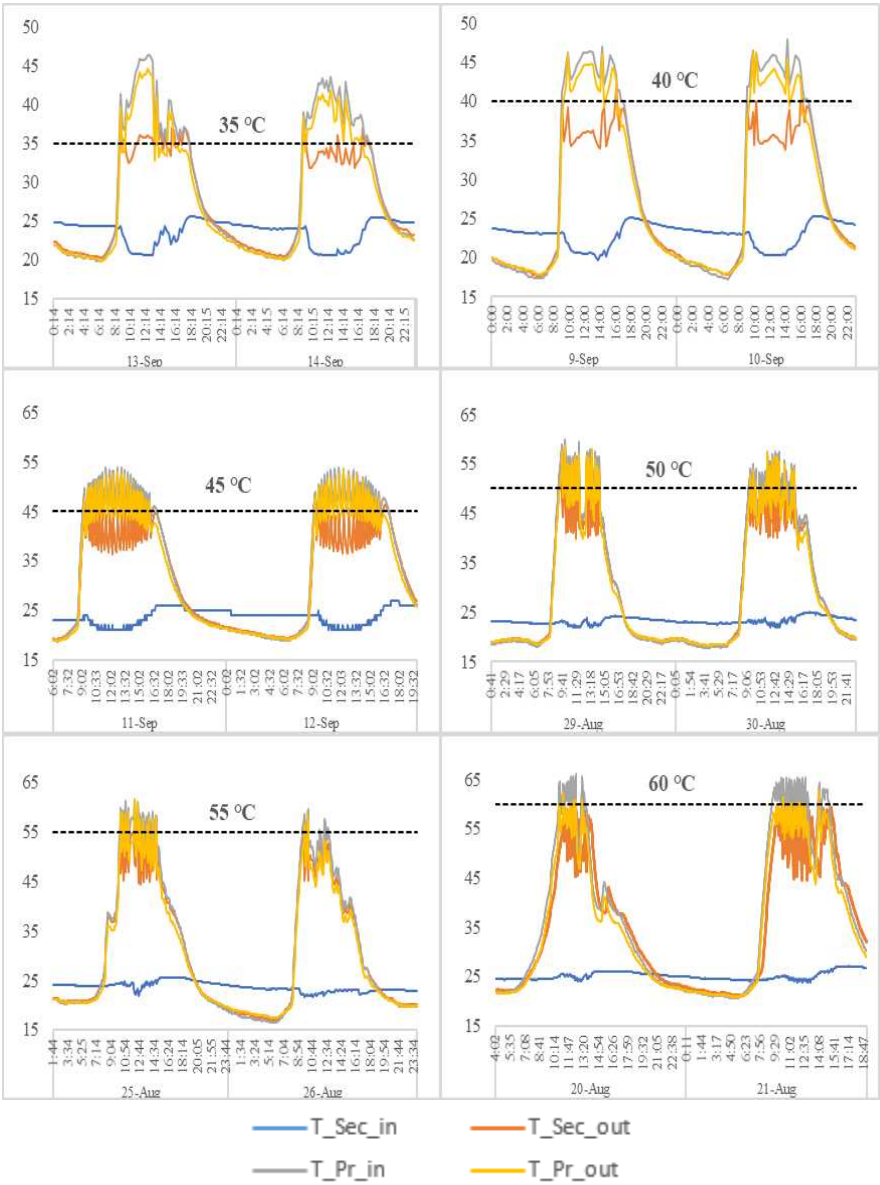


Figure 8
Actual experiments for a different set of output temperatures

The heat exchanger's secondary side output temperature as a function of the primary side input temperature was conducted in a linear correlation with a (0,0) set intercept to estimate the strength of the connection between the most critical parameters of the heat exchanger. As a result, in Figure 9, the coefficients of determination are always higher than 0.9863, indicating the performance of the heat exchanger. It shows that the differences between the two parameters are as low as possible.

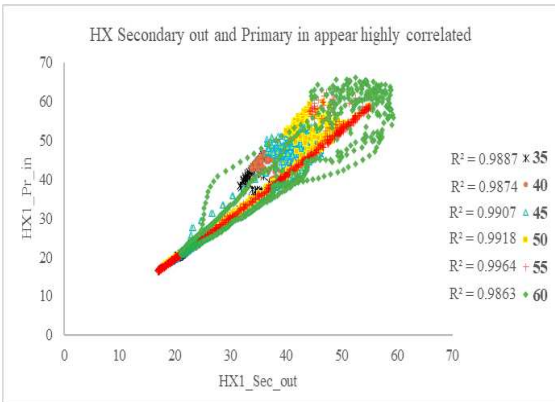


Figure 9
Heat exchanger primary input versus secondary output temperature

In addition, the logarithmic mean temperature difference (LMTD) is a tool in flow systems, most notably in heat exchangers, to determine the temperature driving force for the heat transfer process [29]. For a heat exchanger with a specific heat transfer coefficient and constant area, the larger the LMTD, means more heat is transferred. For a generic heat exchanger with two ends at which the cold and hot streams enter, the LMTD is defined by the following equation:

$$LMTD = \frac{\Delta T_p - \Delta T_s}{\ln\left(\frac{\Delta T_p}{\Delta T_s}\right)} \tag{1}$$

Where ΔT_s are the temperature changes in the heat exchanger's primary and secondary sides, respectively. In our case, the heat exchanger is a counter-current where the streams enter from different ends. It should be noted that the LMTD is a steady-state concept, and it cannot be applied to a dynamic analysis [30].

Figure 10 shows the LMTD of all real cases to estimate the average value. As LMTD is higher, we have better performance of the heat exchanger. It is noted that all cases have average LMTDs higher than eight, while the higher the value means more heat has been transferred between the two sides.

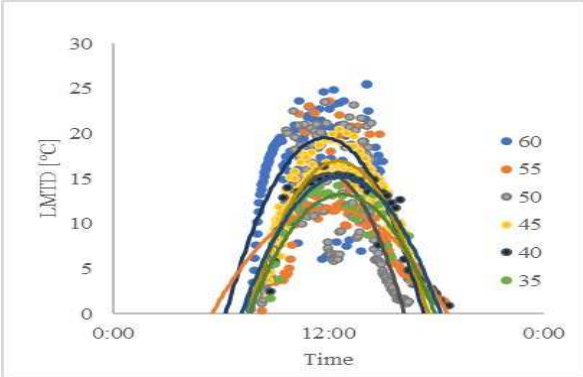


Figure 10
LMTD performance of the heat exchanger

4.3 Parametric Analysis

4.3.1 Effect of Solar Radiation and Mass Flowrate on the Outlet Temperature

Figure 11 shows the outlet temperature variation under varying solar radiations for three different configurations, 30/30, 60/60 and 90/90 litres per hour per square meter mass flow rate. In contrast, the inlet water temperature of the secondary loop was relatively constant at 25°C return temperature. The simulation was conducted in July, and the results show that the 30/30 configuration has the highest output temperature for summertime compared to 60/60 and 90/90 $l/h^{-1}m^{-2}$.

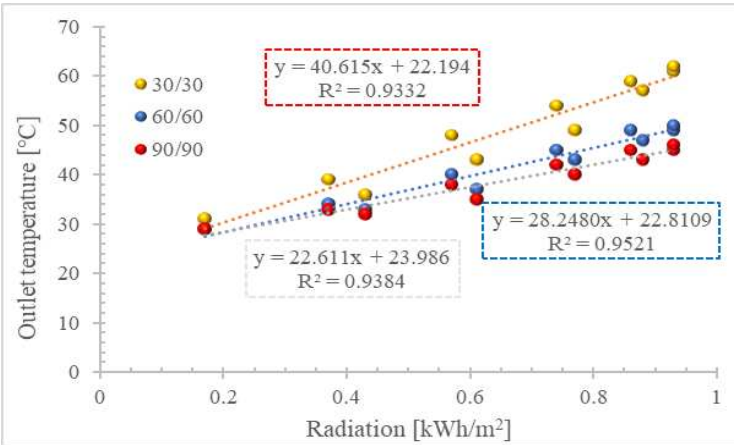


Figure 11
Effect of solar radiation on water outlet temperature for tankless SHIP system at different flowrates

4.3.2 Effect of Solar Gain under Different Mass Flowrate

Figure 12 illustrates the variation of the generated thermal energy fraction from the tankless SHIP system for the specific industrial process under varying solar radiation and mass flowrate. Results show that the 30/90 configuration has the highest annual energy yield by 638.79 kWhm^{-2} and 1,226 kWh for the whole system. In the second place, the whole system has a 30/60 configuration of 604.05 kWhm^{-2} and 1,160 kWh. It is noted from the graph that the second loop must have a higher flow rate compared to the primary loop, which is a result of choosing a closed loop on the primary side. At the same time, the open loop on the secondary side shows that a higher flow rate results in higher yields if we consider a fixed flow rate at the primary loop.

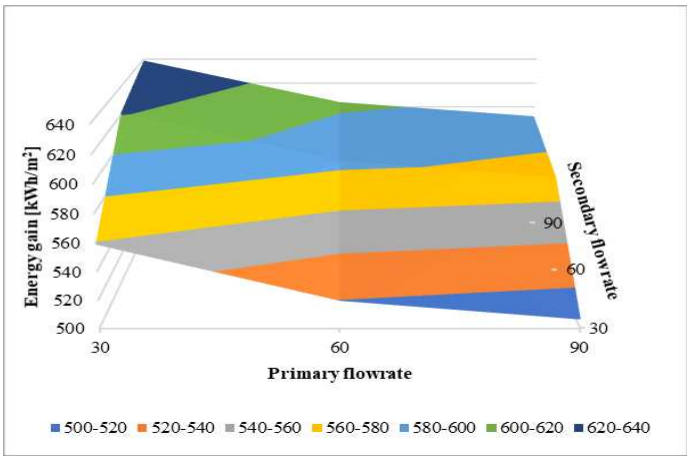


Figure 12
Annual energy yield considering different flowrate configurations

Conclusion

A novel tankless solar heat for industrial process (SHIP) system comprising FPC and an external heat exchanger has been designed. A dynamic simulation model has been developed in T*sol software, which is thus authenticated using prototype implementation. The proposed tankless SHIP system can fulfil the low to medium heat demand for pasteurizing process heat demand. This work's primary objective is to study the possibility of affording solar energy for an industrial process requiring low to medium temperatures at different output temperatures. The case study is considered for a food processing plant in Budapest, Hungary.

The modelling system has been tested experimentally based on the thermodynamics and heat transfer phenomena. The meteorological data has been used from MeteoSyn software in the T*Sol program. Moreover, six different temperature configurations were chosen as a set temperature of 35, 40, 45, 50, 55,

and 60°C, respectively. Each measurement took several days during the summertime of 2021. The meteorological data have been examined between average aggregated data between 1995-2012 and one-year actual experiment weather data. Results show less than 10% difference between the two solar radiation profiles and less than 1% in the ambient air temperature profiles.

On the other hand, a narrower comparison was conducted during the three months of the experiment. The results show a high correlation with more than $R^{2,}$ more significant than 0.7500 for the coefficients of determinations for both solar radiation and ambient temperature profiles. In practice, this methodology and the IMRe controlling system met all set temperature configurations. After that, all coefficients of determination of the primary input temperature and the secondary output temperature have been examined. The results show that R^2 is higher than 0.9800 for all scenarios. Also, the performance of the external heat exchanger was measured by the logarithmic mean temperature difference LMTD, and the results show high performance with bigger than 7.9°C for all scenarios. The root means squared error RMSE was evaluated for the ambient temperature and was around 5%, proving the proposed methodology. Finally, parametric analysis has been evaluated to determine the best flow rate configuration on the primary and secondary sides. The results show that 30/90 has the best performance with 638.79 $kWhm^{-2}$ and 1,226 kWh for the whole system. In conclusion, tankless SHIP systems can serve as a considerable solution for low to medium heat temperatures in the industrial sector and under central European weather conditions.

Acknowledgement

This work was supported by the Stipendium Hungaricum Programme and by the Mechanical Engineering Doctoral School, Hungarian university of agriculture and life sciences , Gödöllő, Hungary. A special thanks to Passzívhaz KFT for the intelligent measurement system IMRe.

References

[1] E. Torres-Reyes, J. J. Navarrete-González, A. Zaleta-Aguilar, and J. G. Cervantes-de Gortari, “Optimal process of solar to thermal energy conversion and design of irreversible flat-plate solar collectors,” Energy, Vol. 28, No. 2, pp. 99-113, 2003, doi: 10.1016/S0360-5442(02)00095-6

[2] I. Luminosu and L. Fara, "Determination of the optimal operation mode of a flat solar collector by exergetic analysis and numerical simulation," energy, Vol. 30, No. 5, pp. 731-747, 2005, doi: 10.1016/j.energy.2004.04.061

[3] A. Shukla, D. Buddhi, and R. L. Sawhney, "Solar water heaters with phase change material thermal energy storage medium: A review," Renew. Sustain. Energy Rev., Vol. 13, No. 8, pp. 2119-2125, 2009, doi: 10.1016/j.rser.2009.01.024

[4] A. S. Wallerand, M. Kermani, R. Voillat, I. Kantor, and F. Maréchal, "Optimal design of solar-assisted industrial processes considering heat pumping: Case study of a dairy," Renew. Energy, Vol. 128, pp. 565-585, 2018, doi: 10.1016/j.renene.2017.07.027

[5] L. Kumar, M. Hasanuzzaman, N. A. Rahim, and M. M. Islam, "Modeling, simulation and outdoor experimental performance analysis of a solar-assisted process heating system for industrial process heat," Renew. Energy, Vol. 151, No. 10, p. 1017, 2021, doi: 10.1016/j.renene.2020.09.062

[6] L. Kumar, M. Hasanuzzaman, and N. A. Rahim, "Global advancement of solar thermal energy technologies for industrial process heat and its future prospects: A review," Energy Convers. Manag., Vol. 195, No. February, pp. 885-908, 2019, doi: 10.1016/j.enconman.2019.05.081

[7] N. Cottrel and E. Menichetti, "Solar Heat for Industrial Processes (SHIP) - State of the art in the Mediterranean region," Obs. Méditerranéen l'Energie, p. 64, 2011

[8] C. Lauterbach, B. Schmitt, U. Jordan, and K. Vajen, "The potential of solar heat for industrial processes in Germany," Renew. Sustain. Energy Rev., Vol. 16, No. 7, pp. 5121-5130, 2012, doi: 10.1016/j.rser.2012.04.032.

[9] K. Vajen, C. Lauterbach, and B. Schmitt, "Solar heat for industrial processes - Potential, technologies and applications," Int. Conf. Sol. energy MENA Reg., No. 2002, pp. 28-34, 2012 [Online]. Available: www.solar.uni-kassel.de

[10] C. A. Schoeneberger, C. A. McMillan, P. Kurup, S. Akar, R. Margolis, and E. Masanet, "Solar for industrial process heat: A review of technologies, analysis approaches, and potential applications in the United States," Energy, Vol. 206, p. 118083, 2020, doi: 10.1016/j.energy.2020.118083

[11] M. Calderoni, M. Aprile, S. Moretta, A. Aidonis, and M. Motta, "Solar thermal plants for industrial process heat in Tunisia: Economic feasibility analysis and ideas for a new policy," Energy Procedia, Vol. 30, pp. 1390-1400, 2012, doi: 10.1016/j.egypro.2012.11.153

[12] S. H. Farjana, N. Huda, M. A. P. Mahmud, and R. Saidur, "Solar process heat in industrial systems – A global review," Renewable and Sustainable Energy Reviews, Vol. 82. pp. 2270-2286, 2018, doi: 10.1016/j.rser.2017.08.065

[13] R. K. Pachauri and L. Meyer, "Climate Change 2014 Synthesis Report," 2014, doi: 10.1177/0002716295541001010

[14] Y. Bhusal, A. Hassanzadeh, L. Jiang, and R. Winston, "Technical and economic analysis of a novel low-cost concentrated medium-temperature solar collector," Renew. Energy, Vol. 146, pp. 968-985, 2020, doi: 10.1016/j.renene.2019.07.032

[15] M. Ghazouani, M. Bouya, and M. Benaissa, "Thermo-economic and exergy analysis and optimization of small PTC collectors for solar heat integration in industrial processes," Renew. Energy, Vol. 152, pp. 984-998, 2020, doi: 10.1016/j.renene.2020.01.109

[16] M. Lados, É. S. Somossy, and T. Tóth, "Financial subsidies and the location decision of solar power plants in Hungary: An empirical investigation," Reg. Stat., Vol. 10, No. 2, pp. 166-185, 2020, doi: 10.15196/RS100207

[17] B. Kumar et al., "Trendline assessment of solar energy potential in hungary and current scenario of renewable energy in the visegrád countries for future sustainability," Sustain., Vol. 13, No. 10, 2021, doi: 10.3390/su13105462

[18] Hungarin Ministry of National Development, National Energy Strategy 2030. Budapest, 2012

[19] T. Khanam, A. Rahman, B. Mola-Yudego, P. Pelkonen, Y. Perez, and J. Pykäläinen, "Achievable or unbelievable? Expert perceptions of the European Union targets for emissions, renewables, and efficiency," Energy Res. Soc. Sci., Vol. 34, No. July, pp. 144-153, 2017, doi: 10.1016/j.erss.2017.06.040

[20] J. Dyduch and A. Skorek, "Go South! Southern dimension of the V4 states' energy policy strategies – An assessment of viability and prospects," Energy Policy, Vol. 140, No. June 2019, p. 111372, 2020, doi: 10.1016/j.enpol.2020.111372

[21] M. Jahangiri, E. T. Akinlabi, and S. M. Sichilalu, "Assessment and Modeling of Household-Scale Solar Water Heater Application in Zambia: Technical, Environmental, and Energy Analysis," Int. J. Photoenergy, Vol. 2021, 2021, doi: 10.1155/2021/6630338

[22] K. Urbán, S. Bíró-szigeti, and B. Pataki, "Technology Road-Mapping of Final-Energy Generator Technologies , for the Industrial Sector , in Developed European Economies Basic Concepts of the Final-Energy Industry," Vol. 19, No. 4, pp. 65-84, 2022

[23] R. Kalbasi et al., "Finding the best station in Belgium to use residential-scale solar heating, One-year dynamic simulation with considering all system losses: Economic analysis of using ETSW," Sustain. Energy Technol. Assessments, Vol. 45, No. February, p. 101097, 2021, doi: 10.1016/j.seta.2021.101097.

[24] P. Kádár, "Pros and cons of the renewable energy application," Acta Polytech. Hungarica, Vol. 11, No. 4, pp. 211-224, 2014, doi: 10.12700/aph.25.04.2014.04.14

[25] M. Jahangiri, A. Alidadi Shamsabadi, and H. Saghaei, "Comprehensive Evaluation of Using Solar Water Heater on a Household Scale in Canada,"

J. Renew. Energy Environ., Vol. 5, No. 1, pp. 35-42, 2018 [Online] Available: http://www.jree.ir/article_88491.html%0Ahttp://www.jree.ir/pdf_88491_51a9a7966f493d8992890ecd5bfcf104.html

[26] S. Pahlavana, M. Jahangirib, A. Shamsabadic, and A. Khechekhouched, "Feasibility study of solar water heaters in Algeria, a review," J. Sol. Energy Res., Vol. 3, No. August, pp. 135-146, 2018

[27] B. József, H. Péter, and D. Zoltán, EXAMINATION OF INTELLIGENT MEASUREMENT SYSTEM (IMRE) APPLICATIONS, no. October. Szeged, 2018

[28] P. Hermanucz and J. Benécs, APPLYING THE INTELLIGEN MEASURING SYSTEM (IMRe) TO A COOLING SYSTEM USED IN THE FOOD INDUSTRY. Gödöllő, Hungary: Szent István Egyetemi Kiadó, 2019

[29] Q. Li, X. Huang, Y. Tai, W. Gao, L. Wenxian, and W. Liu, "Thermal stratification in a solar hot water storage tank with mantle heat exchanger," Renew. Energy, Vol. 173, pp. 1-11, 2021, doi: 10.1016/j.renene.2021.03.105

[30] R. Bálint, A. Fodor, and A. Magyar, "Model-based power generation estimation of solar panels using weather forecast for microgrid application," Acta Polytech. Hungarica, Vol. 16, No. 7, pp. 149-165, 2019, doi: 10.12700/APH.16.7.2019.7.9

Effect of Ultrasound on the Austenite Transformation of Shape Memory Alloys

Ali H. Alhilfi

University of Misan, Baghdad road R4QR+5R8, 62001, Maysan, Iraq
alihameed@stud.uni-obuda.hu

Endre Ruszinkó

Óbuda University, Népszínház St. 8, H-1081 Budapest, Hungary
ruszinko.endre@uni-obuda.hu

Abstract: The paper aims to develop a model describing the shape memory deformation in the presence of ultrasonic impulses. The application of ultrasound can be of great importance, primarily when heating the shape memory alloy elements cannot be carried out for many applications. Experimental results record that acoustic waves induce the austenitic transformation, which manifests itself in a jump-wise increment in the deformation as ultrasound is On. The model presented here is constructed in terms of the synthetic theory of irrecoverable deformation. To catch the phenomena caused by acoustic energy, we enter into the basic equation a new term reflecting the effect of ultrasound on the processes governing the phase transformation. The analytical results fit good experimental data.

Keywords: shape memory alloys; austenitic transformation; ultrasound

1 Introduction

Today Shape Memory Alloys (SMAs) are already commercially applied in many technical fields. SMAs have been developed since the early 1960s, and since then, they have been successfully used for medical [1, 2, 3], robotic [4, 5, 6], aerospace [7, 8, 9], and automobile applications [10, 11].

The superior property of SMA is that it can go through solid-state phase transformations, meaning it can be stretched, bent, heated, cooled, and still remember its original shape. SMAs are widely used for medical implants in medicine due to their kink resistance, stress constancy, high elasticity, and corrosion resistance. If we speak of electronic devices and robotic systems, SMA

actuators, sensors, and controllers have drawn significant attention and interest due to their unique properties and are expected to be equipped in many modern vehicles at competitive market prices [12]. The essential advantage is that active elements (e.g., SMA wire or spring) can be deformed by applying minimal external force and retain their previous form when subjected to certain stimuli such as thermomechanical or magnetic changes. In aerospace, SMA release devices can be actuated slowly, avoiding satellite shock failures. This application is essential for satellites because it can also be used for 'microsatellites' since compact separation devices with minimal SMA release triggers can be made [13].

As mentioned above, SMAs show a particular ability to recuperate the original shape while heating above specific critical temperatures (shape memory effect). In other words, they can recover a large unelastic deformation or create high recovery stress on heating within the temperature range of martensitic transformation. However, heating the shape memory alloy elements cannot be carried out for many applications. Other ways to initiate shape memory effects may be applied in these cases. It is known that strain recovery may be initiated by neutron irradiation, hydrostatic pressure, and ultrasonic action [14]. The last method is the most attractive because ultrasonic vibration does not require expensive equipment like other methods.

This paper concentrates our attention on the effect of ultrasonic vibrations on the shape memory effect (austenitic transformation on heating). Experimental observations on this topic [15, 16, 17, 18, 19, 20] can be summarised as follows (Fig. 1):

I. Ultrasonic vibrations impulsively added to austenitic transformation on heating result in negative strain jumps. In other words, acoustic energy can initiate strain variations of SMA.

II. The magnitude of the strain jumps increases with the ultrasonic vibration amplitude.

III. The effect of insonation strongly depends on the moment when the ultrasound is applied. Acoustic energy results in no effect if it acts outside the austenite transformation temperature range, i.e., before A_s or after A_f. Further, the magnitude of the ultrasound-induced strain jumps is not distributed uniformly within the austenite transformation temperature range. This phenomenon reaches its maximum if the alternate stresses are applied approximately in the middle of the temperature range of phase transformations. It can be explained by the fact that the action of ultrasound manifests itself in full force when the amount of material involved in transformation is maximum. The lattice is very soft during transformation, and the phase boundaries are very movable. In this case, any external action, for instance, alternate stress, results in the appearance of an additional quantity of preferably oriented

domains, which leads to a further strain variation. The number of phase boundaries reaches its maximum when half of the alloy is transformed into the austenite phase [14].

IV. It was found that the series of ultrasonic impulses led to the finish temperature being less than during conventional heating. Since acoustic energy boosts the transformation processes, it is logical to assume that they sooner reach their completeness. In other words, the temperature needed to finish the transformation is partially compensated by ultrasound heating.

V. After switching of ultrasound, the further realization of SME is observed according to the reverse transformation kinetics. However, immediately after the ultrasound is off, some "backsliding" in austenitic deformation, a slight increase of deformation, is observed; 67-77℃ temperature range in Fig. 1. This aftereffect is assumed to be due to a) the decrease in temperature after ultrasound is off and b) the action of ultrasound which "left a trail" in the form of ultrasound-assisted defect conglomeration, reducing the development of the phase transformations. Therefore, while the central portion of acoustic energy converts irreversibly into the phase deformation increment, some fraction of it recovers.

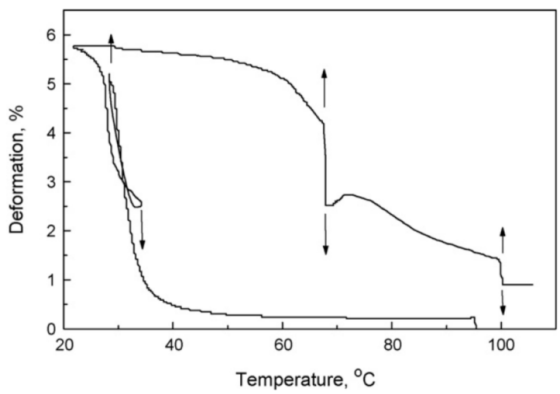


Figure 1

State diagram of NiTi alloy in deformation-temperature coordinate. The sample is subjected to uniaxial tension $\sigma = 30$ MPa . The arrows show the moments of switching-on (↑) and switching-off (↓) of ultrasonic vibrations [16].

Experimental papers clearly show that stress and temperature are equal stimuli for initiating martensite/austenite transformations, i.e., the same mechanical effects can be achieved by employing both stress and temperature. Since ultrasound is a carrier of both these effects – alternating stress and increase in temperature caused by them – the physical substantiation of the phenomena observed above can be summarised as follows [14, 15, 16, 17, 18, 19, 20]:

I. The variation (increase) in austenite deformation can be explained by ultrasonic heating of the sample due to ultrasound waves' energy dissipation and the formation of steady austenite.

II. Acoustic energy increases the mobility of interfaces (phase and domains) by decreasing the efficient friction force caused by alternate stresses.

III. The superposition of alternate stresses induces the movement of interface and martensitic domain boundaries (within the temperature range of reverse martensitic transformations).

IV. The lattice is very soft during transformation, and the phase boundaries are very movable. In this case, any external action, for instance, alternate stress, results in the appearance of an additional quantity of preferably oriented domains, which leads to a further strain variation. The number of phase boundaries reaches its maximum when half of the alloy is transformed into the austenite phase.

This paper aims to develop a model describing analytically the austenitic transformation of NiTi alloy subjected to ultrasound impulses. For the mathematical apparatus, we take the synthetic theory of irrecoverable deformation, which has established itself as a reliable instrument to model many problems in the field of phase transformation [21, 22, 23].

2 Synthetic Theory

The synthetic theory operates with notions used in the Ilyushin deviatoric space, such as stress and strain vector, **S** and **e** [23]. For a three-dimensional subspace (S^3) of the Ilyushin five-dimensional deviatoric space, we have

$$\begin{aligned} &S_1 = \sqrt{3/2}S_{xx},\ S_2 = S_{xx}/\sqrt{2} + \sqrt{2}S_{yy},\ S_3 = \sqrt{2}S_{xz} \\ &e_1 = \sqrt{3/2}e_{xx},\ e_2 = e_{xx}/\sqrt{2} + \sqrt{2}e_{yy},\ e_3 = \sqrt{2}e_{xz} \end{aligned} \tag{1}$$

where S_{ij} and e_{ij} are the stress and strain deviator tensor components, respectively. As a two-level model, the synthetic theory calculates the macroscopic deformation as an averaging (summation) of deformations occurring on the micro level of material. For the macro level, we take the elementary volume of the body, which is considered a point in the mathematical sense. This volume consists of a large number of micro volumes, each being an element of the continuum capable of deforming under external effects. Thus, macrodeformation is defined as the integration (summation) of deformations occurring on the microscopic level of material. As a result, we propose the following formula for strain vector components (e_k) [23]

$$e_k = \iiint_V \varphi_N N_k dV, \;\; k = 1,2,3. \tag{2}$$

The formula above is derived for such stress states when $\mathbf{S} \in \mathsf{S}^3$. In Eq. (2), φ_N is named strain intensity; it is an average measure of deformation within one microstructure element. N_k are components of unit vector ($\mathbf{N}$) giving the orientation of this element, which are defined via spherical angles α, β, and λ in. The range of angles α, β, and λ is $0 \le \alpha \le 2\pi$, $0 \le \beta \le \pi/2$, and $0 \le \lambda \le \pi/2$ [23]. In terms of the Synthetic theory, the development of inelastic deformation is derived via planes in S^3, their positions are defined through angles α, β, and λ as shown in Fig. 2 [23].

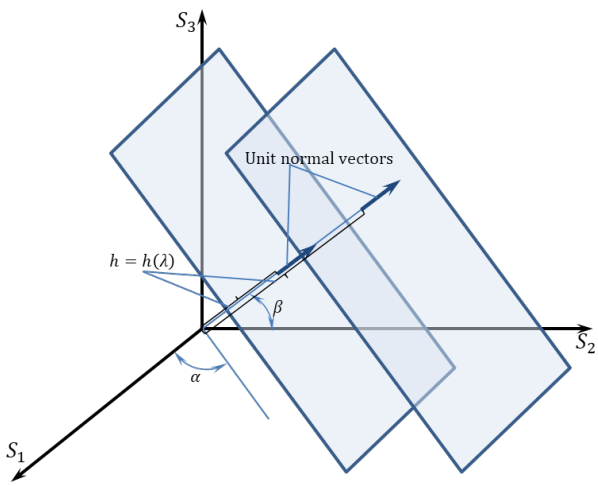


Figure 2
Locations of planes in S^3 given by angles α and β, and the plane distance $h(\lambda)$

Finally, dV from (2) is an elementary set of micro volumes involved in the deformation [23]

$$dV = \cos\beta d\alpha d\beta . \tag{3}$$

Now, we have to define the strain intensity. To apply Eq. (2) to the description of deformations induced by phase transformations, we relate the strain intensity rate to the rate of martensite fraction ($\dot{\Phi}$) [23]:

$$r\dot{\varphi}_N = \dot{\Phi}, \tag{4}$$

where r is the model constant. We define $\dot{\Phi}$ as [23]

$$\dot{\Phi} = -\frac{\dot{T}_e}{M_s - M_f}, \tag{5}$$

where $\dot{T}_e$ is the rate of effective temperature, which will be defined below. This formula holds for martensitic transformation at

$$\dot{T}_e < 0 \text{ and } M_f < T_e < M_s. \tag{6}$$

For austenitic transformation, we write [23]

$$\dot{\Phi} = -\frac{\dot{T}_e}{A_f - A_s}, \dot{T}_e > 0 \text{ and } A_s < T_e < A_f \tag{7}$$

Formulae (5)-(7) give a linear relationship between the martensite fraction and effective temperature, which is widely accepted in the scientific community (Fig. 3).

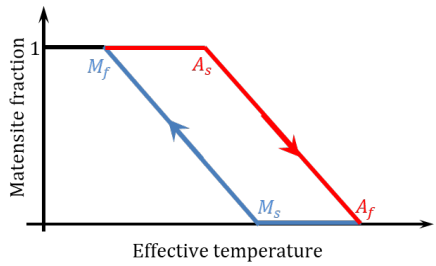


Fig. 3
Φ □ T_e graph plotted via Eqs. (5)-(7)

In Eq. (7), T_e is effective temperature proposed in terms of the structural-analytic model by [24], through the Clausius-Clapeyron equation, as [23]

$$T_e = T\left(1 - D\mathbf{S}\cdot\mathbf{N}\right), \tag{8}$$

where D is the model constant. Formula (8) determines the start/finish temperatures as a function of temperature and acting stress [23, 24]. Fig. 4 demonstrates the change in effective temperature from (8) for differently orientated elements (for simplicity, we set $\lambda = 0$). As one can see, the start and finish of the transformation strongly depend on the orientation of the element we consider.

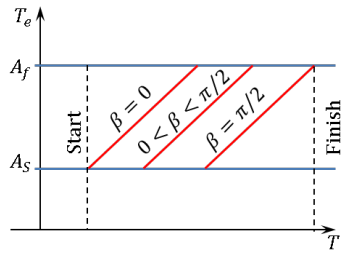


Figure 4
Effective tempreture for diffrent directions ($\mathbf{S} = \text{const}$)

Eq. (8) enables one to account for the shift of the characteristic temperatures caused by loading. Summarising, formulae (5)-(8) define the amount of martensite as a single-valued function of temperature and acting load. The scalar product $\mathbf{S} \cdot \mathbf{N}$ gives the resolved shear stress acting in the element with $\mathbf{N}$-orientation. This fact reflects the well-known fact that external load manifests itself differently depending on how preferable the element's/slip system's orientation is.

With Eqs. (4), (7), and (8), φ_N for austenitic transformaton is

$$r\dot{\varphi}_N = \dot{\Phi} = -\frac{\dot{T}_e}{A_f - A_s}, \tag{9}$$

Differentiating in (8), we get

$$r\dot{\varphi}_N = -\dot{T}_e = -\dot{T}\left(1 - D\mathbf{S} \cdot \mathbf{N}\right) + TD\dot{\mathbf{S}} \cdot \mathbf{N}. \tag{10}$$

In the formula above, the constant r includes $A_f - A_s$.

Consider austenitic transformation on heating when the material is under the action of constant stress, $\dot{\mathbf{S}} = 0$. Eq. (10) gives

$$r\dot{\varphi}_N = -\dot{T}\left(1 - D\mathbf{S} \cdot \mathbf{N}\right) \tag{11}$$

Let us apply the above formula for the case of uniaxial tension when the stress vector, according to (1), has only one non-zero component, $S_1 = \sqrt{2/3}\sigma \equiv S$, and, together with (3), we have [23]

$$r\dot{\varphi}_N = -\dot{T}\left(1 - DS_1N_1\right) = -\dot{T}\left(1 - DS\sin\beta\cos\lambda\right). \tag{12}$$

On integrating in (12), we have

$$r\varphi_N = -T\left(1 - DS\sin\beta\cos\lambda\right) + C \tag{13}$$

The integration constant C is determined from the condition that the austenite transformation terminates ($\varphi_N = 0$) as the effective temperature reaches the

austenite finish temperature, A_f. Since the transformation completion takes place at $\beta = \pi/2$ and $\lambda = 0$, we can calculate the austenitic transformation finish temperature, T_f. Equating T_e from (8) for the specified angle values to A_f, we obtain $T_f = A_f/(1-DS)$. Now, Eq. (13) at $\varphi_N = 0$, $T = T_f$, $\beta = \pi/2$, and $\lambda = 0$ gives $C = A_f$. So

$$r\varphi_N = -\left(T - A_f\right) + TDS\sin\beta\cos\lambda. \tag{14}$$

Fig. 5 schematically shows the $\varphi_N \sim T$ graphs plotted with Eq. (14) for different values of β. Again, it is easy to see that different elements are involved in the transformation for different temperature ranges. Fig. 6 demonstrates the $\varphi_N \sim \beta$ graphs plotted with Eq. (14) for different effective temperatures. As one can see, at the start of martensitic transformation, when the material is in a martensitic state ($\Phi = 1$), φ_N takes positive values over the whole diapason of angle β (lines 1 and 2). As the temperature increases, the domain of angles β with positive strain intensities decreases and finally shrinks to the point $\beta = \pi/2$ where $\Phi = 0$). The boundary values of angles λ and β (points 1, 1', 1" for lines 3-5) are obtained from conditions $\varphi_N = 0$ and $\lambda = 0$

$$\cos\lambda_1 = \frac{1}{DS\sin\beta}\left(1 - \frac{A_f}{T}\right), \quad \sin\beta_1 = \frac{1}{DS}\left(1 - \frac{A_f}{T}\right). \tag{15}$$

Here we assume that $T > A_f$; for $T < A_f$ we let $\beta_1 = 0$ and $\lambda_1 = \pi/2$ (lines 1 and 2 in Fig. 6).

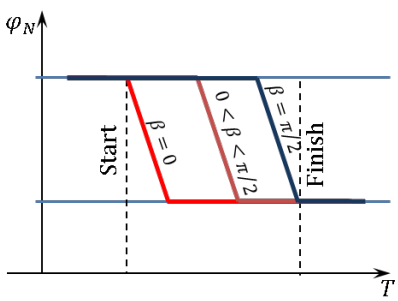


Figure 5
$\varphi_N \sim T$ plots for different angles β ($\lambda = 0$, $\mathbf{S} = \text{const}$).

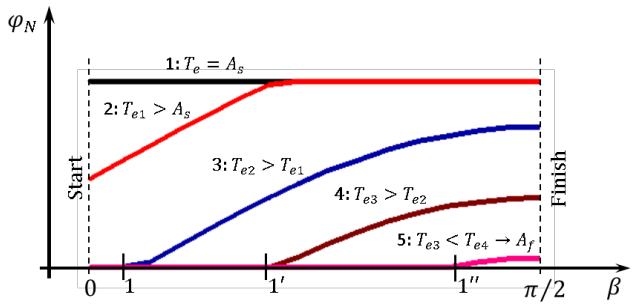


Figure 6
$\varphi_N \sim \beta$ plots for different effective temperatures ($\lambda = 0$, $\mathbf{S} = \text{const}$)

Therefore, the initial stage of martensite-austenite transition is described by Eq. (2) with the integration diapason $0 \le \alpha \le 2\pi$, $0 \le \beta \le \pi/2$, and $0 \le \lambda \le \pi/2$ (lines 1 and 2 in Fig. 6), and in the course of temperature increase, when the domain of non-zero strain intensities shrinks (lines 3-5 in Fig. 6), the integral (2) becomes

$$e = \int_0^{2\pi}\int_0^{\lambda_1}\int_{\beta_1}^{\pi/2} \varphi_N N_1 \cos\beta d\alpha d\beta d\lambda. \tag{16}$$

3 Extension of the Synthetic Theory to the Ultrasound-assisted Shape Memory Deformation

Since effective temperature is the most crucial factor in the progress of phase transformation – T_e directly influences the values of Φ and φ_N –, we propose to extend Eq. (8) by a term reflecting the presence of ultrasound. Following the observations on the effect of ultrasound on austenitic transformation, we propose to shift (increase) the value of effective temperature by a term (U) representing the ultrasound action:

$$T_e = T\left(1 - D\mathbf{S}\cdot\mathbf{N}\right) + U. \tag{17}$$

We decompose the U on two components to reflect the thermal and the mechanical impact of acoustic energy on the kinetics of the austenitic transformation.

$$U = f\left(S_U\right)\cdot g\left(T\right). \tag{18}$$

The first term points out the assisting action of ultrasound due to the generation of alternating stresses that induce and intensify the movement of interface and martensitic domain boundaries. We define $f\left(S_U\right)$ as

$$f\left(S_U\right)=\left(B+e^{-w\left(T-T_i\right)}\right)\int_{A_s}^{T}U_1\left(\mathbf{S}_U\cdot\mathbf{N}\right)dt, \tag{19}$$

where T_i are the temperatures as ultrasound is On; U_1, w, and B are model constants and $\mathbf{S}_U$ is a stress vector whose components are formed via Eq. (1) by the values of alternating stress amplitudes. Considering short-termed ultrasound impulses, f is assumed to increase in a jump-wise manner as the ultrasound is On. Then, it decreases, stabilizing at some value after the ultrasound is Off. The overall effect of the ultrasonic impulses is an increase in f. Although $\mathbf{S}_U=0$ between the ultrasound impulses, $f\left(S_U\right)>0$ throughout the transformation due to the integration in (19) that cumulates positive values during the sonication.

We propose the term $U_1\left(\mathbf{S}_U\cdot\mathbf{N}\right)$ in (19) to comply with experimental data stating that the magnitude of the strain jumps is proportional to stress amplitudes ($\mathbf{S}_U$), and the effectiveness of ultrasound varies depending on the orientation of the microregion considered. It is the scalar product $\mathbf{S}_U\cdot\mathbf{N}$ that reflects the effectiveness of the ultrasound action for a given orientation. The presence of $\left(B+e^{-w\left(T-T_i\right)}\right)$ in (19) reflects the experimental fact that after ultrasound is Off, the so-called aftereffect is recorded, consisting of some deformation increase. As $e^{-w\left(T-T_i\right)}\to 0$, the further realization of SME takes place according to the austenite transformation kinetics, although shifted by B. To summarize, the effect of ultrasound manifests itself in the recoverable and irrecoverable portions of phase deformation.

In order to catch the experimental observation recording that the strain-jump magnitude strongly depends on the temperature when ultrasound is applied, we propose the function $g\left(T\right)$ from (18) in the form of the Agnesi curve:

$$g\left(T\right)=\frac{a^3}{a^2+\left(T-C\right)^2}. \tag{20}$$

While the function f leads to uniform strain increments at a given ultrasonic intensity, the function g gives different strain-jump magnitudes depending on the transformation stage when the acoustic energy acts. The model constants a and C from (20) reflect the material's compliance to react to the ultrasonic impact at different moments of the transformation.

Figure 7 demonstrates the $U(T)$ plot, which shows its step-wise increments at the temperatures as the ultrasound is On. The magnitudes of these increments vary with the temperature.

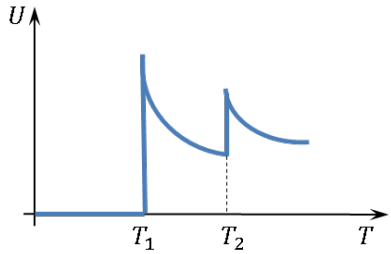


Figure 7
Dependence of U on temperature; ultrasonic impulses are applied at $T = T_1$ and $T = T_2$

Consider the case when a material is fully austenitic ($\Phi = 0$) and subjected to the action of constant uniaxial stress. Let an impulse of longitudinal ultrasound at a given temperature be applied. In this case, the vector $\mathbf{S}_U$, according to (1), has components $\left(\sqrt{2/3}\sigma_m, 0, 0\right)$, where σ_m is the amplitude of oscillating tension-compression stress.

Now, the strain intensity, on the base of Eqs. 14 and (17)-(20), is

$$r\varphi_{NU} = -\left(T - A_f\right) + \left(TDS - U\right)\sin\beta\cos\lambda. \tag{21}$$

The shape memory deformation (e_u), according to Eq. (2), takes the following form (the integral over α gives 2π)

$$e_u = \frac{\pi}{r}\int_0^{\lambda_{1U}}\int_{\beta_{1U}}^{\pi/2}\left[-\left(T - A_f\right) + \left(TDS - U\right)\sin\beta\cos\lambda\right]\sin 2\beta\cos\lambda d\lambda d\beta \tag{22}$$

The boundary angles in (22) are

$$\cos\lambda_{1U} = \frac{1}{\left(DS - U\right)\sin\beta}\left(1 - \frac{A_f}{T}\right),\ \sin\beta_{1U} = \frac{1}{\left(DS - U\right)}\left(1 - \frac{A_f}{T}\right). \tag{23}$$

Like in Eq. (15), we assume that $T > A_f$, otherwise, we let $\beta_{1U} = 0$ and $\lambda_{1U} = \pi/2$.

Figure 8 gives a comparison between $\varphi_{NU} \sim T$ and $\varphi_N \sim T$ plots, which demonstrate entire agreement with experimental observations:

Ultrasound induces the deformation recovery on heating, i.e., intensifies the shape memory deformation. This is due to the term U in Eq. (21) reflecting the promoting action of ultrasound within microelements. In an instant as ultrasound On, the deformation yields a negative increment and, after a short disturbance, follows the kinetics of austenitic transformation. Another effect from U is that the amount of material involved in the transformation increases due to ultrasonic impulse, i.e., by comparing (15) and (23), we have $\lambda_{1U} < \lambda_1$ and $\beta_{1U} > \beta_1$. Besides, formula (20) governs the magnitude of the strain jump depending on the temperature of ultrasonic insonation. Summing the strain intensities in all the micro volumes taking part in the austenitic transformation, Eq. (22), we obtain the deformation on the macroscopic scale.

Since acoustic energy boosts the phase transformation, it ends at less temperature than the ordinary shape memory effect – in Fig. 8, the φ_{NU} line reaches sooner than φ_N the zero value.

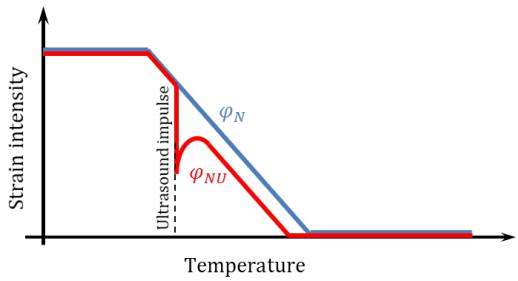


Figure 8
Strain intensity vs. temperature for an ordinary (φ_N) case and with ultrasound (φ_{NU})

Therefore, the extension of the synthetic theory expressed in formulae (17)-(23) leads to a qualitative correspondence with experiments. The next step is to inspect its quantitative correctness.

4 Results. Discussion

Here, our goal is to plot $\varepsilon \sim T$ diagram of NiTi alloy at fixed static stress subjected to ultrasonic insonation and compare it with the experiment. Characteristic temperatures of the alloy measured by differential scanning calorimetry were: $A_s = 323$ K, $A_f = 349$ K [16]. The procedure of tests was the following. The wire sample in the high-temperature austenite was subjected to load (uniaxial tension $\sigma = 30$ MPa that results in deformation of 1.9%) with subsequent cooling. After cooling, the sample was brought into austenitic condition by heating. The heating was performed at a rate of 1 K/min.

Two ultrasonic impulses (each of 9 sec [14-16]) with vibrational amplitude $A = 5$ μm and frequency $f = 22.2$ kHz were produced in the temperature range $A_s - A_f$. The first ultrasonic impulse was produced at $T_1 = 340$ K and the second at $T_2 = 373$ K.

To calculate the alternating stress amplitudes σ_m, we utilize the following relationship

$$\sigma_m = E\frac{2\pi f}{c}A, \tag{24}$$

where E and c is the NiTi alloy's Young modulus and the speed of sound, respectively. Taking into account that these are temperature functions, further, we utilize their average values for $A_s - A_f$ diapason: $E = 62\,\text{GPa}$, $c = 5200$ m/s [25], [26]. As a result, from (24), we have $\sigma_m = 8.3\,\text{MPa}$.

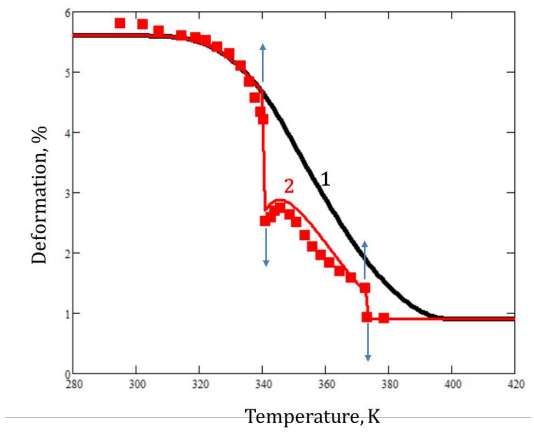


Figure 9

State diagram of NiTi alloy in deformation-temperature coordinate. The sample is subjected to uniaxial tension $\sigma = 30$ MPa. The arrows show the moments of switching-on (↑) and switching-off (↓)of ultrasonic vibrations; ■ –experiment [16], lines – model.

First, we plot the $\varepsilon \sim T$ diagram for the shape memory effect with the above data without ultrasonic action. To do this, we use Eqs. (14)-(16) for plotting line 1 in Fig. 9 with the following model constants: $D = 4.2\times10^{-3}\,\text{MPa}^{-1}$ and $r = 4.9\times10^{2}$ K. At least for the temperature diapason 290-340 K, i.e., before the first ultrasonic impulse, we can conclude that the model curve shows good agreement with the experiment.

The next step is the $\varepsilon \sim T$ diagram in the presence of ultrasonic impulses (line 2 in Fig. 9). Formulae (17)-(23) give the value of strain jumps shown in Table 1. The model results have been obtained with the following values of constants: $w = 1.32 \times 10^{-1}\,\mathrm{K}^{-1}$, $U_1 = 13.83(\mathrm{K} \cdot \mathrm{MPa})^{-1}$, $a = 4.0\,\mathrm{K}$, $C = 360\,\mathrm{K}$, $B = 4.79 \times 10^{-1}$. It must be stressed that the values of constants D and r stay the same as in the previous paragraph. The magnitude of the strain jump at the first impulse is much greater than that at the second. The reason is that the first impulse is done near the middle of the transformation while the second is closer to its end. The same result is obtained analytically because of the function $g(T)$ (20), where the constant C regulates the temperature of the greatest increment in the deformation caused by ultrasound. Therefore, the magnitude of strain jumps is governed by constants U_1 in $f(S_U)$ and a in $g(T)$, and the constant C in $g(T)$ regulates the temperature of the maximum ultrasound effect.

Table 1

Strain jumps due to ultrasound impulses

	Experiment	Model	Relative error, %
1st impulse	1.675	1.973	+15.1
2nd impulse	0.493	0.445	–9.7

The kinetics of $\varepsilon \sim T$ after switching off of ultrasound is also in full accordance with experimental results. Namely, after the strain jump, the model result shows some increase in the deformation, after which the deformation on heating develops according to the reverse transformation kinetics. This result is due to term $\left(B + e^{-w(T-T_i)}\right)$ in $f(S_U)$. Here, constant w governs the duration of the deformation's increase following the ultrasonic action, and B expresses an irrecoverable portion of deformation caused by ultrasound. Thus, the model accounts for another experimental fact stating that the ultrasound-assisted transformation ends at lower temperatures. In our case, the second ultrasonic impulse results in such a drop in deformation that the material achieves a fully austenitic state when the further increase in temperature results in no deformation variation. So, we read from Fig. 9 that the finish temperature after two acoustic impulses is 373 K, while the ordinary $\varepsilon \sim T$ diagram stops its variation at about 400 K.

In summarizing, we may conclude that the extension of the synthetic theory leads to qualitatively and quantitatively correct results.

Conclusion

We developed a model that predicts the kinetics of ultrasound-assisted austenitic transformation. This model was created in terms of the synthetic theory of irrecoverable deformation. Following the results of experimental observations,

stating that the presence of ultrasounds affects/translates the transformation temperatures, we extended the formula for effective temperature by the term accounting for the presence of acoustic energy. As a result, we derived a formula to calculate deformations as a function of temperature at constant static stress under the action of ultrasonic impulses. The analytical results obtained are in full accordance with the experiment. Namely, (i) deformation yields a jump-wise increment at the instant as ultrasound is On; (ii) after ultrasound is Off, the deformation incurs some increase, after which strain-temperature diagram runs according to austenitic transformation kinetics, although beneath the standard curve; (iii) the magnitude of strain jump strongly depends on the temperature when the acoustic impulse is produced – as approaching the middle of the phase temperature range the jump tends to a maximum; (iv) the presence of acoustic energy results in the decrease of the finish temperature. Since the analytical results show good accuracy, the model presented in this paper can be treated as an effective instrument for describing ultrasound-assisted austenitic transformations.

References

[1] Bansiddhi A, Sargeant T, Stupp S and Dunand D (2008) Porous nickel-titanium for bone implants: a review. Acta Biomaterialia 4, pp. 773-782

[2] Morgan N (2004) Medical shape-memory alloy applications—the market and its products. Materials Science and Engineering: A 378(1-2): 16-23

[3] Sun L, Huang WM, Ding Z et al. (2012) Stimulus-responsive shape-memory materials: a review. Materials & Design 33: 577-640

[4] Kim B, Lee MG, Lee YP, Kim Y and Lee G (2006) An earthworm-like micro robot using shape-memory alloy actuator. Sensors and Actuators A: Physical 125(2): 429-437

[5] Qin CJ, Ma PS and Yao Q (2004) A prototype micro-wheeled-robot using shape memory alloy actuator. Sensors and Actuators A: Physical 113(1): 94-99

[6] Wang Z, Hang G, Li J, Wang Y and Xiao K (2008) A micro-robot fish with embedded shape-memory alloy wire actuated flexible biomimetic fin. Sensors and Actuators A: Physical 144(2): 354-360

[7] Hartl D and Lagoudas DC (2007) Aerospace applications of shape-memory alloys. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering 221(4): 535-552

[8] Hartl D, Lagoudas D, Calkins F and Mabe J (2010a) Use of a Ni60Ti shape memory alloy for active jet engine chevron application: I. Thermomechanical characterization. Smart Materials and Structures 19(1): 015020

[9] Hartl D, Mooney J, Lagoudas D, Calkins F and Mabe J (2010b) Use of a Ni60Ti shape-memory alloy for active jet engine chevron application: ii.

Experimentally validated numerical analysis. Smart Materials and Structures 19(1): 015021

[10] Bellini A, Colli M and Dragoni E (2009) Mechatronic design of a shape memory alloy actuator for automotive tumble flaps: a case study. IEEE Transactions on Industrial Electronics 56(7): 2644-2656

[11] Stoeckel D (1990) Shape-memory actuators for automotive applications. Materials & Design 11(6): 302-307

[12] Jani, J. M., Leary, M., & Subic, A. (2014) Shape Memory Alloys in Automotive Applications. Applied Mechanics and Materials, 663, 248-253

[13] Wanhill, R. J. H., & Ashok, B. (2017) Shape memory alloys (SMAs) for aerospace applications. Aerospace Materials and Material Technologies, 467-481

[14] Belyaev, S., Volkov, A., & Resnina, N. (2014) Alternate stresses and temperature variation as factors of influence of ultrasonic vibration on mechanical and functional properties of shape memory alloys. Ultrasonics, 54(1), 84-89

[15] Klubovich, V. V., Rubanik, V. V., Likhachev, V. A., Rubanik Jr., V. V. Dorodeiko, V. G. (1997) Generation of shape memory effect in Ti-Ni alloy by means of ultrasound, in A. Pelton, D. Hodgson, S. Russel, T. Duering (Eds.), Proc of the 2nd International Conference on Shape Memory and Superelastic Technologies, SMST, pp. 59-64

[16] Rubanik Jr., V. V. Rubanik, V. V. Klubovich, V. V. (2008) The influence of ultrasound on the shape memory behavior, MSE A 481-482 620-622

[17] Buchelnikov, V., Pikshtein, I., Grechishin, R., Khudoverdyan, T., Koledov, V., Kuzavko, Y., Nazarkin, I., Shavrov, V. Takagi, T. (2004) Ultrasound-induced martensitic transition in ferromagnetic Ni2.15Mn0.81Fe0.04Ga shape memory alloy, J. Magn. Mater. 272-276 2025-2026

[18] Steckmann, H., Kolomytsev, V. I., Kozlov, A. V. Acoustoplastic effect in the shape memory alloy Ni–Ti–Re at ultrasonic frequency, Ultrasonics 37 (1999) 59-62

[19] Breczko, T., Rubanik, V. V., Rubanik Jr., V. V. (1999) Study of behaviour NiTi-alloy under action of ultrasound, Proc. of SPIE, NDTCS-98 3687 310-312

[20] Bao, M., Zhou, Q., Dong, W., Lou, X., Zhang, Y. (2013) Ultrasound-modulated shape memory and payload release effects in a biodegradable cylindrical rod made of chitosan-functionalized PLGA microsphere, BioMac

[21] Goliboroda, I., Rusinko, K., & Tanaka, K. (1999) Description of a Fe-based shape memory alloy thermomechanical behavior in terms of the synthetic model. Computational materials science, 13(4), 218-226

[22] Rusynko, K. M., & Shandrivskyi, A. H. (1996) Irreversible deformation in the course of a martensite transformation. Materials Science, 31(6), 786-789

[23] Rusinko, A., & Rusinko, K. (2011) Plasticity and creep of metals. Springer Science & Business Media

[24] Likhachov, V. A. and Malinin, V. G. (1993) Structural-Analytic Theory of Elasticity [in Russian] Nauka, St. Petersburg

[25] https://matthey.com/products-and-markets/other-markets/medical-components/resource-library/nitinol-technical-properties

[26] Bradley, D. (1965) Sound Propalation in Near-Stoichiometric Ti-Ni Alloys. The Journal of the Acoustical Society of America, 37(4), 700-702

Simulation of the Position Control of PAM, Using Its' Nonlinear Dynamic Model

István Bíró[1], Monika Trojanová[2], Tomáš Čakurda[2], Zoltán Fabulya[3] and József Sárosi[4]

Department of Mechanical Engineering, Faculty of Engineering, University of Szeged, Mars tér 7, 6724 Szeged, Hungary, biro-i@mk.u-szeged.hu

Department of Industrial Engineering and Informatics, Faculty of Manufacturing Technologies with a Seat in Prešov, Technical University of Košice, 1 Bayerova, 080 01 Prešov, Slovakia; monika.trojanova@tuke.sk; tomas.cakurda@tuke.sk

Department of Engineering Management and Economics, Faculty of Engineering, University of Szeged, Mars tér 7, 6724 Szeged, Hungary, fabulya@mk.u-szeged.hu

Department of Mechatronics and Automation, Faculty of Engineering, University of Szeged, Moszkvai krt. 9, 6725 Szeged, Hungary, sarosi@mk.u-szeged.hu

Abstract: The paper investigates the mathematical, dynamic nonlinear model, of a single Pneumatic Artificial Muscle (PAM). A Fluid muscle from FESTO, is connected vertically as an oscillating non-linear system with 1 degree of freedom. When the muscle is filled with a working medium - compressed air, there is a change in the pressure over time and muscle parameters (change in length, diameter, fiber braiding angle, mass position, etc.). The simulation of the position of mass is the goal of the research described in this article. A mathematical model of the system, considers the change in the spring stiffness and the damping factor. At the same time, the model considers the effect of time pressure changes, on system oscillation. The article provides a few examples, which help specify the positioning model (also identifying values of individual coefficients of quantities occurring in the model).

Keywords: pneumatic artificial muscle (PAM); nonlinear dynamic model; oscillator; positioning; mathematical model

1 Introduction

For robots and manipulators, it is necessary to note several attributes, that should be ensured, for proper functioning: manipulation, automation, programmability, universality, feedback, and others. The importance of the individual properties results from their use. Properties are influenced by the elements, such as design,

actuators, sensors, and control [1]. In addition to conventional types of robotic actuators (e.g., electric or hydraulic), many researchers have also focused on unconventional types, e.g., pneumatic artificial muscle (see. [2-6]). It is a device that consists of two layers and endings, whose task is to generate tensile force after filling the muscle with the working media. Around 1930, the first researcher of the Pneumatic Artificial Muscle (Garasiev) focused mainly on constructing the muscle. His muscle, however, encountered the problem of inadequate material technology, whereby the fibers were unable to produce sufficient elasticity. At present, the most typical representative of this type of actuator - McKibben's artificial muscle, is characterized by softness, flexibility, easy maintenance, and possible use in demanding operating conditions, which can be a significant benefit also for industrial applications. Due to the friction between fibers themselves or between the fibers and tube, its behavior is nonlinear and hysteretic. This deficiency was addressed by FESTO [7], which combined the inner and outer layers in one, thereby prolonging the lifetime of the muscle compared to other muscles and improving its performance. Although the manufacturer has eliminated some undesirable properties, it is still necessary to consider the occurring hysteresis and nonlinearity of the characteristics that make the identification process more difficult.

Identification allows analyzing relationships and processes within the system and creates a model representing this system. Proper positioning of PAM-powered systems is also subject to the process of modeling pneumatic artificial muscles, which is the subject of the research described not only in this article but also by other researchers in the field. Studies by experts in PAM include, e.g., studies of the mechanical properties of PAM described in [8]. These are experimental models describing the mechanics of the behavior of PAM, where the experiment is performed at three different muscle sizes and different loads. Another study describes progress in PAM research [9]. The authors summarized an overview of artificial muscles, artificial muscle division, application of artificial muscles, but at the same time described a study of the control of the artificial muscle algorithm (similar as in the introduction of reference [10]).

Some approaches are currently known for modeling pneumatic artificial muscles. E.g., in [11] is an overview of the commonly used techniques to modeling pneumatic artificial muscles. The most used approach to modeling is experimental. It was used, e.g., in [8, 12-16]. In [17] is experimentally characterize the behavior of McKibben's muscles. The model can predict muscle displacement and force parameters under static conditions. Comparing the measured data with the model showed the model's ability to predict static characteristics. Similarly, [18] describing a 2-muscle FESTO test device. Authors also describe methods used to determine muscle characteristics under static conditions (isobaric, isometric, and isotonic). The measured data from the experimental device were used for the analytical determination of equations of characteristics, which were subsequently used for the simulation of characteristics. An experimental modeling approach is not always practical. In that case, is chosen the analytical method of finding the

solution (especially searching for the analytical model based on the measured values). In [19] is an analytical static model based on Newton mechanics, comparing the model results with the measured data. One of the latest publications in analytical modeling of systems powered by PAMs is [20]. It describes a nonlinear numerical approximation of non-linear static characteristics of the McKibben PAM (namely the dependence of force on deformation and pressure), the output of which are four mutually compared functions. Models and publications that describe the system's dynamics are more time-consuming in terms of research and preparation. The reason is perhaps the occurrence of non-linear characteristics and the higher number of elements in the system. This fact complains about the process of identifying the model under dynamic conditions. However, [21] describes a dynamic model for a system with 2 degrees of freedom, where the nonlinearity of the force-dependent on muscle pressure and deformation is investigated. The force parameter was obtained analytically, followed by experimental testing. [22] describes the creation of dynamic models whose role is to accurately predict the dynamic behavior of muscles. The dynamic model is simulated using the Matlab, and the pressure, contraction, and muscle temperature are tested experimentally. The most recent article describing a 1 degree of freedom system, to derive a motion equation from studying resonant conditions is [23]. This paper derives a nonlinear motion control equation concerning muscle strength as a function of pressure (both static and time-variable).

The monitoring of hysteresis and nonlinearity of the characteristics of pneumatic artificial muscles plays an essential role in achieving the highest possible level of system identification and, consequently, control. E.g., in [24] is described motion control of manipulator that have 2 degrees of freedom. The manipulator is parallel and uses 3 PAMs. Because the authors realize the relationship between muscle pressure and contraction rate, they included a hysteresis characteristic in their research. Therefore, they tried to design a control system with integrated hysteresis compensation. Ref. [25] describes the dependence of friction by speed, the dependence of coefficient friction by pressure, and hysteresis itself. Also, [26] describes a static force model for a hysteresis loop in a force-contraction curve, and it has also been experimentally verified. Modeling and hysteresis compensation is also research in [27], where a generalized Bouc-Wen model is applied for dynamic modeling of hysteresis. Finally, in the literature [28], the authors present new equations for modeling the length of contraction using the sigmoid shape of the characteristics.

There is an enormous amount of applied research where pneumatic artificial muscles are integrated into the purpose-built device. Their application is diverse (from industry in the medical field of prostheses to applications in everyday life). E.g., authors in [29] describe the application of flexible fluidic actuators to grippers or arms. Another applied research is the project, which in [30] presents the design and construction of an electro-pneumatic parallel manipulator with 3 degrees of freedom allows to pick up objects in recycled waste facilities. The last applied

research is active bridge vibration control, which is stretched by pneumatic artificial muscles [31]. The task of the researchers was first to derive a discrete analytical model, which was subsequently verified by non-feedback experiments. Then, based on the model, a feedback control model was designed.

The research described in the article aims to use a nonlinear dynamic model in simulations of position control of the pneumatic artificial muscle. The model can respond to spring stiffness and damping factor changes depending on the muscle contraction (eliminating muscle mass oscillations). A fluidic muscle DMSP-20-400 N with the length of 400 mm and diameter of 20 mm was therefore chosen for the experiments. The individual parts of the article contain the process of derivation of the mathematical model of the system, which considers the change in muscle stiffness and contraction-dependent damping, is described. Finally, the research results are presented on examples of the muscle reaching a specific state at selected coefficients that influence the resulting oscillation.

2 Nonlinear Dynamic Model of the Pneumatic Artificial Muscle

Suppose the pneumatic artificial muscle is connected so that one end is firmly attached and a load with a certain weight m is placed at the other end. In that case, if the muscle is filled with a working medium, a contraction will occur, and the load will be pulled upwards. The muscle thus has one degree of freedom. The pneumatic artificial muscle can accordingly be modeled as an oscillator. The vibration model is shown in Figure 1 (m is mass of the load, x is displacement, g is the acceleration of gravity, c is coefficient of damping, k is stiffness, and F_{spring} is contractile force developed by PAM).

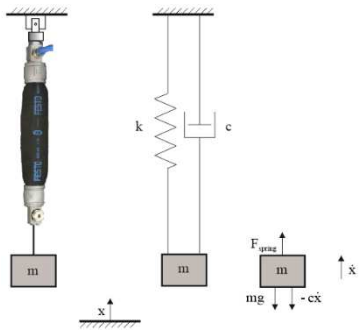


Figure 1

Oscillator formed by a pneumatic artificial muscle and one degree of freedom

In the spring-mass system, where damping is also present, vibrations occur due to a non-linear force, and at the same time, a gravitational force is present in a

vertically downward direction. Its vibration can be described by the following second-order differential equation:

$$m\ddot{x} = F_{spring} - mg - c\dot{x} \tag{1}$$

2.1. Contractile Force F_{spring}

The dimension of the contractile force produced in PAM depends on the parameters of the muscle, namely: the material from which the inner and outer muscle layers are made; geometric parameters of the fibers of these layers; and the pressure of the working medium in the muscle – the applied pressure of the air.

The contractile force is closely related to an important muscle property - contraction. Muscle contraction is the relative displacement with initial length l_0 of PAM and length l when the muscle is contracted. The equation for calculating muscle contraction is shown in Figure 2, which also indicates muscle parameters. While the muscle contraction is changing the length of the muscle, the angle of the muscle braid also changes from α_0 to α and the diameter of the muscle from d_o *to d.*

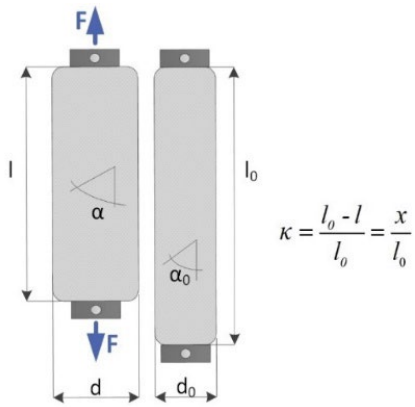


Figure 2
Contraction of pneumatic artificial muscle and its parameters

According to [21], F_{spring} contractile force can be described by an equation containing six unknown constants (a_1, a_2, a_3, a_4, a_5, and a_6) as follows:

$$F_{spring}(p, \kappa) = (a_1 \cdot p + a_2) \cdot exp^{a_3 \cdot \kappa} + a_4 \cdot \kappa \cdot p + a_5 \cdot p + a_6 \tag{2}$$

Where p is the gauge (applied), pressure and κ is a contraction.

It has to be noted that the damping coefficient depends on displacement for this reason, (1) can be rewritten into the following form:

$$m\ddot{x} = F_{spring}\big(\kappa(x)\big) - mg - c(\kappa(x))\dot{x} \tag{3}$$

The damping coefficient represents a nonlinear function of the displacement x. Therefore, to get values of $c(\kappa(x))$, it is needed to know the hysteresis based on force-contraction curves. Hysteresis occurs due to friction which occurs when a muscle contracts between individual muscle layers, layer fibers, or when muscle ends and layers touch.

2.2. Hysteresis of Force-Contraction Curves on Constant Pressures

In Figure 3, the scheme of the testbed for pressure, force, and position measuring is shown that consists of the following sensors and regulators: Motorola MPX5999D pressure sensor, LINIMIK MSA 320 incremental encoder, Kaliber 7923 load cell, and Festo VPPM-6L-L-1-G1/8-0L6H-V1N-S1C1 proportional pressure regulator (PPR). Acquisition and monitoring of measured data are developed in LabVIEW.

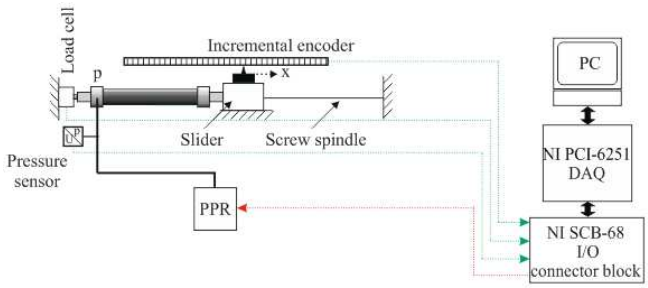


Figure 3

Experimental stand for measuring pressure, force, and position

Fluidic muscle type DMSP-20-400N is built into the rig. One of the ends of the muscle is fixed while the other is movable. The free end is connected to an incremental encoder (with a resolution of 0.01 mm) by a spindle. The list of technical parameters of fluidic muscle used in research is in the following table:

Table 1

List of technical parameters of DMSP 20-400N Festo fluidic muscle

Parameter	Value
Inside diameter	20 mm
Nominal length	400 mm
Maximal lifting force	1 500 N
Maximal operating pressure	600 kPa
Maximal permissible contraction	25 % of nominal length
Maximal permissible pretensioning	4 % of nominal length
Ideal temperature of the environment	from -5 °C to 60 °C

Firstly, the force-contraction function for various compressions is measured. The applied pressure is kept at a constant value while the screw spindle changes the length of PAM. Measuring at every point is repeated five times and statistically averaged. The hysteresis loop in the force-contraction curve is obtained when the procedure in the opposite direction is repeated. Approaching of experimentally measured hysteresis curves by Equation (2) is demonstrated in Figure 4. The figure shows that the function of static force and contraction depends on pressure. The maximum static force is reached by increasing the length of PAM. The zero-force is achieved at the minimum length of the muscle, i.e. at its maximum contraction. The muscle length is expressed in the figure in three basic positions. Nominal length - at zero muscle contraction. Minimal length - at maximum muscle contraction. Total length - presented as a negative contraction on the graph. Muscles from FESTO are made of durable but also flexible materials, which allow the muscle to have parameters such as Maximal permissible pretensioning and Maximal permissible contraction (it is possible to observe pre-tensioning in the muscle, based on which the contraction appears to be negative in the graph).

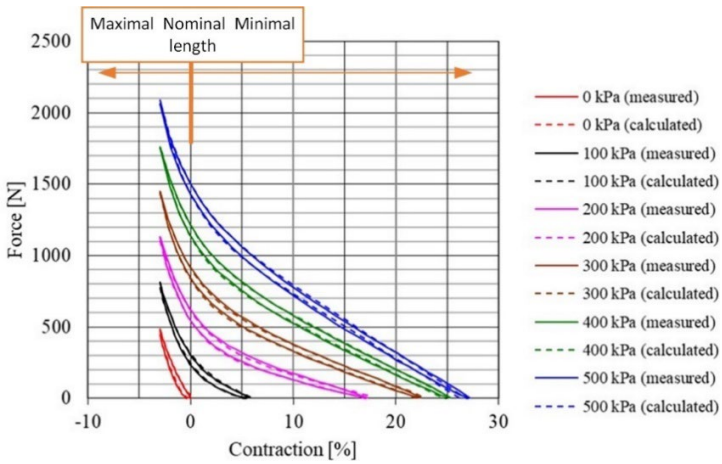


Figure 4

Hysteresis of force-contraction curves on constant pressures (measured and calculated by equation (2))

Figure 4 graphically presents the presence of hysteresis in the investigated system (PAM); therefore, it is necessary to optimize the branches of hysteresis based on Equation 2 (optimization of individual constants of Equation 2). Tables 2 and 3 contain results of optimization concerning the upper and lower branches of hysteresis. For upper branch $R^2 = 0.9995$ ($R = 0.9997$) as well as for the lower one $R^2 = 0.9991$ ($R = 0.9995$) accuracy are obtained.

Table 2

Parameters a_1, a_2, a_3, a_4, a_5, and a_6 of (2) for an upper branch of the hysteresis curve

Parameters	a_1	a_2	a_3	a_4	a_5	a_6
Values	-4.35573	281.22370	-0.32866	-9.27035	302.20107	-263.69185

Table 3
Parameters a_1, a_2, a_3, a_4, a_5, and a_6 of (2) for a lower branch of the hysteresis curve

Parameters	a_1	a_2	a_3	a_4	a_5	a_6
Values	3.65565	232.27984	-0.37855	-8.95540	290.68164	-277.50222

2.3. Damping Coefficient

The difference between one hysteresis surface (U_0 defines the area below the upper branch) and the other hysteresis surface (U_1 defines the area below the lower branch) is called the hysteresis surface ΔU ($\Delta U = U_0 - U_1$). If the hysteresis surface ΔU and the surface under the upper branch U_0 are added, this ratio gives Lehr's damping coefficient. However, Lehr's damping coefficient can also be expressed as the ratio of the damping coefficient and the critical damping coefficient.

Based on the above, Lehr's damping coefficient can be written as

$$\zeta = \frac{\Delta U}{U_0} = \frac{c}{c_{cr}} \tag{4}$$

Subsequently, the critical damping coefficient c_{cr} can be given as

$$c_{cr} = 2 \cdot \sqrt{k \cdot m} \tag{5}$$

Where the stiffness k for constant pressure can be derived from (2) as:

$$k = k(\kappa) = \frac{dF(l)}{dl} = \frac{dF(\kappa)}{l_0 \cdot d\kappa} = \frac{1}{l_0} \cdot \frac{dF(\kappa)}{d\kappa} = \frac{1}{l_0} \cdot \frac{dF_{spring}(p,\kappa)}{d\kappa} =$$

$$= \frac{1}{l_0} \cdot \frac{d\left((a_1 \cdot p + a_2) \cdot exp^{a_3 \cdot \kappa} + a_4 \cdot \kappa \cdot p + a_5 \cdot p + a_6\right)}{d\kappa} = \frac{(a_1 \cdot p + a_2) \cdot a_3 \cdot exp^{a_3 \cdot \kappa} + a_4 \cdot p}{l_0} \tag{6}$$

The damping coefficient c can be given as

$$c = \zeta \cdot c_{cr} = \zeta \cdot 2 \cdot \sqrt{k \cdot m} = \zeta \cdot 2 \cdot \sqrt{\frac{m}{l_0} \cdot [(a_1 \cdot p + a_2) \cdot a_3 \cdot exp^{a_3 \cdot \kappa} + a_4 \cdot p]} \tag{7}$$

2.4. Dynamic Properties of PAM

Based on Figure 1 and (1), moving a load with a mass of 42 kg for a displacement of 72.0 mm at a pressure of 500 kPa, acceleration, velocity, displacement, and F_{spring} force as a function of time for the 1-DOF oscillatory system (described in Figure 1) can be seen in Figures 5-8. The functions were determined and created in Simulink.

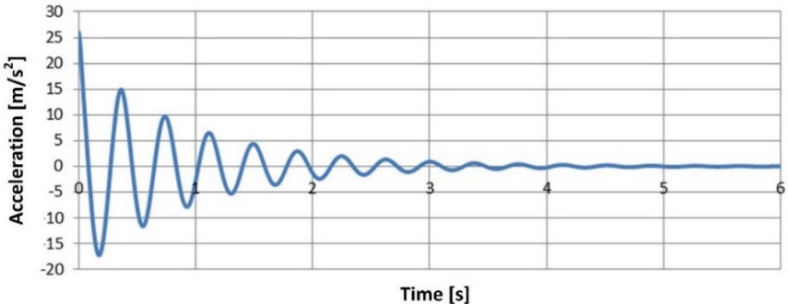


Figure 5
The acceleration-time function of the one-degree-of-freedom oscillatory system (pressure of 500 kPa, load of 42 kg, and displacement of 72.0 mm)

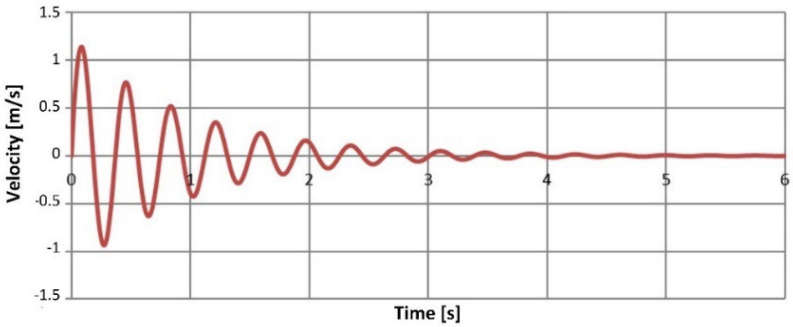


Figure 6
The velocity-time function of the one-degree-of-freedom oscillatory system (pressure of 500 kPa, load of 42 kg, and displacement of 72.0 mm)

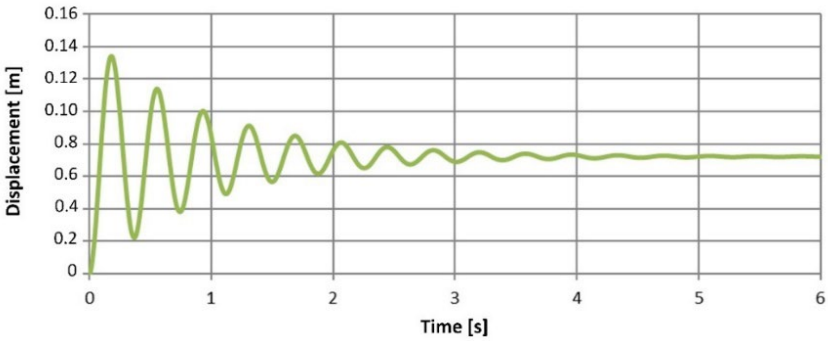


Figure 7
The displacement-time function of the one-degree-of-freedom oscillatory system (pressure of 500 kPa, load of 42 kg, and displacement of 72.0 mm)

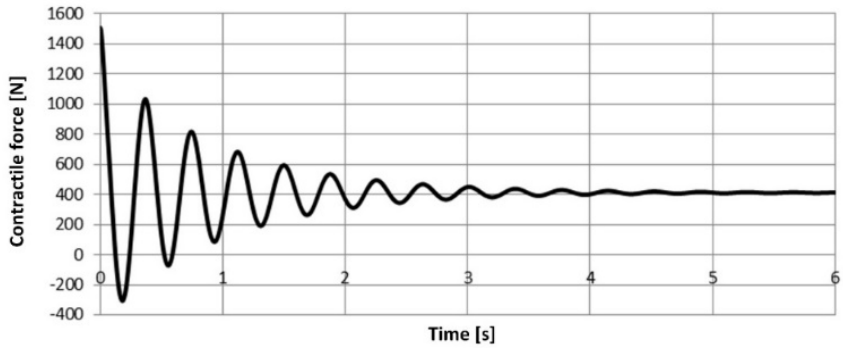


Figure 8
F_{spring} force-time function of the one-degree-of-freedom oscillatory system (pressure of 500 kPa, load of 42 kg, and displacement of 72.0 mm)

3 Nonlinear Dynamic Model of the Pneumatic Artificial Muscle Excited by Time-Varying Pressure

In the previous chapter of the article, Equation 3 was considered with respect to constant pressure. However, in case of taking into consideration the effect of variable pressure, the motion Equation (3) of the oscillating system in an upright position can be written down in the modified form:

$$m\ddot{x} - F_{spring}[\kappa(x), p(t)] + c[\kappa(x), p(t)]\dot{x} + mg = 0 \quad \text{or…}$$

$$m\ddot{x} - F_{spring}[\kappa(x), p((x(t))] + c[\kappa(x), p(x(t))]\dot{x} + mg = 0 \tag{8}$$

The second-order nonlinear differential Equation (8) can be solved solely in a numerical way. Applied algorithms to the initial value problem in MS Excel can be seen in Table 4 below.

Table 4
Applied algorithms in MS Excel for solving a differential equation

t	$\ddot{x}(\dot{x}, x)$	$\dot{x}$	x
t_o	$\ddot{x}_o(\dot{x}_o, x_o)$	$\dot{x}_o$	x_o
t_1	$\ddot{x}_1(\dot{x}_1, x_1)$	$\dot{x}_1 = \dot{x}_o + \ddot{x}_o(t_1 - t_o)$	$x_1 = x_o + \dot{x}_1(t_1 - t_o)$
t_2	$\ddot{x}_2(\dot{x}_2, x_2)$	$\dot{x}_2 = \dot{x}_1 + \ddot{x}_1(t_2 - t_1)$	$x_2 = x_1 + \dot{x}_2(t_2 - t_1)$
t_3	….	….	….
t_4	….	….	….

For example, there is a task to lift a mass m = 50 kg and keep its position on a level of 50 mm by applying a pneumatic artificial muscle. Then, at a pressure of 160 kPa, the motion in the vertical direction starts. According to Figure 9, the pressure in PAM has changed uniformly from 160 kPa to 400 kPa. Therefore, needed times in the investigated versions are 0 s, 0.1 s, 0.2 s, 0.3 s, and 0.4 s, respectively.

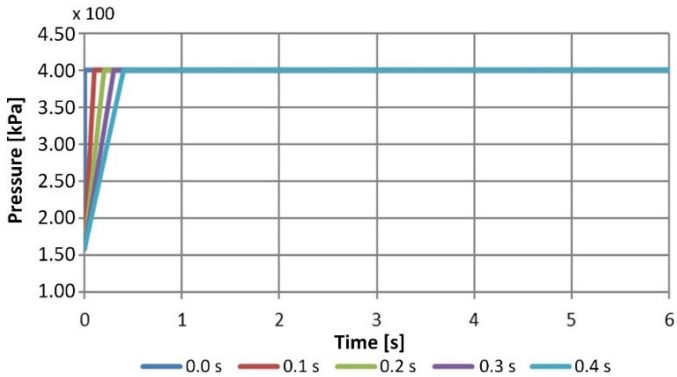


Figure 9
p(t) diagrams of PAM in case of different intensity of pressure ratio

In Figure 10, *x(t),* diagrams of PAM can be seen. In each case, the mass reaches the level of 50 mm (accurate value: 49.54 mm). To achieve an expected precise position within 0.01 mm, pressure control is needed. Seeing Figure 10, it has to be noticed that there is a relationship between the intensity of pressure ratio and amplitudes of oscillating mass.

In figure 11, $F_{spring}(t)$ diagrams of PAM in the case of different intensity of pressure ratio can be seen. After damping of oscillation, the spring force in PAM equals the weight of the mass (accurate value: 490.5 N). It is important to notice that by increasing the interval of increase of pressure – similarly to change of amplitudes - the change of spring force in PAM decreases significantly.

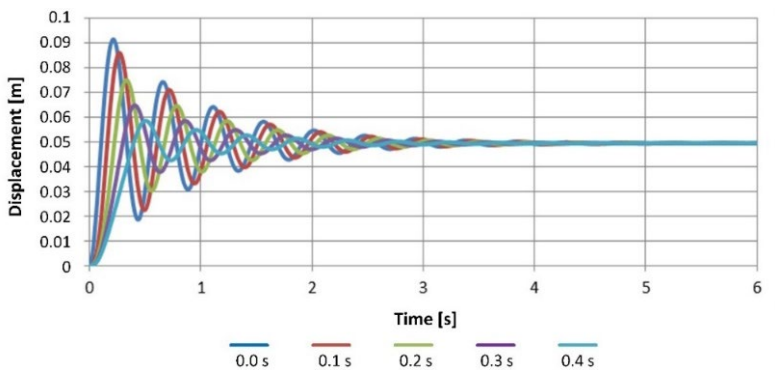


Figure 10
x(t) diagrams of PAM in case of different intensity of pressure ratio

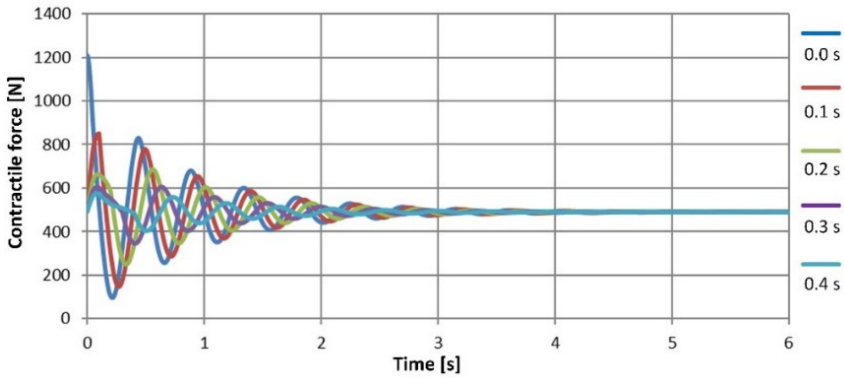


Figure 11

$F_{spring}(t)$ diagrams of PAM in case of different intensity of pressure ratio

4 Simulation of the PAM Dynamics Using the Nonlinear Dynamic Model

Mathematic model (8) of PAM is suitable for positioning by pressure control: initial pressure p_o is adjustable arbitrarily (proposed interval: 0-600 kPa). The change of pressure is controlled by the difference of actual and demanded the position of mass:

$$p_{n+1} = p_n + b(x_{dem} - x) \tag{9}$$

In Equation (9), p_n and p_{n+1} are pressures in PAM after nth, respectively n+1-th time step, b is the intensity factor of pressure ratio, x_{dem} is the demanded position. Finally, x is the actual position of the mass point after nth time step.

3.1. First Task

From position x= 0 mm to lift 10 kg mass by PAM to the demanded level of 40 mm. The initial pressure is p_o= 10^5 Pa; the value intensity factor of pressure ratio is b=0.01. In Figure 12-13, the pressure in PAM and the mass position in the function of time can be seen in a time interval of 0 to 6 seconds.

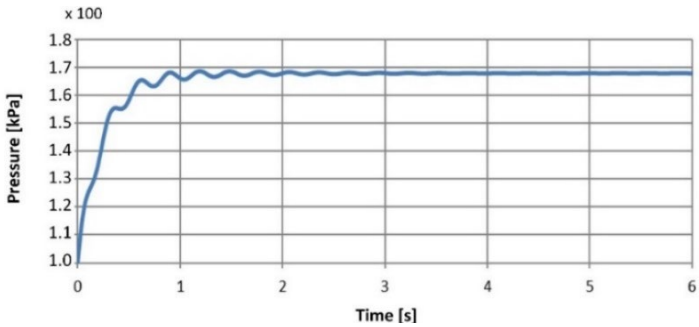


Figure 12
Pressure in PAM in the function of time, position changes from 0 mm to 40 mm

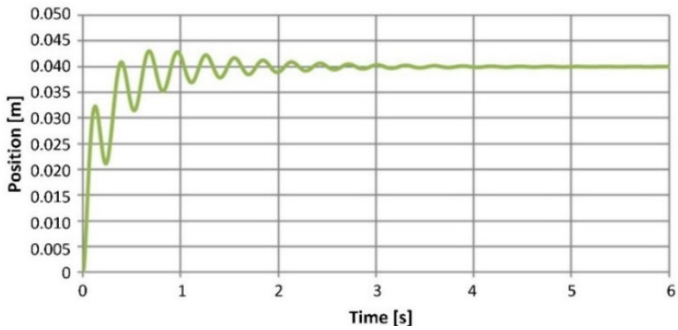


Figure 13
Position of mass in the function of time, position changes from 0 mm to 40 mm

3.2. Second Task

The mass point starts from position x= 40 mm. Its demanded level of 30 mm (Figure 14-15). Initial pressure p_o=1.678 x 100 kPa is needed to keep it at the initial position. The value of the intensity factor of pressure ratio is b=0.012.

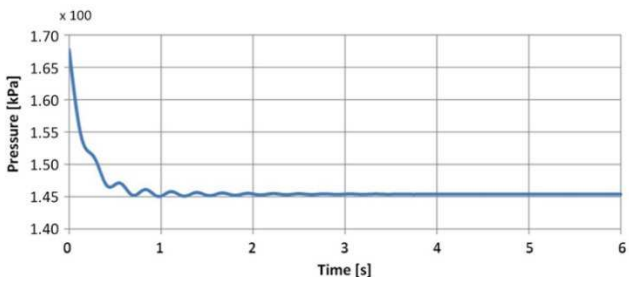


Figure 14
Pressure in PAM in the function of time, position changes from 40 mm to 30 mm

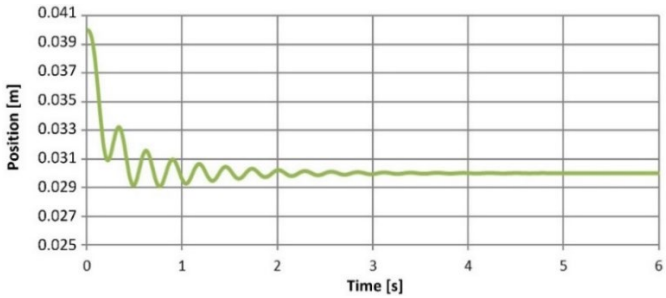


Figure 15
Position of mass in the function of time, position changes from 40 mm to 30 mm

3.3. Third Task

Finally, 25 kg mass starts from position x=0 mm. Its demanded level of 50 mm (Figure 16-17). Initial pressure p_o=2 x 100 kPa. The value of the intensity factor of pressure ratio is b=0.005.

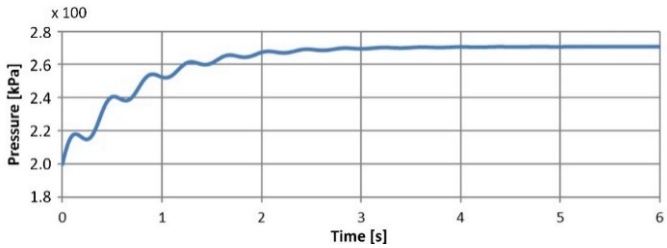


Figure 16
Pressure in PAM in the function of time, position changes from 0 mm to 50 mm

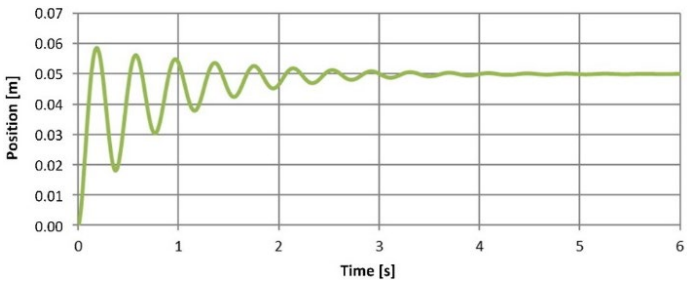


Figure 17
Position of mass in the function of time, position changes from 0 mm to 50 mm

3.4. Subsequent Tasks

To investigate the accuracy of this method, let us study further examples: the pressure to keep 10 kg mass at the level of 50 mm exactly is 1.91768913274509 x 100 kPa. The initial pressure is a little bit lower: p_o=1.91768913274 x 100 kPa (Figure 18). For this reason, the mass point starts downward and, after damping oscillation - as a result of pressure control - stops precisely at the demanded level of 50 mm. The initial and final positions are the same. The temporary disorder is caused by intentionally modified initial pressure.

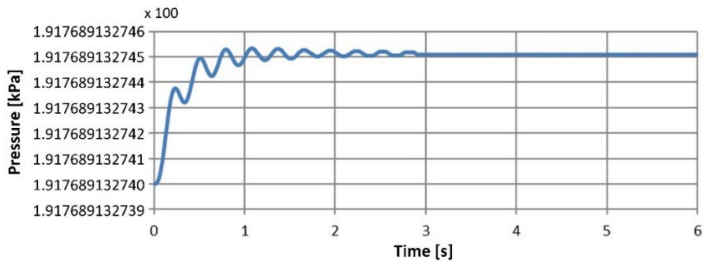


Figure 18
Pressure in PAM in the function of time, initial and final positions are 50 mm

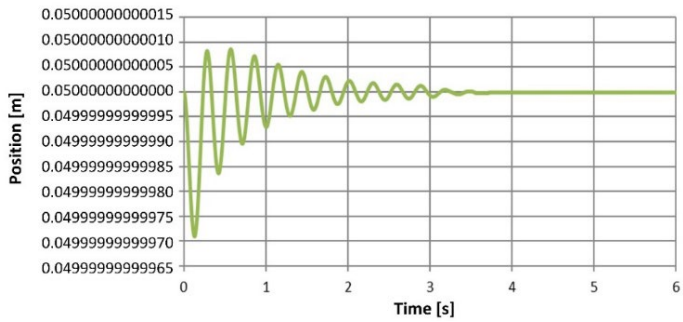


Figure 19
Position of mass in the function of time, initial and final positions are 50 mm

The role of *b* intensity factor of pressure ratio is quite important. Its determination happens empirically in the example above *b*=0.01. If it is modified to 0.005, diagrams are also limited as a result (Figure 20-21). The pressure increase is slower, and the final position is slightly under the one demanded.

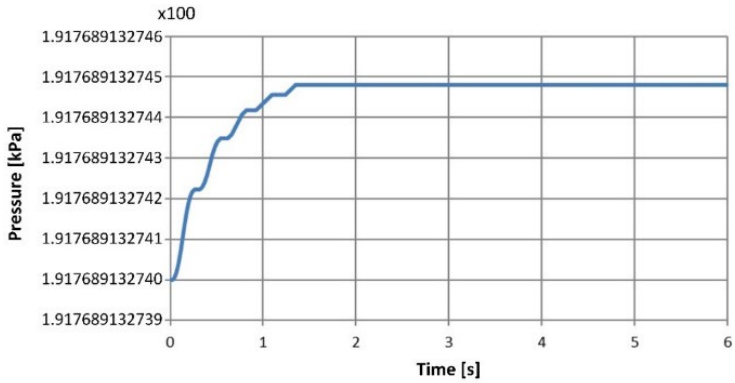


Figure 20

Pressure in PAM in the function of time, the intensity factor of pressure ratio *b*=0.01

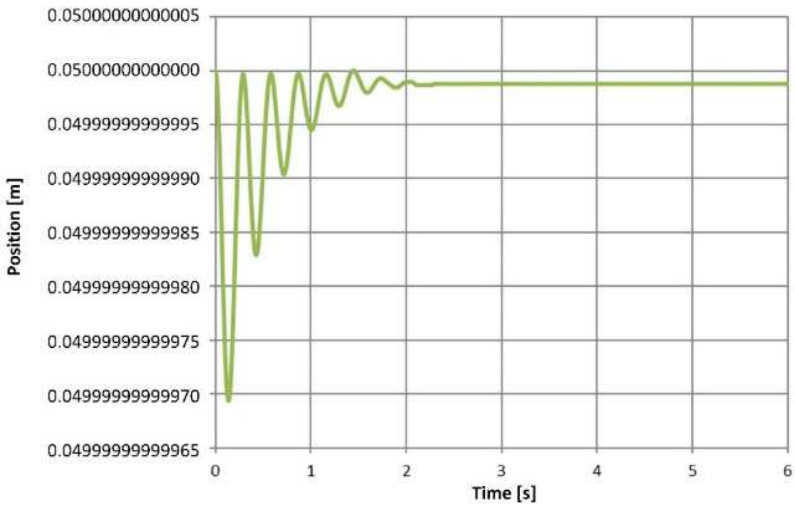


Figure 21

Position of mass in the function of time, the intensity factor of pressure ratio *b*=0.01

Forms of curves are significantly different when the intensity factor of pressure ratio is *b*=0.02. In this case, the stimulating effect of pressure changes is stronger than the internal damping action of PAM. Because the pressure change is controlled by the position change of the mass point, it is excited on the natural frequency. Despite damping, the result is similar to phenomena of resonance, but in this case, the amplitude is getting bigger and bigger, exceeding the usual linear connection (Figure 22-23).

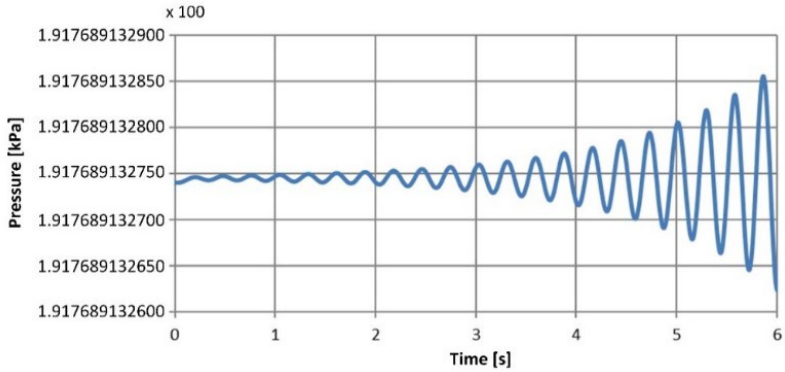


Figure 22
Pressure in PAM in the function of time, the intensity factor of pressure ratio b=0.02

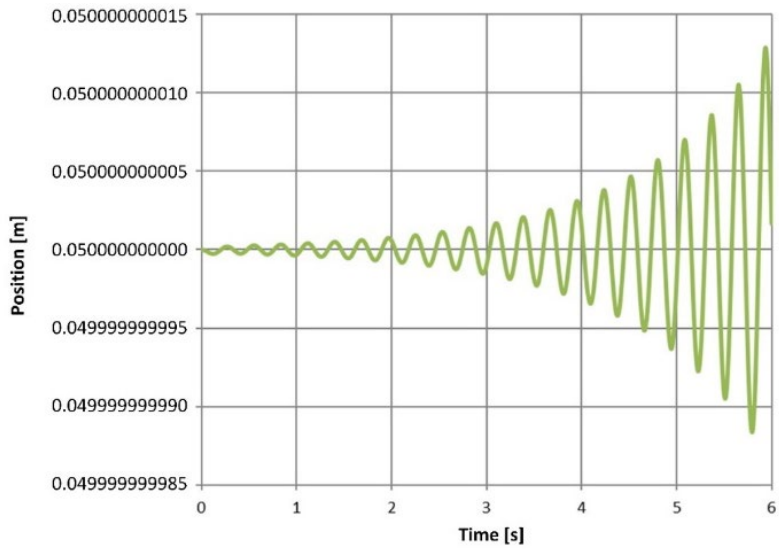


Figure 23
Position of mass in the function of time, the intensity factor of pressure ratio b=0.02

There is a border between the two cases mentioned above when the intensity factor of pressure ratio is b=0.01585. Therefore, the task is to lift a mass m=10 kg from level 49 mm and keep its position on the level of 50 mm. In this case, the amplitudes of excited-damped oscillation in the function of time are the same. The initial pressure p_o=1.91768732075 x 100 kPa is needed to keep the mass point at the level of 50 mm (Figure 24-25).

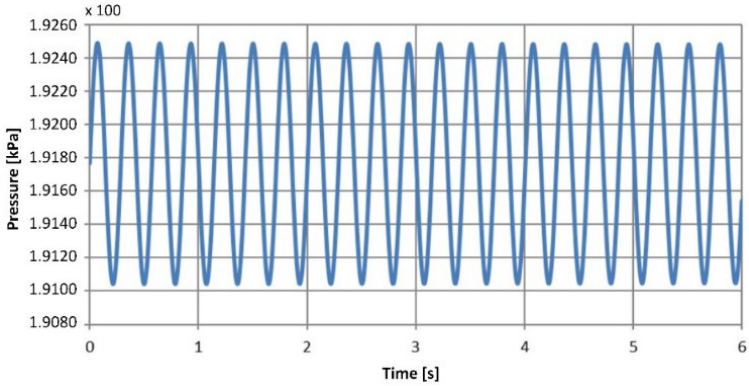


Figure 24
Pressure in PAM in the function of time, the intensity factor of pressure ratio b=0.01585

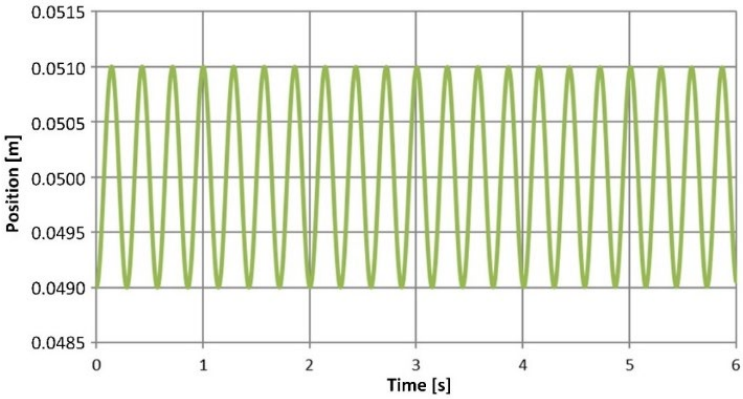


Figure 25
Position of mass in the function of time, the intensity factor of pressure ratio b=0.01585

Conclusions

The presented mathematical model of the vertically mounted nonlinear oscillating system (Figure 1), consisting of mass and excited by in time-variable pressure elastic pneumatic artificial muscle, can take into consideration the change of spring stiffness and damping factor, in the function of contraction, as well as the effect on the oscillation of in time-variable pressure. The value of the pressure and the intensity of its change, significantly affects the dynamic behavior of the weakly damped oscillating system. This mathematical model is suitable for precise and quick positioning, without oscillating different mass points using pressure control. During the pressure control, the following quantities are considered: Position difference of mass points between actual and those demanded, velocity, acceleration and jerk of the mass.

Acknowledgement

The project was supported by the projects VEGA no. 1/0700/20 Object Recognition Techniques with Convolutional Neural Networks and KEGA no. 055TUKE-4/2020 Transfer of knowledge from research on digitization of production processes into study programs of the Faculty of Manufacturing Technologies.

References

[1] Siciliano, B. and Khatib, O.: Springer Handbook of Robotics, *Springer*, 2016, p. 2259

[2] Zhao, J. et al.: Position Control of a Pneumatic Muscle Actuator Using RBF Neural Network Tuned PID Controller, *Mathematical Problems in Engineering*, 2015, pp. 1-16

[3] Ahn, K. K. and Anh, H. P. H.: Comparative Study of Modeling and Identification of the Pneumatic Artificial Muscle (PAM) Manipulator Using Recurrent Neural Networks, *Journal of Mechanical Science Technology*, 2008, Vol. 22, No. 7, pp. 1287-1298

[4] Sárosi, J. et al.: Static Force Model-Based Stiffness Model for Pneumatic Muscle Actuators, *International Journal of Engineering Research in Africa*, 2015, Vol. 18, pp. 207-214

[5] Pitel', J. and Tóthová, M.: Dynamic modeling of PAM based actuator using modified Hill's muscle model, *In Proceedings of the 14th International Carpathian Control Conference (ICCC)*, Poland, May 2013, pp. 307-310

[6] Kang, B.-S. et al.: Dynamic Modeling of Mckibben Pneumatic Artificial Muscles for Antagonistic Actuation, *2009 IEEE International Conference on Robotics and Automation*, Japan, May 2009, pp. 182-187

[7] FESTO Fluidic Muscle DMSP/MAS Available online: https://www.festo.com/rep/en_corp/assets/pdf/info_501_en.pdf (accessed on 10 November 2019.

[8] Wickramatunge, K. C. and Leephakpreeda, T.: Study on Mechanical Behaviors of Pneumatic Artificial Muscle, *International Journal of Engineering Science,* 2010, Vol. 48, No. 2, pp. 188-198

[9] Luo, Z. et al.: Advances in Research on Artificial Muscles Technology and Its Control Algorithm, *In Proceedings of the 2010 2nd International Conference on Advanced Computer Control*, China, March 2010, Vol. 3, pp. 48-51

[10] Davis, S. et al.: Enhanced Modelling and Performance in Braided Pneumatic Muscle Actuators, *I. J. Robotic Res.,* 2003, Vol. 22, pp. 213-228

[11] Kelasidi, E. et al.: A Survey on Pneumatic Muscle Actuators Modeling, *In Proceedings of the 2011 IEEE International Symposium on Industrial Electronics*, Poland, June 2011, pp. 1263-1269

[12] Hošovský, A. et al.: Dynamic Characterization and Simulation of Two-Link Soft Robot Arm with Pneumatic Muscles, *Mechanism and Machine Theory*, 2016, Vol. 103, pp. 98-116

[13] Hošovský, A. and Židek, K.: Experimental Validation of Nominal Model Characteristics for Pneumatic Muscle Actuator, *Applied Mechanics and Materials*, 2014, Vol. 460, pp. 1-12

[14] Wickramatunge, K.C. and Leephakpreeda, T.: Empirical Modeling of Pneumatic Artificial Muscle, *In Proceedings of the Proceedings of the International Multi Conference of Engineers and Computer Scientists (IMECS 2009)*, Hong Kong, March 2009, pp. 1726-1730

[15] Cao, J. et al.: A New Dynamic Modelling Algorithm for Pneumatic Muscle Actuators, *In Proceedings of the Intelligent Robotics and Applications,* 2014; pp. 432-440

[16] Zhang, Z. and Philen, M.: Pressurized Artificial Muscles, *Journal of Intelligent Material Systems and Structures,* 2012, Vol. 23, pp. 255-268

[17] Kothera, C. S. et al.: Experimental Characterization and Static Modeling of McKibben Actuators, *Journal of Mechanical Design*, 2009, Vol. 131, pp. 1-10

[18] Takosoglu, J. E. et al.: *Determining the Static Characteristics of Pneumatic Muscles, Measurement and Control*, 2016, Vol. 49, pp. 62-71

[19] Doumit, M. et al.: Analytical Modeling and Experimental Validation of the Braided Pneumatic Muscle, *IEEE Transactions on Robotics,* 2009, Vol. 25, No. 6, pp. 1282-1291

[20] Tóthová, M. et al.: Numerical Approximation of Static Characteristics of Mckibben Pneumatic Artificial Muscle, *International Journal of Mathematics and Computers in Simulation*, 2015, Vol. 9, pp. 228-233

[21] Sarosi, J. et al.: Dynamic Modeling of a Pneumatic Muscle Actuator with Two-Direction Motion, *Mechanism and Machine Theory*, 2015, Vol. 85, pp. 25-34

[22] Doumit, M. D. and Pardoel, S.: Dynamic Contraction Behaviour of Pneumatic Artificial Muscle, *Mechanical Systems and Signal Processing*, 2017, Vol. 91, pp. 93-110

[23] Kalita, B. and Dwivedy, S. K.: Nonlinear Dynamics of a Parametrically Excited Pneumatic Artificial Muscle (PAM) Actuator with Simultaneous Resonance Condition, *Mechanism and Machine Theory*, 2019, Vol. 135, pp. 281-297

[24] Ito, A. et al.: Motion Control of Parallel Manipulator Using Pneumatic Artificial Actuators, *In Proceedings of the 2010 IEEE International Conference on Robotics and Biomimetics*, China, December 2010, pp. 460-465

[25] Tondu, B.: Modelling of the McKibben Artificial Muscle: A Review, *Journal of Intelligent Material Systems and Structures,* 2012, Vol. 23, pp. 225-253

[26] Sárosi, J.: Elimination of the Hysteresis Effect of PAM Actuator: Modelling and Experimental Studies, *Tehnički vjesnik*, 2015, Vol. 22, No. 6, pp. 1489-1494

[27] Aschemann, H. and Schindele, D.: Comparison of Model-Based Approaches to the Compensation of Hysteresis in the Force Characteristic of Pneumatic Muscles, *IEEE Transactions on Industrial Electronics*, 2014, Vol. 61, No. 7, pp. 3620-3629

[28] Al-Ibadi, A. et al.: Valuable Experimental Model of Contraction Pneumatic Muscle Actuator, *In Proceedings of the 2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR)*, Poland, August 2016, pp. 744-749

[29] Gaiser, I. et al.: Compliant Robotics and Automation with Flexible Fluidic Actuators and Inflatable Structures, *In INTECH,* 2012, pp. 567-608

[30] Laski, P. A. et al.: Design of a 3-DOF Tripod Electro-Pneumatic Parallel Manipulator, *Robotics and Autonomous Systems,* 2015, Vol. 72, pp. 59-70

[31] Bleicher, A. et al.: Model-Based Design and Experimental Validation of Active Vibration Control for a Stress Ribbon Bridge Using Pneumatic Muscle Actuators, *Engineering Structures,* 2011, Vol. 33, No. 8, pp. 2237-2247

Further Evolution of Mortality Prediction with Ensemble-based Models on Hungarian Myocardial Infarction Registry

Péter Piros[1], Rita Fleiner[1], András Jánosi[2], Levente Kovács[1]

[1] Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary,
piros.peter@uni-obuda.hu, fleiner.rita@nik.uni-obuda.hu,
kovacs.levente@nik.uni-obuda.hu

[2] Gottsegen National Cardiovascular Center, Haller u 29, 1096 Budapest, Hungary,
andras.janosi@gokvi.hu

Abstract: In the current study, we present a new approach to predict 30-day and 1-year mortality of patients hospitalized with acute myocardial infarction. The dataset of this research is originated from Hungarian Myocardial Infarction Registry, a full, real-world, unfiltered database of myocardial infarctions from year 2014 to 2016 (n = 47,391). The new approach is based on ensembling and uses the prediction capability of different (already ensembled, in some cases) models like Random Forest, General Boosting Machine, Neural Network and Generalized Linear Model. We previously presented more different modelling techniques with the same target on the same dataset, and this new ensemble-based way of prediction proved to be the best among all the others. By numbers, this means 0.856 ROC AUC (area under the receiver operating characteristic curve) for the 30-day, and 0.839 ROC AUC for the 1-year mortality, both measured on validation datasets. We came to the conclusion that the combination of machine learning algorithms and regression models results the best performance in mortality prediction on the dataset of HUMIR.

Keywords: Myocardial infarction; Mortality prediction; Machine learning; Ensemble classifier

1 Introduction

In *Heart Disease and Stroke Statistics*, American Heart Association annually reports, that approximately every 40 seconds, an American will have an myocardial infarction (MI) - they did the same in the recent statistics titled *2022 Update* [1]. The estimated annual incidence of MI is 605,000 new attacks and 200,000 recurrent attacks, they reported. The overall prevalence for MI is 3.1% in US adults (>= 20

years of age). Males have a higher prevalence of MI than females for all age groups except 20 to 39 years of age. MI prevalence is 4.3% for males and 2.1% for females.

In addition, *Heart disease* (which can lead to myocardial infarction) is still at the first position of the ten leading cause of death, followed by cancer, unintentional injuries, chronic lower respiratory diseases, stroke and Alzheimer disease, respectively.

In the area with numbers like these, mortality prediction can and should play a very important role in the hand of physicians: with validated models, it becomes possible to select patients with high-risk of death and use this information in the process of treatment. Using new, real-life datasets to extract hidden information can lead to more effective treatment and prevention.

Reliable, high-quality datasets are mandatory to build and train any type of predictive model. Hungarian Myocardial Infarction Register (HUMIR) project was introduced in 2010, initially collected AMI information from five districts of Budapest and the county of Szabolcs-Szatmár-Bereg. In 2014, the Hungarian government selected it as the official myocardial database and obligated all hospitals in the territory of Hungary to report all MI-cases to HUMIR. In the recent years, around 15,000 new patients got registered per year and until December 2022, the 94 participating hospitals reported 157,724 cases in 142,439 patients.

The dataset of this research is originated from HUMIR, it is a full, real-world, unfiltered database of myocardial infarctions from year 2014 to 2016 (n = 47,391). The features of the dataset consists of three attribute groups: General information about the patient (Group 1), Previously reported diseases (Group 2) and Information about the pre- and in-hospital treatment (Group 3). All relevant features of each group will be discussed later in *Section 3*.

In previous researches we developed several machine-learning models based on Decision Tree, Neural Networks (NN), Logistic Regression, Random Forest (RF), and Generalized Boosted Models algorithms to predict 30-day and 1-year mortality on the same dataset. The results achieved with these methods were published in several conferences and papers.

The aim of the current paper is to provide an ensemble-based modelling technique in the field of mortality prediction which is trained on the introduced HUMIR-dataset and combines the predictive power of the constituent learning algorithms. The idea behind the approach that we are working on the same dataset both here and in related researches is the following: we are trying to establish an order in the list of different modelling techniques by keeping the dataset fixed and trying to maximize the prediction capability of each of our models. Here, we present the details and results of the ensemble-based technique, and, as a result, we can compare the prediction power of the different learning algorithms.

The rest of the paper is organized as follows. In Section 2, we summarize our previous efforts with different modelling algorithms, and quote some related research results from other authors. Section 3 discusses the tools and methods involved in the current study, especially the background of ensemble modelling. Then, in Section

4, we present the new results and compare them with all our formers'. Finally, we conclude with a summary.

2 Related Work

2.1 Previous Results

In one of the previous works, we overviewed [2] the health care registries in Europe and found that although several databases store information about patients and diseases, only a few exist that focus directly on myocardial events and treatments. Three European projects were investigated: Myocardial Ischaemia National Audit Project (MINAP) in England, Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated According to Recommended Therapies (SWEDEHEART) in Sweden and National Registry of Acute Myocardial Infarction in Switzerland (AMIS Plus) in Switzerland. Then, we stated that from 2014 HUMIR operated as the official hungarian myocardial register, and we examined the changes of completeness and validity of data stored in HUMIR. Completeness was calculated with the official numbers of National Health Insurance Fund of Hungary (the central official organ of health insurance, supervised by the Government of Hungary; Hungarian acronym: OEP). As a result, by 2016, the completeness reached around 84%. As the ensemble-based research is based on the same dataset, this value of completeness applies to the current research as well.

After receiving the dataset from HUMIR, we compared the relative performance of our three initial models, which were based on decision tree, neural network, and logistic regression techniques [3]. Area under the receiver operating characteristic curve (ROC AUC) was used in all cases for evaluating performance. For 30-day mortality, we achieved an average of 0.788 for decision tree models, 0.837 for neural net models and 0.836 for regression models on training set (on validation sets: 0.774, 0.835 and 0.834, respectively). In the case of 1-year mortality, these averages were 0.754 for decision tree models, 0.8194 for neural net models and 0.8191 for regression models (on validation sets: 0.743, 0.8179 and 0.8176, respectively). So, differences were non-significant between our neural network and regression, but both signicantly outperformed our decision trees.

In the next study [4], we investigated if an order could be declared between different tuning approaches on decision tree models. 1-year mortality was selected as target variable and K-fold cross validation, repeated cross validation and bootstrap were used to find the optimal parameters for each model on the dataset of HUMIR. The differences were measured in 10 different cases with increasing, randomly selected number (starting from n = 300, until n = 18,000) of records. It was found that a relatively small difference exists, but K-fold cross validation proved to be the best before repeated cross validation and bootstrap.

Then, we involved a new technique to the research: the predictive power of our newly developed Random Forest models were compared with the result of our de-

cision tree [5]. ROC AUC values of Random Forest models for predicting 30-day mortality were 0.843 and 0.847 (training and validation set), while for the 1-year models these were 0.835 and 0.836, respectively. These meant to be a significant difference: Random Forest models were at least 5% better than the decision tree models, but in some cases the improvement is above 9%. This amount of difference between these two tree-based solution could lead to serious relevance in clinical environment.

Generalized Boosted Model (GBM) was the next learning technique which we used to train models [6]. The ROC AUC values of the new GBM-models for 30-day mortality were 0.847 and 0.839 (training and validation set), while for the 1-year models these were 0.828 and 0.821, respectively. The numbers represented a strong and stable learner: the standard deviation of ROC AUC values between models of different imputations were 0.0035 for the 30-day models, and 0.0038 for the 1-year models, both calculated on the validation datasets.

The difference of predictive performance (measured in ROC AUC values) between our RF and GBM models were between 0.5% and 0.9%, except in the case of 1-year model on the validation dataset: it was 1.7% compared to the RF results. Our conclusion said that GBM almost reached the performance of the RF models.

2.2 Ensembled Modelling

Ensembled modelling is one of the most promising area of machine learning-based predicting. In different domains researchers try to combine the advantages of individual classifiers to produce a strong learner. In the current subsection we summarize the results of some of the most-related articles.

Latha et al. [14] used ensembled modelling on the Cleveland Heart Disease Database to improve the accuracy of prediction of heart disease risk. They used weak classifiers like decision tree (C4.5), Bayesian network, Naïve Bayes, Random forest and neural networks to combine them in different ensembled-based modelling techniques like Boosting, Bagging, Stacking and Majority vote. This comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart disease. As a result, a comparison of the various ensembling strategies revealed that the accuracy of the weak classifiers could be increased by a maximum of 7.26%.

Austin el al. found [15] that improvements in the misclassification rate using boosted classification trees were at best minor compared to when conventional classification trees were used. They analysed short-term (30-day) mortality in two cohorts of patients hospitalized with either acute myocardial infarction (N = 16,230) or congestive heart failure (N = 15,848). They observed minor to modest improvements to sensitivity, with only a negligible reduction in specificity.

In another study [16] on the same datasets, Austin el al. evaluated the improvement that is achieved by using ensemble-based methods, including bootstrap aggregation (bagging) of regression trees, random forests, and boosted regression trees. They

found that ensemble methods offered substantial improvement in predicting cardiovascular mortality compared to conventional regression trees; but conventional logistic regression models that incorporated restricted cubic smoothing splines had even better performance. An example of ROC AUC values from their study: on the "EFFECT Follow-up" database, their models achieved the following results by ROC AUC: regression tree: 0.767, bagged trees: 0.820, random forest: 0.843, Boosted trees (depth four): 0.852, Logistic regression: 0.852, Logistic regression—Splines: 0.858, Logistic regression—GRACE score: 0.826.

A neural network ensemble method was proposed [17] by Das et al.. Three independent neural networks models were used (Levenberg–Marquardt, scaled conjugate gradient and Pola–Ribiere conjugate gradient algorithms) as primary learners, and the final, ensembling layer combined their results with averaging. The investigated database contained 303 complete samples. Although they didn't published the predictive performance of the individual models, the final model gained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values on the validation dataset.

Subramanian et al. were also focused on heart failure mortality and used partial patient data from the dataset of Vesnarinone Evaluation of Survival Trial [18]. On the data of 963 patients, they established three logistic regression models to predict survival and an ensemble model learned by boosting. On of the major finding of the study is that their ensemble model performed significantly better than the standard approach of logistic regression. As authors discuss, the reason for this significant increase in predictive accuracy is that "an ensemble of models adjusts better for the biological variability inherent in clinical studies that are derived from patient data."

Although the previous examples were focusing on heart failure and mortality prediction, researchers gain advantages of ensemble modelling in various fields: Bagging, Random Forests and Extra Trees were used by [19] Petkovic et al. when they addressed the task of feature ranking for hierarchical multi-label classification. Extra Tree is similar to RF, with two main differences: instead of using bootstrap replicas, Extra Trees use the whole original sample; and the selection of cut points is random and not an optimum split, like in RF [20]. Three feature ranking scores like Symbolic, Genie3 and the Random Forest Score were investigated and authors found the first two scores yield relevant feature ranking. In the domain of medical image processing, Tóth et al. [21] described an efficient 3D visualization framework in connection with an ensemble-based decision support system.

3 Materials and Methods

3.1 Fundamentals of Ensembled Modelling

Ensembled modelling as a strategy based on the idea that if we combine the predictive performance of different classifiers, it can produce a stronger learner. *Bagging* also known as *Bootstrap aggregation*, *Boosting* and *Stacking* are the main classes of

ensemble learning methods.

1. In *Bagging*, from the original dataset new datasets (called *bootstrap samples* or *bootstrap replicates*) are selected with replacement; we train the models on each of them; and finally the outcome is calculated with averaging (in case of regression) or majority vote (in case of classification).

2. In *Boosting*, simple, 2-3 level depth trees are used and we build the models trying to predict based on the prediction error of the previous tree. The two types are: ADA Boosting and Gradient Boosting.

3. In *Stacking* different types of individual machine learning models are applied (*1-st level learner*) and trained on the same, original dataset, then we combine the prediction results of them in an upper level (*meta-learner* or *second-level learner*).

In the list of previously published models, our RF models is similar to *Bagging* category (although there are some differences between RF and bagged models); and our GBM models belong to *Boosting* category: we were using a given number of decision tree to construct a final, better learner.

In this paper, we are focusing on *Stacking*, as we are using different types of first-levels learners then we try to exploit the common predictive power of them in an upper level.

The schematic overview of Stacking is depicted on Figure 1. As it shows, there can be any number of 1st-level learners, they are trained on the full, original dataset and produce their "local predictions". These different predictions serve as inputs for the meta-learner who attempts to combine these predictions to have the best possible final outcome. As can be seen, the 1-st level learners have to be fully trained and the local predictions have to be made *before* the Meta-learner starts to operate.

In the current study, our 1-st level learners are RF, GBM and NN, while the Meta-learner is Generalized Linear Model, so the ensembled model is a combination of machine learning algorithms and regression models. The modelling structure of the current research is explained and visualized in details in *Chapter 3.4*.

3.2 Patient Record

As we are working on the same dataset to have comparable results, the structure of a patient record is the same as it was in our previous researches and consists of the following groups and fields (the following categorisation is made by the authors of this research to make referencing easier):

1. Group 1: General information about the patient (*Event ID*, *Patient ID*, *If the patient alives*, *Date of death*, *Gender*, *Date of birth*, *ZIP code*)

2. Group 2: Previously reported diseases (*Myocardial infarction*, *Heart failure*, *Hypertension*, *Stroke*, *Diabetes*, *Peripheral vascular disease*, *Hyperlipidaemia*, *Smoking*)

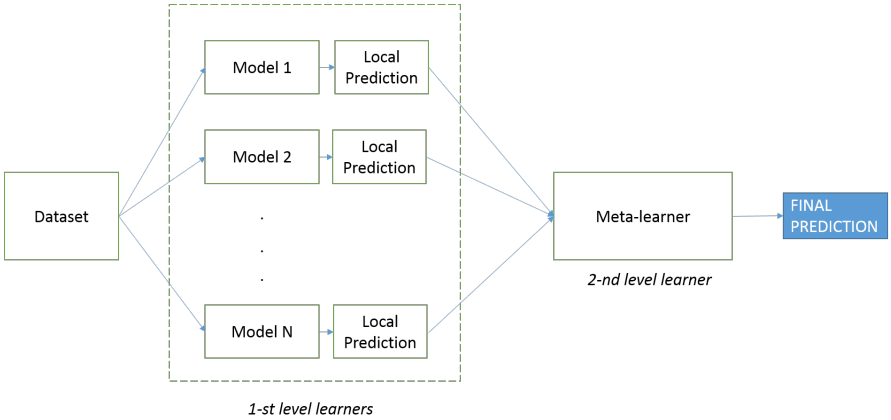


Figure 1
Schematic overview of Stacking

3. Group 3: Information about the pre- and in-hospital treatment (*Prehospital reanimation*, *Cardiogenic shock*, *Percutaneous Coronary Intervention*, *Level of creatinine*, *Diagnosis*, *Treatment ID*, *Date of admission*, *Creatinine*)

We had two fields without holding any relevant information in the given context: *Event ID*, *Patient ID*, they were both eliminated. *Level of creatinine* was almost an empty field (was filled only in 2.8% of the rows), so it was eliminated as well. Finally, a total of 21 fields were involved as the features of our dataset.

3.2.1 Target Variables

As mortality usually examined in short and long run by physicians, our dependent variables (target variables) were the 30-day and the 1-year mortality. Since these values were not initially present in the dataset, the following simple technique was used both here and in our previous works: they were calculated as the date range between two of our fields, the date of hospital admission and the date of the death.

3.3 Missing Values and Imputation

The presence of missing values proved to be essential in our researches. These values in percentage for each field are shown in Table 1 (table contains only the attributes where at least one missing value is present):

To handle the issue with missing data, multiple imputation using Fully Conditional Specification and Bayesian linear regression was applied with 5 imputations and 5 iterations, leaving the final, prepared dataset size at $n = 47{,}391$.

As a result, 5 different sub-datasets were created, and on each we performed the full process of modelling for both the 30-day and 1-year mortality, as it can be seen in the *Modelling structure* subsection.

Table 1
Rates of missing values of attributes

Attribute name	Missing value rate (%)
Hypertension	2.5
Myocardial infarction	4.3
Diabetes mellitus	4.4
Stroke	5.3
Prehospital reanimation	5.8
Creatinine	6.0
Heart failure	7.4
Cardiogenic shock	8.3
Peripheral vascular disease	9.9
Hyperlipidaemia	18.4
Smoking	39.4

3.4 Modelling Structure

After generating the imputations, training and validation sets were created on each imputations with maintaining the original distribution of the target variables. The trainings were used as the input data of the models (on these, the algorithm performed boosting to find the optimal hyperparameters for the given model); while the validation datasets were used to manually measure the prediction performance in ROC AUC.

ROC AUC was applied to select the optimal parameters using the largest value. For each parameter-combination, a bootstrap based validation with 10 resampling iterations were used on the training set to obtain a reliable estimate of model performance.

The Modelling structure can be visualized in three figures: on Figure 2, the full modelling structure is visualized, while the next two figures focus on the separate sections in a more detailed way.

Figure 3 depicts the first step: the connection between the original dataset, the imputations, the target variables and the models as inputs of the ensembled models. It contains only one case (RF model) out of the three, but the same processes were performed for GBM and NN as well.

After we finally had all the 5 (number of imputations) * 2 (number of target variables) * 3 (number of model types) = 30 models, we could go on with the ensembling phase. Figure 4 depicts the connection between the initial models and the ensembed ones.

3.5 Software Environment

As in the connected previous researches, we used R as an open-source software environment for statistical computing and graphics [7].

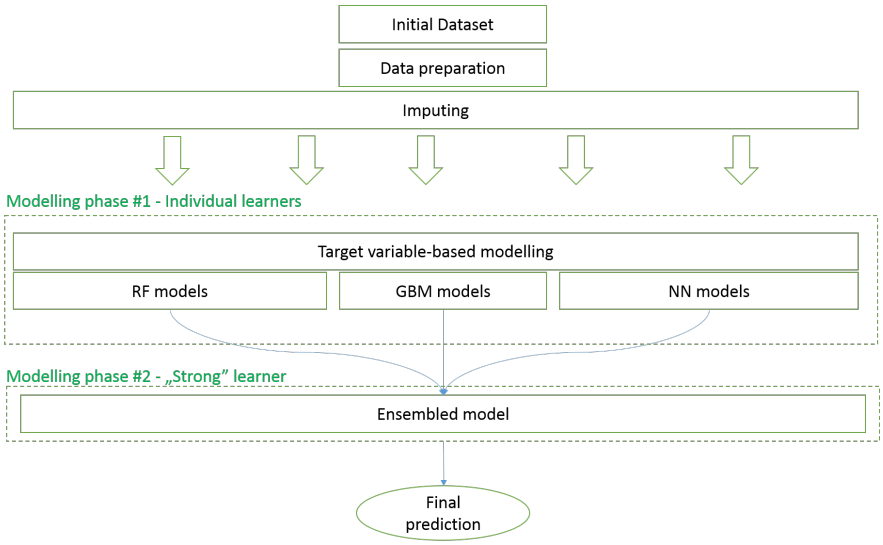


Figure 2
Modelling structure - Overview

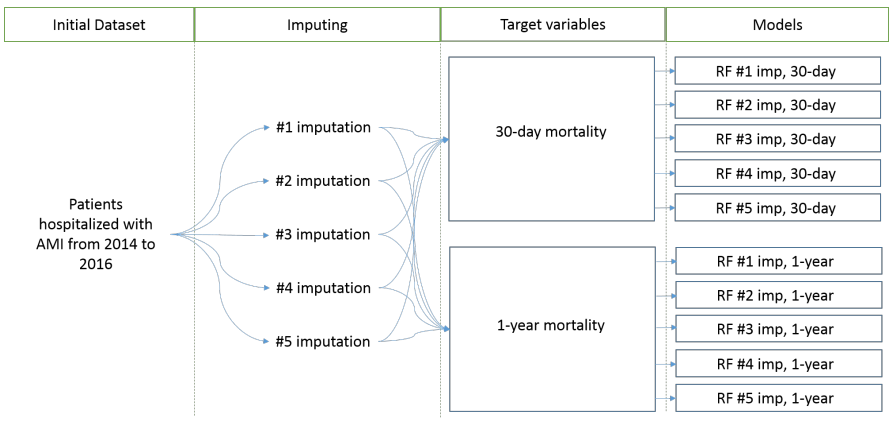


Figure 3
Modelling structure - Step 1

For the three initial models the following packages were applied: randomForest [8], GBM [9] and rms [10] (RF, GBM, LR, respectively), while the ensembling methods were handled by caretEnsemble [11] package. For resampling and training the models, we used Caret [12] package. To deal with missing data and imputation, mice [13] package was used.

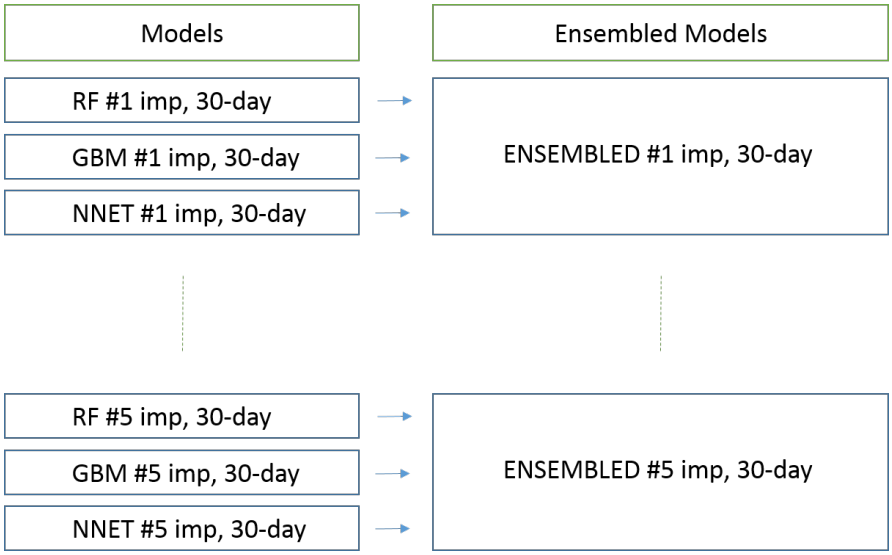


Figure 4
Modelling structure - Step 2

3.6 Hardware Environment

During the investigation, we experienced that the *usual* hardware environment was not suitable for Ensembled modelling on this size of dataset. Since with an *average* configuration *(Intel Core i3 processor, 12 GB memory)*, the training times were around 5 hours with neural networks, we chose *Amazon Web Services* with *EC2* instances with the following parameters: 16 vCPU, 70 ECU, 64 GB memory (*m5.4xlarge* configuration). On this configuration, the average training time for a given model was below 25 minutes.

This hardware configuration differs from the previously published RF and GBM models', but this fact doesn't have any affect on the prediction power.

4 Results and Discussion

4.1 Prediction Capability

In Table 2 and Table 3 we summarized the ROC AUC values of the individual and ensembled models for 30-day and 1-year mortality. All values were calculated on the corresponding validation datasets.

Figure 5 depicts the performance of all the four models in a ROC curve while numerical differences between the methods with 99.2% confidence intervals are shown on Figure 6, both for a randomly selected case (30-day mortality as target variable and

Table 2
ROC AUC values of the 30-days models, validation set.

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5	Avg
GBM	0.8411	0.8381	0.8375	0.8443	0.8346	0.8391
RF	0.8499	0.8472	0.8416	0.8528	0.8436	0.8470
NN	0.8358	0.8334	0.8353	0.8398	0.8326	0.8354
Ensembled	0.8592	0.8542	0.8517	0.8602	0.8522	0.8555

Table 3
ROC AUC values of the 1-year models, validation set.

	Imp. #1	Imp. #2	Imp. #3	Imp. #4	Imp. #5	Avg
GBM	0.8169	0.8202	0.8251	0.8178	0.8246	0.8209
RF	0.8323	0.8332	0.8392	0.8312	0.8384	0.8349
NN	0.8134	0.8166	0.8234	0.8140	0.8224	0.8180
Ensembled	0.8358	0.8371	0.8439	0.8349	0.8432	0.8390

the first imputation was selected). Table 4 reports the standard deviation between the ROC AUC values of the separate models trained on the different imputations.

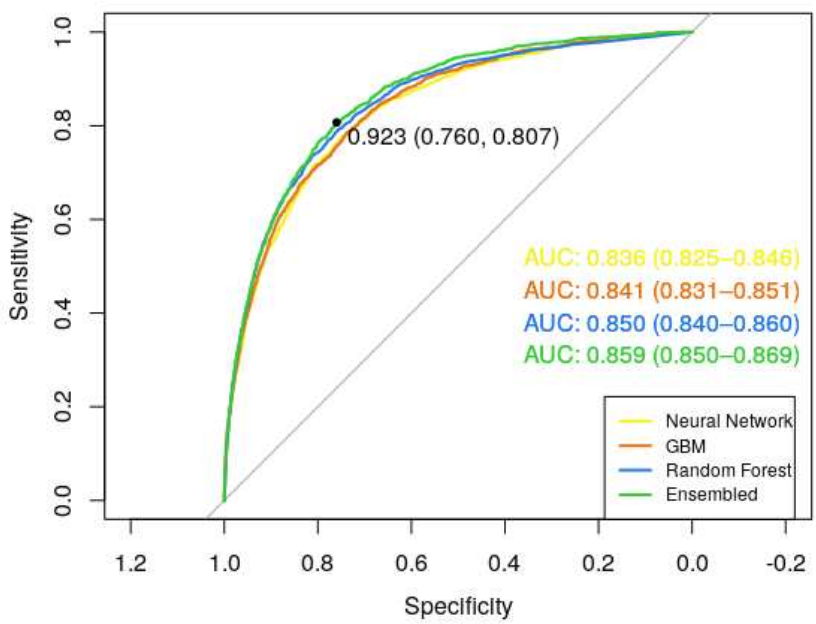


Figure 5
Performance of our Neural Network, Random Forest, Generalized Boosted and Ensembled models. Target: 30-day mortality, dataset: Imputation #1, validation set.

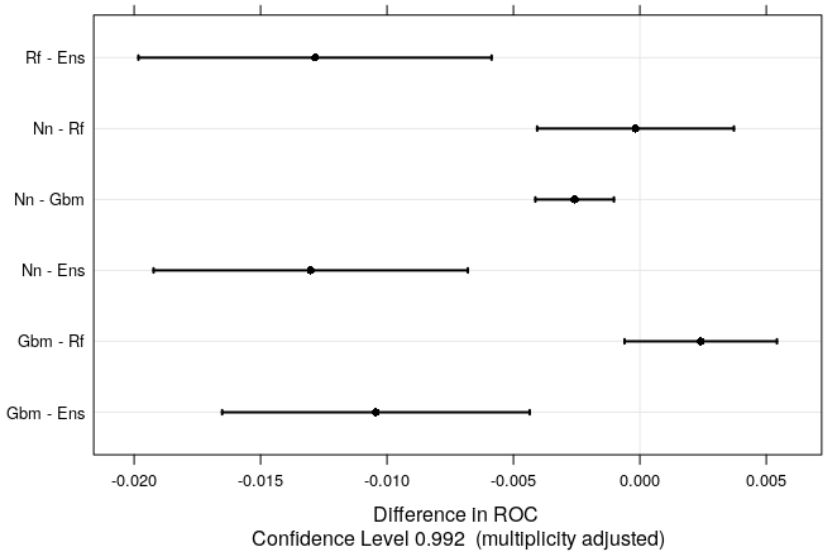


Figure 6
Numerical differences between our Neural Network, Random Forest, Generalized Boosted and Ensembled models. Target: 30-day mortality, dataset: Imputation #1, validation set.

Table 4
Standard deviation of the ROC AUC values of imputations per model type and target variable.

	GBM	RF	NN	Ensembled
30-day models	0.0037	0.0045	0.0028	0.0040
1-year models	0.0038	0.0037	0.0047	0.0043

4.2 Variable Importance

Variable importance, in general, refers to a measure of how much a model uses a given variable to make accurate predictions. In this subsection, we are dealing with the variable importance values for each individual and the ensembled model.

Since the definitions and the methods of calculating the variable importance in separate model types differ, instead of listing the exact feature importance values for each model type, we deal with relative importance: the position of the given feature on the list of the most important fields. With using relative importance it becomes possible to compare the most important features of different models types, i.e. we can make a global order between the variables over the different models.

As we used multiple imputations, we aggregated variable importance values in the imputations: summed up all the relative importance values for each field for a given target variable and for a given model type, then divide this value by the number of imputations. The resulted value represents the relative importance of the given

feature, and in this number, all the imputations added their effects.

The aggregated and relative values of feature importance in descending order for the 30-day models are the following:

1. GBM: Cardiogenic shock (36.3), Age (21.1), Abnormal level of creatinine (10.4), Percutaneous Coronary Intervention (6.7), Prehospital reanimation (6.6)

2. Random Forest: Age (31.1), Cardiogenic shock (14.2), Smoking = never (13.5), Smoking = quit (13.3), Hyperlipidaemia (6.6)

3. Neural net: Age (19.8), Cardiogenic shock (15.2), Percutaneous Coronary Intervention (9.6), Abnormal level of creatinine (9.1), Prehospital reanimation (7.4)

4. Ensembled: Age (26.1), Cardiogenic shock (15.7), Smoking = never (8.6), Smoking = quit (8.4), Abnormal level of creatinine (7.4)

The aggregated and relative values of feature importance in descending order for the 1-year models are the following:

1. GBM: Age (34.1), Cardiogenic shock (16.9), Abnormal level of creatinine (11.9), Percutaneous Coronary Intervention (10), Heart failure (7.8)

2. Random Forest: Age (36.6), Smoking = never (12.4), Smoking = quit (12), Cardiogenic shock (8.2), Abnormal level of creatinine (6.5),

3. Neural net: Age (23.6), Cardiogenic shock (10.9), Percutaneous Coronary Intervention (10.7), Abnormal level of creatinine (9.1), Prehospital reanimation (7)

4. Ensembled: Age (30.8), Cardiogenic shock (10.6), Abnormal level of creatinine (8.3), Percutaneous Coronary Intervention (7.3), Smoking = never (7.3), Smoking = quit (8.4), Abnormal level of creatinine (7.4)

Conclusions

As can be seen in *Section 4.1*, for both target variables, the ensembling technique proved to be the best among GBM, RF and NN models. In addition, it outperforms not only the Decision Tree but the regression models published by the current authors in previous papers and conferences.

In case of the 30-day models, the improvements were 1.64% 0.85% and 2.01% (compared to GBM, RF and NN, respectively), while in case of the 1-year models, these were 1.81%, 0.41% and 2.10% (compared to GBM, RF and NN, respectively), all calculated and compared on the validation datasets.

Although the improvement is typically 1-2%, in the area of cardiovascular diseases, this difference can play a significant role in the hand of physicians when they try to select patients with high-risk of death.

As mentioned in *Section 3*, in the data preparation process we applied multiple imputation using Fully Conditional Specification and Bayesian linear regression. We can state, that the reported standard deviation between the imputations verifies the operation of the imputations. At 5% significance level, the differences between the results of the imputations were non-significant. To assess the differences, pairwise differences of performance measures were calculated and checked if they're equal with zero in expected value using a Welch corrected t-test with Bonferroni adjustment for multiplicity. This means, that next to our RF and GBM models, the Ensembled models also represent a stable learner.

Comparing the most important features from the view of Ensembled model, there were two common items in the list of top 5 features between the ensembled and the constituent models, namely *Age* and *Cardiogenic shock*; *Abnormal level of creatinin* is present in GBM's, NN's and Ensembled's list, while derived fields of *Smoking* appear also in RF and Ensembled.

According to the general results of several papers, that an ensembled model can boost up the predictive performance of the individual constituent models, we can confirm this finding in the case of our models on the dataset of HUMIR. This result is in line with referenced researches in *Section 2*. As a final conclusion, we have found that the combination of machine learning algorithms and regression models results the best performance in mortality prediction on the dataset of HUMIR.

References

[1] Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., ... & American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. (2022). Heart disease and stroke statistics—2022 update: a report from the American Heart Association. Circulation, 145(8), e153-e639.

[2] Fujita, H. (2017, September). An overview of myocardial infarction registries and results from the hungarian myocardial infarction registry. In IOS Press (Vol. 297, p. 312).

[3] Piros, P., Ferenci, T., Fleiner, R., Andréka, P., Fujita, H., Főző, L., ... & Jánosi, A. (2019). Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry. Knowledge-Based Systems, 179, 1-7.

[4] Piros, P., Fleiner, R., Ferenci, T., Kovács, L., & Jánosi, A. (2019, May). Comparing the predictive power of decision tree models with different tuning approaches on Hungarian Myocardial Infarction Registry. In 2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI) (pp. 326-331). IEEE.

[5] Piros, P., Fleiner, R., & Kovács, L. (2020, June). Random Forest-based predictive modelling on Hungarian Myocardial Infarction Registry. In 2020 IEEE 15th International Conference of System of Systems Engineering (SoSE) (pp. 525-530). IEEE.

[6] Piros, P., Fleiner, R., & Kovács, L. (2020, November). Finding improved predictive models with Generalized Boosted Models on Hungarian Myocardial Infarction Registry. In 2020 IEEE 20th International Symposium on Computational Intelligence and Informatics (CINTI) (pp. 179-184). IEEE.

[7] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

[9] Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2019). gbm: Generalized Boosted Regression Models. R package version 2.1.5. https://CRAN.R-project.org/package=gbm

[10] Frank E Harrell Jr (2018). rms: Regression Modeling Strategies. R package version 5.1-2. https://CRAN.R-project.org/package=rms

[11] Zachary A. Deane-Mayer and Jared E. Knowles (2019). caretEnsemble: Ensembles of Caret Models. R package version 2.0.1. https://CRAN.R-project.org/package=caretEnsemble

[12] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). caret: Classification and Regression Training. R package version 6.0-77. https://CRAN.R-project.org/package=caret

[13] Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/.

[14] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.

[15] Austin, P. C., & Lee, D. S. (2011). Boosted classification trees result in minor to modest improvement in the accuracy in classifying cardiovascular outcomes compared to conventional classification trees. American journal of cardiovascular disease, 1(1), 1.

[16] Austin, P. C., Lee, D. S., Steyerberg, E. W., & Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods?. Biometrical journal, 54(5), 657-673.

[17] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680.

[18] Subramanian, D., Subramanian, V., Deswal, A., & Mann, D. L. (2011). New predictive models of heart failure mortality using time-series measurements and ensemble models. Circulation: Heart Failure, 4(4), 456-462.

[19] Petkovic, M., Dzeroski, S., & Kocev, D. (2020). Feature ranking for hierarchical multi-label classification with tree ensemble methods. Acta Polytechnica Hungarica, 17(10), 129-148.

[20] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.

[21] Tóth, J., Tornai, R., Labancz, I., & Hajdu, A. (2019). Efficient visualization for an ensemble-based system. Acta Polytechnica Hungarica, 16(2), 59-75.

A Simplified Approach to Distributed Message Handling in a CQRS Architecture

Munonye Kindson, Martinek Péter

Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics
H-1111 Budapest, XI. Egry Joszef u. 18, Budapest, Hungary
Email: kindson.munonye@edu.bme.hu*; martinek@ett.bme.hu

Abstract: Architecting distributed information system is not a trivial task. This is especially true when a relatively novel design pattern such as Command Query Responsibility Segregation and Event Sourcing is applied. This research aims to describe a simplified approach to the handling of three kinds of messages involved in a CQRS architecture: Commands, Events, and Queries. An evolutionary approach to microservices design was applied to study the information flow within the architecture and establish the reliability properties. Using this approach, a prototype of a Basic Order Fulfillment System was designed and software complexity analysis was applied. The results obtained showed a significant reduction in complexity. The metrics indicated that the approach proposed in this research not only simplifies the process of distributed message handling, but has lower overall complexity than conventional microservices design methods. Therefore, with further refinement, the model can be a standard for building Event Driven Systems.

Keywords: CQRS; Event Sourcing; Command Handling; Aggregates; Event Handling; Microservices; Query Handling

1 Introduction

CQRS is a relatively new architectural pattern for software design based on Command Query Separation (CQS) This concept was proposed by Bertrand Meyer in the late 90s in the context of Object-Oriented Software Construction [1]. The idea of CQS is the separation of operations on application objects into two categories: (1) methods that modify the state of the object are called commands and (2) methods that retrieve the current state of the object are called queries. Naturally, such partitioning adds additional complexity relative to a single partition architecture since both the commands and the queries would have to be in sync to ensure the underlying data store remains in a consistent state.

Command Query Responsibility Segregation (CQRS) which derives from CQS is therefore a design pattern that partitions the commands and queries into separate services based on microservices. It is an Event Driven Architecture (EDA), where communication between the services is achieved via messages [2] [3]. This makes it possible to communicate state changes in an application to other parts of the application anytime when changes in the state occur. These changes occur as a result of a command modifying the read store. Such modification needs to be updated in tandem in the write store. Hence, Event Sourcing (ES) is used with CQRS to enable synchronization between the read data store and the write data store each time a change occurs [4]. Therefore, ES and CQRS are usually used together, though this is not a requirement [5]. However, some challenges need to be overcome. This includes the complexity of the CQRS/ES architecture, performance optimization [6], and managing data conversion across data stores [7]. This research focuses on mitigating complexity-related challenges by simplified message handling (SMH).

The rest of this paper is arranged as follows: Chapter 2 covers the architectural details of the CQRS pattern including aggregates specification, commands, events, and queries. Related works are discussed in Chapter 2. The methodology applied for this research is explained in Chapter 3. In Chapter 4, the analysis, design, and implementation process of a prototype based on the CQRS is covered with a focus on commands, events, and query handling. Chapter 5 presents the results and discussion. Finally, the summary and conclusion sections provide a summary of achievements, limitations, and areas of further research.

2 Theoretical Overview and Related Works

This chapter provides a general detail of the CQRS pattern and establishes the need for efficient message handling.

2.1 Aggregate and Entities

An aggregate is the primary building block for implementing the command model (write-model) in a CQRS-based system. In the context of CQRS, aggregate has variously been defined as a group of entities or functional entities [8] grouped within a single ACID transaction [9]. But a more encompassing definition is:

"an aggregate is a functional entity or group of functional entities in a specific domain and processed together within a consistent boundary such that it is always kept in the consistent state".

The concept of aggregates is generally understood by considering the relational entities in a conventional three-tier architecture. These entities are @Entity annotated classes that map to tables in a relational database. In CQRS, we can consider aggregates as @Aggregate annotated classes that map to aggregate in an

event store and as such, the state of the aggregate is derived (sourced) from the collection of events. Table 1 provides some differences between an entity used in relational systems and an aggregate used in CQRS applications. A summary would be that an aggregate is always made up of an entity (or some entities) but not vice versa. A key point is that aggregates are not instantiated by a 'command-handling constructor'. This constructor publishes an event that is event-sourced by an event-sourcing handler and the aggregate identifier must be set in the event-sourcing handler of the very first event published by the aggregate.

Table 1
Aggregate in CQRS vs Entity in RDBMS

Aggregates in CQRS	Entities in Three-Tier Architecture
Annotated with @Aggregate annotation	Annotated with @Entity annotation
Persisted in an Event Store	Could be stored in a Relational DB
Contains Command Handlers	Contains normal methods
Part of the command model	Implemented as part of the query model
Contains only relevant attributes	Contains all attributes

2.2 Message Handling

The CQRS pattern is a message-driven architecture that leverages the use of message objects. A message is a unit of information or an object that can be passed from one component to another. Therefore, messages have to be defined just like other classes in the application. Messaging-based architecture offers the following benefits:

- maintainability is improved since explicit messaging focuses on message design,
- Message objects can be stored and used subsequently for processing
- Explicit messaging eases the transparent distribution of information to remote components

In this research, the handling of three specific types of messages is considered with each playing a different role. Command messages are those that express an intent to mutate the application state and therefore are routed to a specific destination. Query messages just like queries in relational databases are requests for some data. Event messages are messages that provide some notification that an operation has been performed. Normally, events are generated when a command is carried out. Types of messages in a CQRS architecture are provided in Table 2.

Table 2
Types of Messages in a CQRS Architecture

Message Type	Details	Method	Handler
Commands	Changes state of the application	CommandHandler	*handle()*

Events	Notification that some action occurred	EventHandler	*on()*
Queries	Request for data	QueryHandler	*on()*

2.3 Related Works

Some of the existing works in the area of CQRS/ES are provided in this section. In related research on Determining Critical Properties of the CQRS pattern, the authors of the thesis with the title '*Correctness for CQRS Systems'* worked on determining the proof that a system is correctly implemented [10]. They applied a verification model to determine the critical properties of the system whose presence could determine a correct implementation thereby bridging the gap between system specification and actual implementation. So, the focus was not on mitigating complexity but on determining functionality and correctness. Another research is in the area of using CQRS and Event Sourcing to enhance transaction performance for distributed systems. This research covers the application of CQRS and Event Sourcing on distributed systems as a way to mitigate issues related to monolithic application design such as Create, Read, Update and Delete (CRUD) approach. Performance improvements were also highlighted via the creation of read-optimized views from data denormalization [11]. Additionally, it was shown that persisting of prebuild aggregates could also improve performance.

Research on the application of CQRS to Point Trading System highlights the specific techniques and tools for CQRS pattern in a point trading system [12]. This was achieved in combination with the Actor model and event sourcing to achieve scalability and performance. In this research, aggregates are encapsulated as actors with state and behaviour and therefore can be manipulated in a message-driven way via messages. The aggregates can then be reconstructed by events replay from an audit log which serves as an event source.

2.3.1 Application to Medical Information System

The authors in this research proposed an approach for handling the challenge of fragmented data from several tables which is the case in many medical information systems. Therefore, the CQRS is recommended as a method for denormalizing the read database to reduce data access time for frequent queries [13]. Model-driven approach was used for data modelling, mapping, and code generation. Finally, they proved how a synchronization component can be used for the migration of data from the main database to a read-database using an existing data replicator.

2.3.2 Use of Distributed Command Bus

The book '*Practical Microservices Architectural Patterns*'[2] shows how a distributed command bus can be used for relaying a command generated in one service to another service. Clustering event bus and other components including Advanced Message Queuing Protocol (AMQP) was used to route command and

event between 5 services. This, however, introduced some complexity to the architecture.

Of the related works cited, a key limiting factor is the complexity of the CQRS architecture compared to the 3-tier architecture arising from the addition of more artifacts to ensure synchronization between the command and the query aspects of the system. This is what this research aims to solve.

3 Methodology

The methodology applied in this research follows the algorithm for the minimization of Finite State Machines (FSM) [14]. So, if the message handling within the CQRS architecture can be specified using an FSM, then a reduction algorithm can be applied to eliminate complexity. Therefore, an architecture with x components could be scaled down to y, where $y < x$ without losing any functional requirements.

3.1 Basic Message Handling Model

Figure 1 indicates a general message-handling architecture for three-layer architecture (TLA) and CQRS architecture. Two message types are handled for the layered pattern: request message and response message. For the CQRS, the three messages handled are commands, events, and queries. For the TLA, messages are routed through 4 components while for the CQRS, 8 components are indicated. Therefore, the complexity of the CQRS is obvious from the added interfaces.

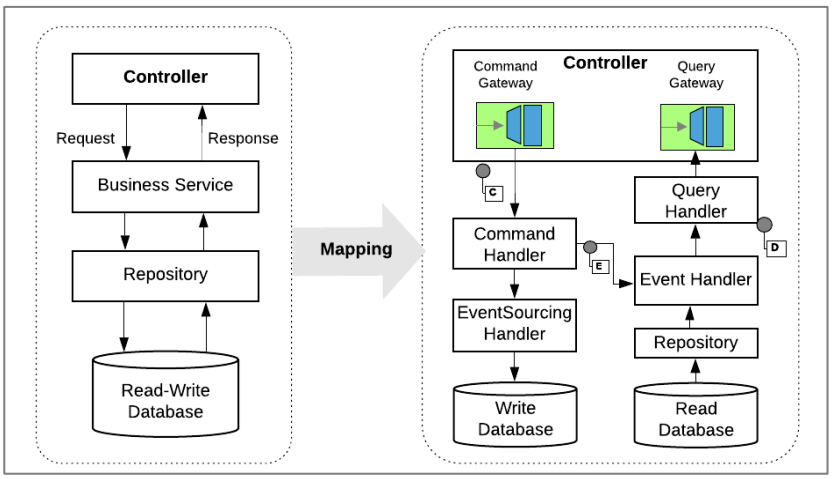


Figure 1
Simplified Message Mapping

To successfully model commands, events, and query handling in CQRS and ES-based systems, it is necessary to first identify the building block of the architecture. Then, information flow across the fabric of the architecture could be formally represented. The following three sections outline the core components playing a role in CQRS and ES. Since events are just notifications of an occurrence of some operation, event handling overlaps with both commands, queries, and common components.

3.2 Command/API Components Modelling

The command components are architectural pieces involved in the execution of a command message from the instantiation throughout the life cycle of the command. These are presented in Figure 2. Codes have been assigned to enable representation using property graphs.

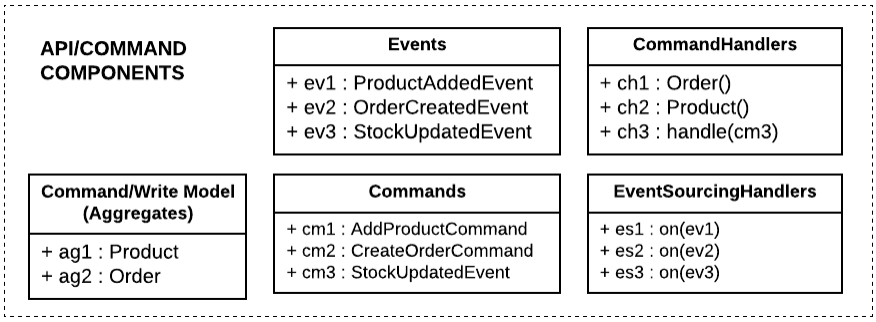


Figure 2
Command Handling Components

Based on Figure 2, any system consisting of k commands, m events, and m aggregates can be represented using equations 1 to 3:

$$Commands\ cm = \{cm_1, cm_2, \dots, cm_k\} \quad (1)$$

$$Events\ ev = \{ev_1, ev_2, \dots, ev_m\} \quad (2)$$

$$Aggregates\ ag = \{ag_1, ag_2, \dots, ag_n\} \quad (3)$$

If there are x commands and y command handlers, then $x = y$. Similarly, if there are x' event and y' event-sourcing handlers, then $x' = y'$. It is also necessary to note that the first command in an aggregate is the *'command handling constructor (CHC)'* since it instantiates the aggregate. Therefore, it has the name of the aggregates. Other command handlers have the name *handle ()*. In Figure 2, the two CHCs include *order ()* and *product ()*. The set of event-sourcing handlers and command handlers is represented in Equations 4 and 5.

$$EventSourcing\ Handlers\ es = \{es_1, es_2, \dots, es_m\} \quad (4)$$

$$Command\ Handlers\ \ ch = \{ch_1, ch_2, \dots, ch_n\} \quad (5)$$

Both the event sourcing handlers and the command handlers are generally defined in the aggregates. The command handlers handle commands publishing events while the event sourcing handlers allow the aggregates to be sourced from their events [15]. Therefore, state changes must be set in the event sourcing handler.

3.3 Query components Modelling

These components are responsible for ensuring that queries originating from external interfaces are routed in an efficient manner and the resulting response is sent back to the user. These are shown in Figure 3

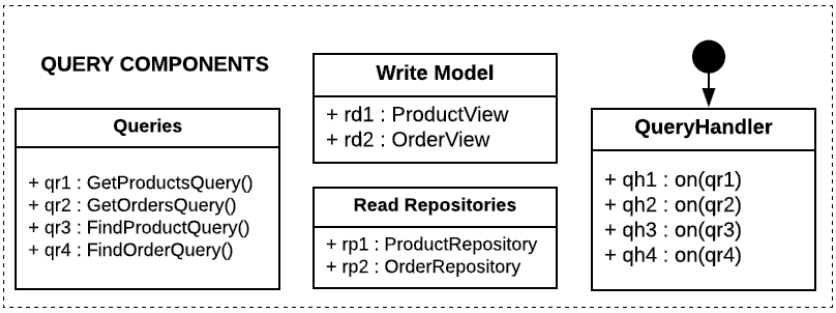


Figure 3
Query Handling Components

The number of query handlers (qh) as can be seen from Figure 3 equals the number of queries. Also, the number of read models equals the number of repositories. These can be represented using the four equations below:

$$Queries\ qr = \{qr_1, qr_2, \dots, qr_n\} \tag{6}$$

$$Query\ Handlers\ qh = \{qh_1, qh_2, \dots, qh_n\} \tag{7}$$

$$Read\ Models\ rd = \{rd_1, rd_2, \dots, rd_m\} \tag{8}$$

$$Repositories\ rp = \{rp_1, rp_2, \dots, rp_m\} \tag{9}$$

3.4 Events/Common Components Modelling

These are components involved in both commands and queries and general message routing. They include the gateways (*gw*), buses (bus), and controller endpoints (*ce*). Events have also been included in this section. They are simply a notification of some operations performed. These are given in Figure 4.

The controller endpoints are the methods that get executed when an *HTTP* request is received based on a *url* pattern. The number of methods would vary based on the system requirement. 6 have been provided here. The event handlers are the actual methods that get executed when an event is emitted: either dispatched from an aggregate as a result of command handling or published from some other object.

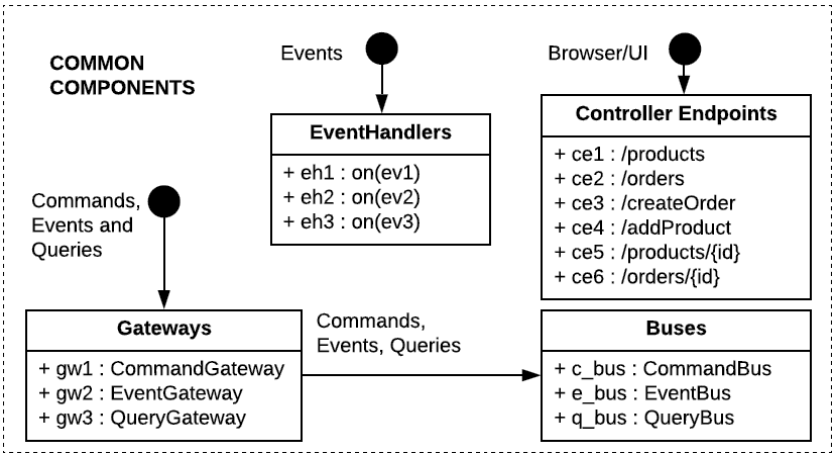


Figure 4
Common Components

The gateways are used to send messages to the buses while the buses relay the messages between components. These artifacts are represented as follows:

$$Gateways\ gw = \{e_{gw},\ q_{gw},\ c_{gw}\} \tag{10}$$

$$Buses\ bus = \{e_{bus},\ q_{bus},\ c_{bus} \tag{11}$$

$$Controller\ Endpoints\ ce = \{rd_1, rd_2, \dots, rd_l\} \tag{12}$$

$$Event\ Handlers\ eh = \{eh_1, eh_2, \dots, eh_m\} \tag{13}$$

3.5 Basic Query Message Handling

To be able to derive the state transitions for the message handling, a very basic message handling is defined: Query Messages for the CQRS which correspond to a normal GET request for TLA. The steps are represented as s_i where i is the number of the specific states as follows:

- A request is sent from a UI service to the controller endpoint (s_1)
- The controller interprets this request as a GET operation and routes it to the business service (s_2)
- The Service receives the request and executes a specific method (s_3)
- The request reaches the repository interface (s_4)
- The repository loads the database driver and connects to the database (s_5)
- Data is retrieved from the database by executing an SQL Select statement (s_6)
- The retrieved payload is sent back to the UI as an HTTP response (s_7)

The equivalent FSM is given in Figure 5. The five states are represented using the components of the architecture.

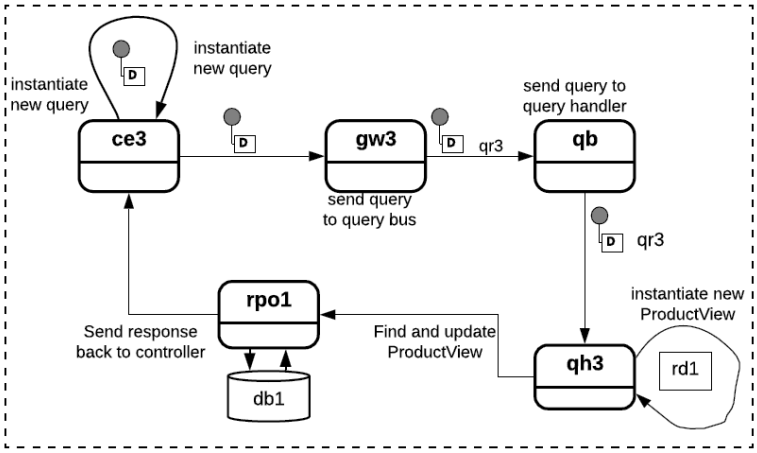


Figure 5
Query Handling in CQRS

3.6 Commands and Events Message Handling

Having considered query handling, this section now models the handling of the two other message types: commands and events. Figure 6 shows, the state machine for the order fulfilment process. The request for new order creation originates from the browser/UI and hits the controller endpoint (*cm1*). The command cm1 is sent through the command gateway (*gw1*) into the command bus (*c_bus*). Since the command handler (*ch1*) is defined in the order aggregate (ag1), the command bus sends the command to ch1. The command is executed and a new event *ev2* is published to the event bus. At the same time, the event-sourcing handler *es2* receives the event and updates the state of the aggregate. The event handler (*eh2*) for *ev2* is defined in *OrderProjector* and therefore gets fired. The *eh2* uses rpo1 to save the new order to the read datastore. Then a stock-updated event *ev3* is fired which is handled by *eh3*. The *eh3* uses *rpo2* to find the entity, update the state end save it back to the read data store. Figure 6 provides the message routing for this process. The detailed use case algorithm is provided in the next section.

Additional operations could be easily modelled using this simple approach. CRUD operations such as the addition of a new product and deleting and updating an order all follow the same approach. Requirements such as customized queries (QR), could simply be designed by defining more query classes in the projector files and adding the relevant rest endpoints (ce) to the controller file. This approach simplifies the CQRS/ES design and eliminates the associated complexity as would be shown in the implementation.

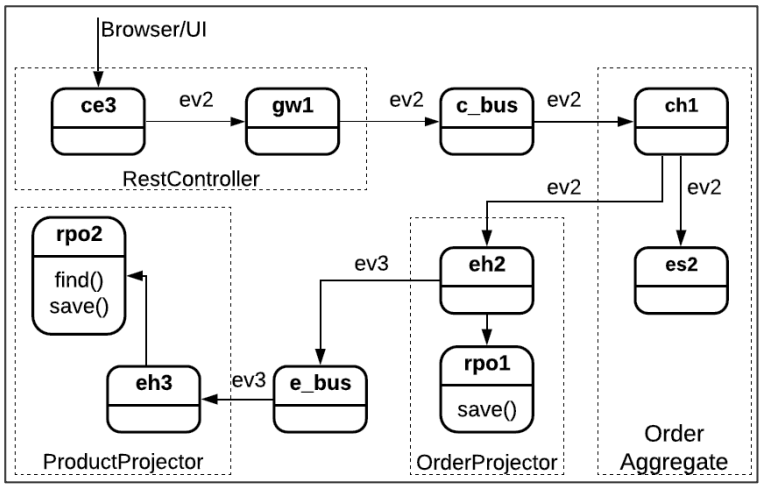


Figure 6
Command and Events Handling

3.7 Use Case Algorithm for Simple Order Processing

In this section, a scenario for message handling which combines all three types of messages is presented. The process is given for CQRS and then analysed and simplified using the FSM minimization algorithm.

The steps are given as:

A user places an order by clicking a link in the User Interface (UI). The request gets to the order controller (*OrderController.java*).

The order controller interprets this request as a post request and instantiates a *CreateOrderCommand* using the request *MultiValueMap* request parameter. This new command is sent through the command gateway to the command bus using the *send ()* async method of the command gateway.

The command is delivered via the command bus to the command handler in the aggregate (*Order.java*).

The command handler, which is a command-handling constructor, is called after that. It invokes the aggregate life,cycle *apply ()* method. This method is used when an event message needs to be published. Hence, a new *OrderCreatedEvent* is created using the details provided in the command and passed to the apply method. The event is published within the scope of the *Order* aggregate.

Then the event sourcing handler is called to ‘event-source’ the aggregate, i.e. to set the state of the new aggregate instance.

The event sourcing of the Order fires the event handler for the *OrderCreatedEvent* defined in the OrderProjector (*OrderProjector.java*). A new *OrderView*(read model) object is instantiated and saved to the database using the repository's *save ()* method.

With this, a *StockUpdatedEvent* object is instantiated using the details for the *OrderCreatedEvent*. This object is published through the event gateway to the event bus.

The event handler for *StockUpdatedEvent* defined in the ProductProjector (*ProductProjector.java*) receives the event from the event bus. Based on the events data, it finds the ProductView (read model) by calling the repository's *findById ()* method. It updates its stock and saves back the changes using the repositories *save ()* method.

Also, the event sourcing handler for the *StockUpdatedEvent* defined in the product aggregate (Product.java) gets fired to set the new state of the particular product aggregate

4 Analysis and Design

This chapter details the design and analysis process and techniques in the prototyping of an e-commerce order fulfilment process. The use case is an example of distributed command and event handling using Axon Framework [2]. We first design a basic CQRS architecture based on a single microservice. Next, it is extended by adding distributed command handling functionality. But first, here is a brief highlight of the key design and implementation tools used.

4.1 Axon Platform

Axon is a platform (Axon Framework + Axon Server) that enables the application of the CQRS and ES architectural style in developing modular, microservices-based solutions. Java 8+ was used for this research and the implementation was done using Maven. Axon abstracts all asynchronous behaviour via the use of executors, which decouple task submission from the internal process of how each task is run.

4.2 Scenario 1 – Basic Message Handling Design

A single CQRS-based microservices is described based on an e-commerce application. The service allows users to place an order. When this happens, a new order is created, and the product stock is depreciated based on the quantity in the order.

Based on the model outlined in Chapter 3, an evolutionary design approach is adopted which begins with the design of a single structured monolithic service. Then, it is partitioned based on the specification.

The message handling architecture in Figure 7 corresponds to the algorithm in section 3.7 of Chapter 3 and therefore would not be repeated. However, the actual implementation structure of the application is presented here.

The structured monolith consists of three packages: controllers, *coreapi*, read_model(command), write_model, and query. These packages are then later taken apart into separate microservices to form a distributed system.

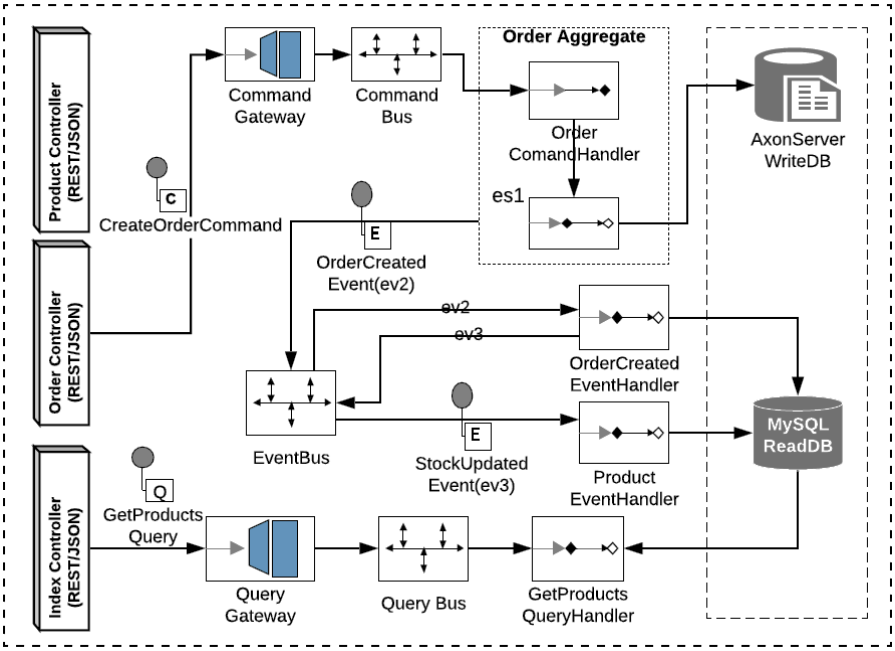


Figure 7

Message Handling Implementation with Axon 4.3

coreapi: The command and events are defined in this package under two respective Kotlin files, *commands.kt* and *events.kt*. These files contain data classes specifying the commands and the events.

write_model: The package contains the implementation of the aggregates. In this use case, two aggregates are identified: Product and Order defined in the *product.java* and *order.java* files respectively.

read_model: This is what is stored in the read database. In this case, it is a relational database.

query: The query package contains the *api.kt* file which has a definition of the queries and repositories. This package also contains the projectors. This is a component where the event handlers and query handlers are defined.

controller: This package contains all the rest controller classes annotated with @RestControllers. Methods available to HTTP endpoints are defined in the controller files.

4.3 Scenario 2 – Distributed Message Handling

In this scenario, the order processing is extended to distribute across JVMs and nodes. The processes are split into six different microservices as shown in Figure 8.

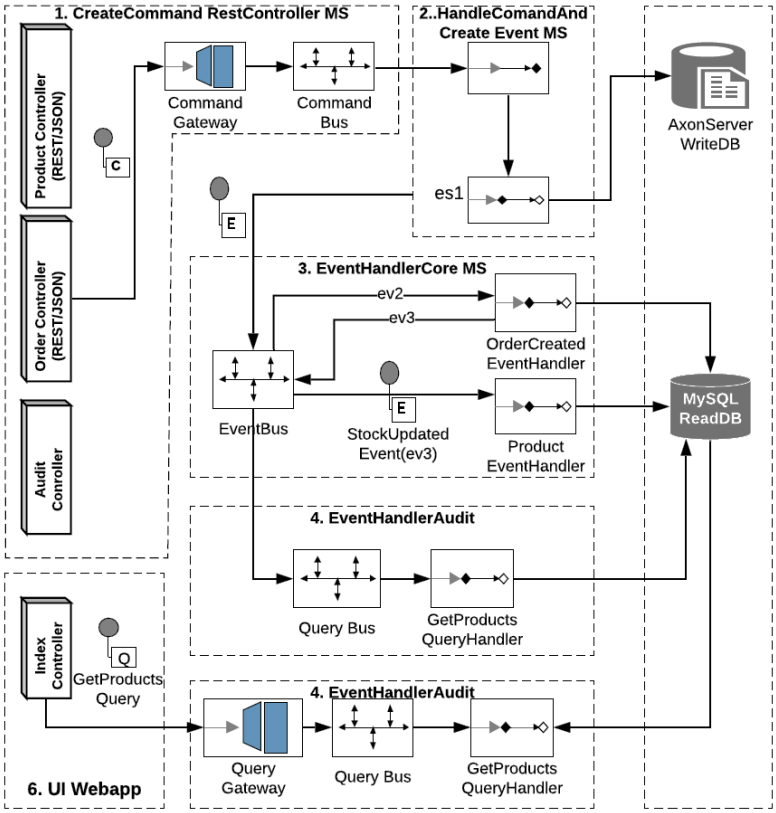


Figure 8
Distributed Command and Event Handling Implementation

The design of the distributed message handling follows from the structured monolithic design. For the distributed architecture, the contents of different packages are then built into an independent service.

UI-App μ-service: This is a UI service built with Angular.

Controller μ-service: This service is similar to the structured monolith except that this sends the *CreateOrder* command to a distributed command bus.

HandleCommand μ-service: This service receives commands coming from the Controller service. Also, it creates and sends out events to the distributed event bus.

HandleEvent μ-service: This service handles events published by the *CreateCommand* service

Audit μ-service: This service also gets notified when a *StockUpdated* event is published by the HandleCommand service

Common μ-service: Contains common configuration used by all the other services.

4.4 Measuring Complexity

As a proof of concept for the model used in this research, we adopt a formal strategy for the measurement of system complexity. Complexity in application development is defined as a measure of the resources expended by a system while interacting with a piece of software to execute a specified task [16]. There are also other variations of this definition, but the basic idea remains the same: a measure of the resources (time and efforts) expended to carry out some operation. The following two methods are used:

First, we use the weighted methods per class (WMC) which is part of the McCabe structure complexity metrics as defined in [17]. The second complexity measure used is known as software science [18] which provides a metric based on the program controls structure, program size, and the nature of the module interfaces.

4.4.1 McCabe Structure Complexity Metric

This approach to complexity measurement considers the program as a directed graph where the edges are lines of control flow while vertices are the linear segments of the program code. This is a computation of the WMC for each of the classes that make up the application. For each class *C*, there are M_1, M_2, … Mn methods defined, where each of the methods has complexity C1, C2, … Cn respectively. The WMC is given by

$$WMC = \sum C_i (i = 1, \dots n) \tag{14}$$

The complexity *C* of method *M* is calculated based on the flow graph derived from the business logic. With the flow graph, a complexity metric V(G) is calculated using the number of edges e, number of nodes n, and number of connected graphs p. The metric is given as:

$$V(G) = e - n + 2p \tag{15}$$

For our prototype services, the flowchart is presented in Figure 8 in Chapter 3. To be able to calculate the required metrics, we need to transform the flowchart into a data flow diagram (DFD). The corresponding data flow diagram is modelled according to the Gane and Sarson notations [19] which represents processes using rounded-corners squares.

4.4.2 Halstead Software Science Metric

This is a measurement of software complexity based on the code structure making up the modules. The Halstead approach computes three complexity metrics: the program Volume (V), the program difficulty (D), and the effort (E).

This measurement is based on the size of the program, the program module interfaces, and the control structure regardless of program comments on the stylistic components making up the code such as naming conventions and indentations.

The first step to evaluating the Software Science metric is to determine the functions making up the program. Then, the number of operators and operands is then determined. With these, we can then compute the volume of the program V and the program difficulty D. The SS metric are defined as follows. Let

- n_1 be the count of distinct operators
- n_2 be the count of distinct operands
- N_1 is the total number of operators
- N_2 is the total number of operands

Then the volume of the program, V is given by:

$$V = (N_1 + N_2)\log_2(n_1 + n_2) \tag{16}$$

And the program difficulty D is given by:

$$D = \frac{n_1 + N_2}{2n_2} \tag{17}$$

The value of D describes the complexity of the program and therefore a higher value of D indicates a higher complexity and vice versa. For this research, we first evaluate the complexity measures for each of the six object services. Then, we compute the parameters also for the combined microservices architecture. The detailed results are presented in Chapter 5.

5 Results and Discussion

This chapter presents the results in terms of complexity reduction achieved using the methodology proposed in this research. Complexity in applications is a measure of the resources consumed by a system while interacting with a piece of software in the execution of a specific task [18]. We focus on two:

- Structured complexity metrics
- Software Science metrics

The results of our methods relative to the existing approach for CQRS/ES are presented. The existing scenario for order fulfilment service can be found on the GitHub repository [20].

5.1 Structured Complexity (SC) Metrics

In this section, the values of the complexity metrics are presented based on the data flow diagram (*dfg*), G of each approach. The three values needed to determine the overall complexity are number of edges (e), number of nodes (n), and number of connected graphs (c). Based on these variables, the overall complexity V is calculated as $V(G) = n - e + c$. Table 3 shows the structured complexity metric values as calculated for the three approaches: structured monolith, existing microservice, and simple message handling. It could be observed that the structured monolith gave a value of 3 while the microservices architecture has a value of 12 which means a 4-fold increase in complexity. The approach in this research indicates a complexity value of 10, which is an improvement over the existing microservices design. However, thins the SC metrics are largely based on *dfg*. The next experiment provides a more complete result which includes program volume, difficulty, and effort as well.

Table 3
SC metrics for design approaches

Metrics for complete microservice	**Metrics**		
	Monolithic	**Existing**	**Our Approach**
Number of edges, e	12	21	**17**
Number of nodes, n	9	16	**13**
Number of connected graphs, c	6	17	**14**
Overall Complexity	3	12	**10**

5.2 Software Science (SS) Metrics

For the first part, the primary components of SS metrics for distributed messaging are tabulated and evaluated. Table 4 provides *N, n, D,* V, and E metrics for both approaches: the existing approach and the simplified message handling approach. Table 4 indicates that our simplified approach required more effort (540.9) than the monolithic approach (525.43) but less than the Microservices (MS). However, our approach provides the least difficulty (D) value of 1.03 against 1.2 and 1.14 for monolith and MS. The reduction in the program volume and difficulty is likely as a result of autoconfiguration provided by Spring and Axon which abstracts certain aspects of the application that would have required addition coding efforts.

An example is the *CommandGateway's send ()* method which uses the *CommandBus* underneath to perform the actual dispatching of commands.

Table 4
SS Metrics comparison

Metrics for complete microservice	Metrics		
	Monolithic	**Microservices**	**SMH**
Number of unique operators – n1	14.00	17.00	15.00
Number of unique operands – n2	25.00	29.00	33.00
Total number of operators – N1	37.00	38.00	41.00
Total number of operands – N2	46.00	49.00	53.00
Volume of program – V	438.69	480.55	524.99
Program difficulty – D	1.20	1.14	1.03
Effort – E	526.43	546.83	540.90

Table 4 indicates the SS metrics for individual services. In the case of the UI Service, the key difference is that it has been built with JavaScript based on AngularJS for the existing architecture. As for the SMH approach, Angular 9 has been used with the actual *HTTP* requests placed in a different *HttpClient service* file written in *TypeScript*. Table 5 shows the SS metrics computed by individual services. An important aspect of Table 5 is the *Common* microservices. This microservice becomes optional in our simplified approach due to autoconfiguration provided by the Spring Framework and therefore reduces the efforts and difficulty. Additionally, it could be noted that the metrics for the Controller MS remain the same in both Table 5 and Table 6 due to having to use the same endpoints and request methods in both approaches.

Table 5
Complexity metrics for Microservice

Microservices	Complexity Metrics						
	n1	**n2**	**N1**	**N2**	**V**	**D**	**E**
UI-App	8	12	14	14	121.01	0.92	110.93
Controller	4	7	11	8	65.73	0.86	56.34
HandleCommand	9	8	15	14	118.54	1.44	170.40
HandleEvent	10	8	13	13	108.42	1.44	155.85
Audit	6	3	6	9	47.55	2.50	118.87
Common	8	9	12	12	98.10	1.11	109.00

The same complexity metrics were calculated for the simplified message handling. Lower values of D and E have a slightly higher value of V indicated in Table 6. The UI-App with a volume of 116.76, a difficulty of 0.83, and an effort of 97.2 perform better than the MS and monolithic approaches since the UI was based on Angular 9 with Node Package Manager (*npm*) which used TypeScript instead of JavaScript. TypeScript is regarded as an extension of JavaScript and enables large-

scale development while addressing the failures of JavaScript [21]. Similar improvement was obtained for other services as well. It could also be noted that the highest program volume is recorded for the UI-App but has the least difficulty. The Audit microservice provides the highest difficulty (2.00) but the least volume (51.89) due to being based on three preceding services and required classes already defined.

Table 6
Complexity metrics for a simplified approach

Microservices	**Complexity Metrics**						
	n1	**n2**	**N1**	**N2**	**V**	**D**	**E**
UI-App	6	12	14	14	116.76	0.83	97.30
Controller	4	7	11	8	65.73	0.86	56.34
HandleCommand	9	8	14	13	110.36	1.38	151.75
HandleEvent	10	7	11	10	85.84	1.43	122.62
Audit	7	4	6	9	51.89	2.00	103.78
Common	0	0	0	0	0.00	0.00	0.00

Comparing the results of Table 5 and Table 6 (simplified approach), it could be observed that our approach not only simplifies the design process but abstracts some of the complexities via autoconfiguration.

Conclusion and future work

This research has provided a realistic approach to managing the complexity associated with the design of a CQRS and ES-based architecture. This was achieved, by first adopting a formal model to represent the components of a CQRS system. The data flow within the architecture was simplified by isolating the commands, events, and query components. The message handling was represented using a generic notation which can be extended to cover more complex systems. Implementation was achieved via an evolutionary approach from a structure monolithic system to a full-fledged microservices architecture. The novel approach of this research was also successfully tested with a use case of the order fulfilment application.

With respect to the CQRS and ES pattern, there are still more areas of further work. For instance, while this approach provided interesting results in a controlled lab-scale network environment, it would also be necessary to evaluate its performance in a cloud-based environment.

Additionally, a key area of interest for CQRS is the management of transactions distributed across services as in a Saga pattern. It is believed that the simplified approach of this research would also mitigate such challenges. However, this is currently being researched and the results would be provided subsequently as well.

References

[1] B. Meyer, "Object-Oriented Software Construction SECOND EDITION." Accessed: Jul. 09, 2020. [Online]. Available: http://www.tools.com

[2] B. Christudas, *Practical Microservices Architectural Patterns*. 2019

[3] K. Machado, R. Kank, J. Sonawane, and S. Maitra, "A Comparative Study of ACID and BASE in Database Transaction Processing," Vol. 8, No. 5, 2017, [Online]. Available: http://www.ijser.org

[4] "State Machine Design, Persistence and Code Generation using a Visual Workbench, Event Sourcing, and CQRS MSc Dissertation Author : Se ´ an Fitzgerald A thesis submitted in part fulfilment of the degree of MSc Advanced Software Engineering in Computer Sci," 2012

[5] P. L. Meena, S. P. Sarmah, and A. Sarkar, "Sourcing decisions under risks of catastrophic event disruptions," *Transp. Res. Part E Logist. Transp. Rev.*, Vol. 47, No. 6, pp. 1058-1074, 2011, doi: 10.1016/j.tre.2011.03.003

[6] Z. Long, "Improvement and Implementation of a High Performance CQRS Architecture," *Proc. - 2017 Int. Conf. Robot. Intell. Syst. ICRIS 2017*, pp. 170-173, 2017, doi: 10.1109/ICRIS.2017.49

[7] M. Overeem, M. Spoor, and S. Jansen, "The dark side of event sourcing: Managing data conversion," in *SANER 2017 - 24th IEEE International Conference on Software Analysis, Evolution, and Reengineering*, Mar. 2017, pp. 193-204, doi: 10.1109/SANER.2017.7884621

[8] G. Maddodi, S. Jansen, and M. Overeem, "Aggregate architecture simulation in event-sourcing applications using layered queuing networks," *ICPE 2020 - Proc. ACM/SPEC Int. Conf. Perform. Eng.*, No. 1, pp. 238-245, 2020, doi: 10.1145/3358960.3375797

[9] "Introduction - Axon Reference Guide." https://docs.axoniq.io/reference-guide/ (accessed Jul. 03, 2020)

[10] H. Kamil, "Correctness for CQRS Systems: Elicitation and validation," 2012, [Online]. Available: www.kth.se/csc

[11] S. O. Diakov, T. E. Zubrei, and A. S. Samoidiuk, "Application of Event Sourcing and CQRS in Distributed Systems," No. 34, pp. 16-22, 2019

[12] Y. Zhong, W. Li, and J. Wang, "Using event sourcing and CQRS to build a high performance point trading system," *ACM Int. Conf. Proceeding Ser.*, pp. 16-19, 2019, doi: 10.1145/3317614.3317632

[13] P. Rajković, D. Janković, and A. Milenković, "Using CQRS pattern for improving performances in medical information systems," *CEUR Workshop Proc.*, Vol. 1036, pp. 86-91, 2013

[14] J. M. Pena and A. L. Oliveira, "Incompletely Specified Finite State Machines," *Comput. Des.*, Vol. 18, No. 11, pp. 1619-1632, 1999

[15] "Aggregate - Axon Reference Guide." https://docs.axoniq.io/reference-guide/implementing-domain-logic/command-handling/aggregate (accessed Jul. 14, 2020)

[16] T. Mens, "Research trends in structural software complexity," *Semant. Sch.*, 2016, [Online]. Available: http://arxiv.org/abs/1608.01533

[17] H. Tu, W. Sun, and Y. Zhang, "The research on software metrics and software complexity metrics," *IFCSTA 2009 Proc. - 2009 Int. Forum Comput. Sci. Appl.*, Vol. 1, pp. 131-136, 2009, doi: 10.1109/IFCSTA.2009.39

[18] J. P. Kearney, R. L. Sedlmeyer, W. B. Thompson, M. A. Gray, and M. A. Adler, "Software complexity measurement," *Commun. ACM*, Vol. 29, No. 11, pp. 1044-1050, 1986, doi: 10.1145/7538.7540

[19] D. Moody, "The physics of notations: Toward a scientific basis for constructing visual notations in software engineering," *IEEE Trans. Softw. Eng.*, Vol. 35, No. 6, pp. 756-779, 2009, doi: 10.1109/TSE.2009.67

[20] "practical-microservices-architectural-patterns/Christudas_Ch12_Source/ch12/ch12-02/Ax2-Commands-Multi-Event-Handler-Distributed at master Apress/practical-microservices-architectural-patterns GitHub." https://github.com/Apress/practical-microservices-architectural-patterns/tree/master/Christudas_Ch12_Source/ch12/ch12-02/Ax2-Commands-Multi-Event-Handler-Distributed (accessed Jul. 13, 2020)

[21] G. Bierman, M. Abadi, and M. Torgersen, "Understanding TypeScript," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, Vol. 8586 LNCS, pp. 257-281, doi: 10.1007/978-3-662-44202-9_11

Convergence Rate for Distributed Macro Calibration of Sensor Networks based on Consensus

Maja Stanković, Dragan Antić, Saša Nikolić

University of Niš, Faculty of Electronic Engineering, Department of Control Systems, Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
majasta@telekom.rs, dragan.antic@elfak.ni.ac.rs, sasa.s.nikolic@elfak.ni.ac.rs

Abstract: In this paper asymptotic convergence rate of a blind distributed macro-calibration algorithm for sensor networks based on consensus is analyzed, assuming the presence of both communication and measurement noise. Convergence to consensus with probability one (w.p.1) and in the mean square sense (m.s.s.) is proved for the sensor gain correction algorithm using general stochastic approximation arguments. Starting from an original formulation of the convergence rate of quadratic Lyapunov functions, an expression for the asymptotic convergence rate of the gain correction algorithm is derived. Using this result, convergence w.p.1 to consensus of the offset correction algorithm is proved under an additional nonrestrictive condition on the step size of the algorithm. In the practically important case when one of the sensors is taken as a reference, convergence w.p.1 of all corrected gains and offsets to the given reference value is proved. Simulation results are provided in order to illustrate characteristic properties of the algorithm.

Keywords: sensor networks; macro calibration; stochastic consensus; stochastic approximation; convergence rate

List of acronyms: a.s. - almost surely; ARMA - autoregressive moving average; e.g. - exempli gratia, meaning for example; i.e. - id est, meaning that is; m.s.s. - mean square sense; w.p.1 - with probability 1

1 Introduction

Sensor networks represent one of very important parts of Cyber-physical systems, Internet of things and complex large scale Control systems [1, 2]. *Sensor network calibration* is one of very actual problems connected to wide deployment of large *sensor networks* in industry, robotics and diverse multidisciplinary fields [3, 4]. In this case, individual calibration of each sensor (*micro-calibration*) cannot be effectively implemented. The so-called *macro-calibration* appears as one of practically efficient and theoretically attractive approaches, treating a given sensor

network as a whole [6, 7]. In this sense, special attention has been paid recently to *distributed methods* for macro-calibration, not requiring centralized measurements and/oractions; *blind distributed macro-calibration* methods represent, theoretically and practically a great challenge since they do not require information about the stimuli [3, 7, 8].

A promising and methodologically consistent approach to *blind distributed macro-calibration* has been proposed in [9-12]. Starting from a sensor model characterized by two parameters (gain and offset), the algorithm proposed therein is of *distributed gradient-type*, providing equal corrected sensor outputs by ensuring convergence to *consensus* of *all the corrected sensor gains and offsets*. In practice, identical readings of all sensors provide: 1) uniformly good measurements for all the sensors in a network, and 2) an important possibility to obtain ideal calibration of the whole network by individual micro-calibration of only one sensor selected as a reference. Convergence of the algorithm has been proved, but without any insight into its convergence rate [9, 11].

Following the line of thought specific for time-synchronization algorithms [13], in [14, 15] a new approach to distributed blind macro-calibration, different from the results presented in [9-12], has been proposed. The algorithm consists of two recursive schemes, in which one is used for independent correction of *sensor gains*, and another for correction of *sensor offsets,* relying on the current gain correction results. It has been demonstrated that this algorithm has theoretical and practical advantages over the previously proposed ones [14, 15, 16]. Having in mind that the recursions for gain and offset correction can be considered as *distributed stochastic approximation algorithms*, efforts have been done to prove convergence to consensus under different operating conditions using their general theoretical background, e.g. [17]. As a result, geometric convergence to consensus has been proved in the noiseless case [14]. In [15], the previously obtained results were substantially generalized by proposing an asynchronous algorithm, working under the presence of additive communication and measurement noises. However, the corresponding theoretical analysis was focused dominantly on the gain correction, as the most critical part. Also, in [16] a new nonlinear calibration algorithm has been proposed, providing high robustness in practice w.r.t. noise samples containing outliers of impulsive type.

As a result, *simultaneous operation* of gain and offset correction in a *stochastic environment* containing both communication and measurement noise has still remained insufficiently clarified for the algorithm proposed in [14, 15]. Such a regime has been completely understood only in the noiseless case as a consequence of the exponential convergence rate. It turned out that the convergence to consensus of the offset correction scheme in a stochastic environment depends on the *convergence rate* of the gain correction scheme, which is far from being exponential. This situation requires a methodologically new tool for the theoretical analysis, since the existing general results cannot be successfully applied, e.g. [11, 15, 17-26].

This paper treats the distributed calibration algorithm proposed in [14, 15] under stochastic disturbances in a new and methodologically original way, trying to formulate the *convergence rate to consensus* of the gain correction scheme, and to prove exactly, for the first time, *convergence to consensus* with probability one (w.p.1) for both gain and offset correction in simultaneous operation. The autonomous scheme for gain correction will be analyzed first using the methodology from [15], generalizing the deterministic approach presented in [14] (Theorem 1 below). Proof of *convergence* w.p.1 and in the mean square sense (m.s.s.) *to consensus* is derived in a novel way. After showing that the methodology from [17], Paragraph 3.1, (proposed in the context of a.s. convergence) may be applied to the analysis of appropriate Lyapunov functions (Theorem 2 below), *convergence rate to consensus* of the gain correction scheme is formulated, as an indispensable precondition for treating the problem of convergence of the offset correction algorithm. Using a selected methodology from [1, 18] and the obtained expression for the rate of convergence of the gain correction algorithm [17], a proof of the offset correction algorithm convergence w.p.1 to consensus has been obtained (Theorems 4 and 5 below).

A special section (Section 5) is devoted to an important extension of the results obtained within Section 4 to the practically important case in which one of the network nodes is selected as a reference. In this case, the whole calibration algorithm succeeds in making all the asymptotical corrected gains and offsets equal to the selected ones w.p.1 and in the m.s.s., see [14, 15, 16]. The corresponding proofs are provided, representing an additional original contribution of the paper.

All in all, a major part of the presented theoretical results can be considered new, providing a new insight into the distributed blind calibration algorithms proposed in [14, 15]. They also provide a new general methodology for determining convergence rate of quadratic Lyapunov functions (Theorem 2), extending an idea from [17]. We hope that the results will fill not only an obvious gap in the theoretical understanding of the basic calibration algorithms from [14, 15] themselves, but also an analogous gap in the general understanding of complex dynamic consensus schemes (having in mind that the algorithm from [14, 15] can be formally considered as a special form of complex dynamic consensus schemes working in a stochastic environment [15]). Notice that the theoretical results from [9-12], having a similar character as the above, cannot be applied to the algorithm treated in the paper, having in mind that: 1) the algorithm itself is basically different, 2) convergence rate has not been analyzed and 3) the analysis methodology developed therein is not applicable to the algorithm from [14, 15]. Namely, the two-way coupling between the corrected gains and the corrected offsets adopted therein implies utilization of mathematical tools (e.g., related to diagonal dominance of the matrices characterizing the system) that are not applicable in the case of the algorithms treated in this paper. On the other hand, numerous existing results have shown that the algorithm from [14, 15] has a superior convergence rate in practice.

A number of simulations illustrate the behavior of the algorithm under different assumptions and operating conditions, enabling getting a better feeling of its behavior in a stochastic environment. These simulations do not serve for direct validation of the derived theoretical results, having in mind their nature - they are strictly valid asymptotically and provide, as usual [17, 24, 25], upper limits of the mean square distance w.r.t. consensus points.

2 Problem Formulation

Assume that there exist *n* distributed sensors measuring a real signal $x(t)$, where t is discrete time$t = 0,1,\ldots$ Sensor outputs are defined by the *standard linear stochastic model*

$$y_i(t) = \alpha_i x(t) + \beta_i + \zeta_i(t) \tag{1}$$

where the sensor *gain* α_i and the sensor *offset* β_i are unknown constants and $\zeta_i(t)$the *measurement noise* always present in real signals (its properties will be specified later), $i = 1, \ldots, n$ [14, 15].

As it is usual [9, 11, 12, 14], we will assume that each sensor implements an *affine calibration transformation* which generates the *corrected (modified) sensor output*

$$z_i(t) = a_i y_i(t) + b_i = g_i x(t) + f_i + a_i \zeta_i(t) \tag{2}$$

where a_i and b_i are the *calibration parameters* to be determined according to the chosen calibration goals, while $g_i = \alpha_i a_i$ represents the *corrected (modified) gain* and $f_i = \beta_i a_i + b_i$ the *corrected (modified) offset*. In general, the calibration goal is to keep g_i close to one and f_i close to zero.

It is assumed that the sensors are mutually interconnected, so that the sensor network is represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes (sensors) and $\mathcal{E}$ the set of directed links (edges) (i, j) (node i sends messages to node j). Let $\mathcal{N}_i^{in}$ represent the set of *in-neighbours* of node i and $\mathcal{N}_i^{out}$ the set of its *out-neighbours* [19].

Remark 1. *In the literature, blind macro-calibration of sensor networks in the sense of finding such parameters* a_i *and* b_i *from (2) which ensure some global properties of a network as a whole, usually assumes a static model (1) and synchronous time [5-14]. Asynchronous functioning based on broadcast gossip is tretaed in [15]. Eventual communication time delays appearing as a consequence of propagation through the network are considered small w.r.t. the variations of the measured signal. In wireless sensor networks, this assumption is simply fulfilled when the nodes from* $\mathcal{N}_i^{out}$ *receive information from node i by broadcasting. Otherwise, blind macro-calibration can easily loose sense, as in the case of large and unpredictable communication delays.*

3 Gain Correction

Following the methodology presented in [14, 15], the algorithm for estimating the correction parameters a_i is derived using the *increments of the local measurement signals*:

$$\Delta y_i(t) = y_i(t) - y_i(t-1) = \alpha_i \Delta x(t) + \Delta\zeta_i(t) \tag{3}$$

where $\Delta x(t) = x(t) - x(t-1)$ and $\Delta\zeta_i(t) = \zeta_i(t) - \zeta_i(t-1)$ are the signal and measurement noise increments. The algorithm (not depending on b_i) is derived starting from the following *local criteria*, aimed at making the *sensor output increments equal*:

$$J_i^a(a_i) = \sum_{j\in\mathcal{N}_i^{in}} \gamma_{ij}\, E\{\left(\Delta z_j(t) - \Delta z_i(t)\right)^2\} \tag{4}$$

where $\gamma_{ij} > 0$ are *scalar weights reflecting the relative importance of the nodes j* from the in-neighbourhood of node i and $\Delta z_i(t) = a_i \Delta y_i(t),\ \ i = 1, \dots, n$. The expression for the gradient of the criterion (4) w.r.t. a_i can be used directly as a generator of *stochastic gradient algorithms* for estimating calibration parameters a_i[14, 15]. In such a way, one obtains the following set of recursions

$$\hat{a}_i(t+1) = \hat{a}_i(t) + \delta(t) \sum_{j\in\mathcal{N}_i^{in}} \gamma_{ij}\, \varepsilon_{i,j}^{\Delta}(t) \Delta y_i(t) \tag{5}$$

where $\hat{a}_i(t)$ is an estimate of a_i, $\delta(t) > 0$ is the step size of the algorithm common for all the nodes, $\varepsilon_{i,j}^{\Delta}(t) = \Delta\hat{z}_j(t) + \xi_{ij}(t) - \Delta\hat{z}_i(t)$, $\Delta\hat{z}_i(t) = \hat{a}_i(t)\Delta y_i(t)$,and$\xi_{ij}(t)$ represents the *additive communication noise* between the nodes j and i (to be specified later). In general, one adopts that $\hat{a}_i(0) = 1$ [14]. Notice that the algorithm is distributed (based on local information), linear and simple for real time implementation.

The algorithm (5) can also be expressed by using the corrected gains $g_i(t) = \alpha_i \hat{a}_i(t)$:

$$g_i(t+1) = g_i(t) + \delta(t) \sum_{j\in\mathcal{N}_i^{in}} \gamma_{ij} \left(g_j(t) - g_i(t)\right) \alpha_i^2 \Delta x(t)^2 + \delta(t) n_i'(t), \tag{6}$$

where$n'_i(t) = \sum_j \gamma_{ij} \{\left[\hat{a}_j(t)\Delta\zeta_j(t) - \hat{a}_i(t)\Delta\zeta_i(t) + \xi_{ij}(t)\right]\left(\alpha_i \Delta x(t) + \Delta\zeta_i(t)\right) + \left(g_j(t) - g_i(t)\right) \Delta x(t) \Delta\zeta_i(t)\}$.

The last relation is important for a proper understanding of the algorithm behavior. The whole measurement noise term is composed of two parts, the first of which is zero-mean, and the second is not (having in mind that $E\{\hat{a}_i(t)\Delta\zeta_i(t)^2\} \neq 0$). An analogous phenomenon can be found in system identification algorithms when the measurement noise corrupts the dynamic system output, leading to biasedness of the parameter estimates [20]. In the case of the above calibration algorithm, this effect is much more pronounced: the algorithm may loose its fundamental capability of achieving consensus, and may become inapplicable to calibration. This problem can be efficiently overcome by applying *instrumental variables*, a very popular tool

for system identification [20]. In our case, instrumental variables have to be *correlated with the measured signalx(t)*, and *uncorrelated with the random noise* $\Delta\zeta_i(t)$, so that the delayed measured noisy signals by at least two sampling instants represent a simple and efficient solution [15]. If $Z_i(t)$ denotes the instrumental variable of the i-th node at the instant t, $Z_i(t) = y_i(t-\tau)$ simply solves the problem for $\tau \geq 2$, having in mind that $\Delta y_i(t-\tau) = \alpha_i \Delta x(t-\tau) + \Delta\zeta_i(t-\tau)$ and that, evidently, $\Delta\zeta_i(t-\tau)$ is uncorrelated with $\Delta\zeta_i(t)$ for $\tau \geq 2$, assuming a realistic situation that $\Delta x(t)$and $\Delta x(t-\tau)$ are correlated.

Consequently, our calibration algorithm based on instrumental variables, is defined by

$$\hat{a}_i(t+\tau) = \hat{a}_i(t) + \delta(t) \sum_{j \in \mathcal{N}_i^{in}} \gamma_{ij}\, \Delta\varepsilon_{i,j}^{\Delta}(t) Z_i(t) \tag{7}$$

where $Z_i(t) = \Delta y_i(t-\tau)$, $\tau \geq 2$. The algorithm consists, in fact, of τ interlaced recursions at the network level, all having identical asymptotic behaviour [23]. The recursion (6) for $g_i(t)$ based on instrumental variables becomes now

$$g_i(t+\tau) = g_i(t) + \delta(t) \sum_{j \in \mathcal{N}_i^{in}} \gamma_{ij} \left(g_j(t) - g_i(t)\right) \alpha_i^2 \Delta x(t) \Delta x(t-\tau) + \delta(t) n_i(t), \tag{8}$$

where $n_i(t) = \sum_j \gamma_{ij} \{\left[\hat{a}_j(t)\Delta\zeta_j(t) - \hat{a}_i(t)\Delta\zeta_i(t) + \xi_{ij}(t)\right]\left(\alpha_i \Delta x(t-\tau) + \Delta\zeta_i(t-\tau)\right) + \left(g_j(t) - g_i(t)\right) \Delta x(t) \Delta\zeta_i(t-\tau)\}$.

A compact form of the algorithm (8) at the network level is

$$g(t+\tau) = [I + \delta(t)\Delta x(t)\Delta x(t-\tau)\alpha^2 \Gamma] g(t) + \delta(t) N(t) \tag{9}$$

where $g(t) = [g_1(t) \dots g_n(t)]^T$, $N(t) = [n_1(t) \dots n_n(t)]^T$ and $\alpha = \mathrm{diag}\{\alpha_1 \dots \alpha_n\}$, while,

$$\Gamma = \begin{bmatrix} -\sum_j \gamma_{1j} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & -\sum_j \gamma_{2j} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & -\sum_j \gamma_{nj} \end{bmatrix}$$

is a matrix of weights having the form of the Laplacian of the underlying directed graph [19, 21, 22].

The convergence analysis given below is based on the general results related to the stochastic approximation method [17, 24, 25, 26] and the properties of dynamic consensus schemes in stochastic environments [18, 21].

We will adopt the following assumptions:

(A.1) Graph $\mathcal{G}$ possesses a central node [22],

(A.2) $\delta(t) > 0$; $\sum_{t=0}^{\infty} \delta(t) = \infty$; $\sum_{t=0}^{\infty} \delta(t)^2 < \infty$,

(A.3) For some $\tau \geq 2$, $E\{\Delta x(t)\Delta x(t-\tau)|\mathcal{F}_t\} \geq c_1 > 0$, $E\{\Delta x(t)^4|\mathcal{F}_t\} \leq c_2 < \infty$, where $\mathcal{F}_t$ is the minimal σ-algebra generated by $x(t), x(t-1)$

(A.4) Signal $\Delta\hat{z}_j(t)$ (or $\hat{z}_j(t)$) sent by node j is received by node i together with the additive noise $\xi_{ij}(t)$ (or $\eta_{ij}(t)$); all random communication noise processes are zero mean and i.i.d. with finite variances.

(A.5) Measurement noise sequences $\{\zeta_i(t)\}$ are mutually independent and composed of zero mean, finite variance i.i.d. random variables.

(A.6) Communication noise, measurement noise and the signal are mutually independent.

Theorem 1 (Convergence of the gain correction algorithm). *Let assumptions (A.1) - (A.6) be satisfied. Then, the estimate $g(t)$ generated by (9) tends w.p.1 and m.s.s. to a random vector $w\mathbf{1}, \mathbf{1} = [1 \dots 1]^T$, where w is a random variable ($E\{w^2\} < \infty$).*

Proof: Let $T = \left[\mathbf{1} \;\; T_{n\times(n-1)}\right]$, where $\text{span}\left(T_{n\times(n-1)}\right) = \text{span}(\alpha^2\Gamma)$. Having in mind properties of Γ expressed by (A.1), it follows that

$$T^{-1}\alpha^2\Gamma\, T = \begin{bmatrix} 0 & \vdots & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & \vdots & \Gamma^* \end{bmatrix}$$

where Γ^* is a Hurwitz matrix, with all the eigenvalues in the left half plane [21, 27]. Define the transformed vector $\tilde{g}(t) = T^{-1}g(t)$, where $\tilde{g}(t) = \left[\tilde{g}(t)^{[1]} \;\; \tilde{g}(t)^{[2]T}\right]^T$, with $\tilde{g}(t)^{[1]} = \tilde{g}_1(t)$ and $\tilde{g}(t)^{[2]} = [\tilde{g}_2(t) \cdots \tilde{g}_n(t)]^T$.

After applying transformation T to (9), one obtains

$$\tilde{g}(t+\tau) = [I + \Delta x(t)\Delta x(t-\tau)T^{-1}\alpha^2\Gamma T]\tilde{g}(t) + \delta(t)T^{-1}N(t) \tag{10}$$

or

$$\tilde{g}(t+\tau)^{[1]} = \tilde{g}(t)^{[1]} + \delta(t)N(t)^{[1]} \tag{11}$$

$$\tilde{g}(t+\tau)^{[2]} = [I + \delta(t)\Delta x(t)\Delta x(t-\tau)\Gamma^*]\tilde{g}(t)^2 + \delta(t)N(t)^{[2]} \tag{12}$$

where $N(t)^{[1])}$ and$N(t)^{[2]}$ are obtained from the vector $T^{-1}N(t)$ as its first and the remaining $(n-1)$ components, respectively, satisfying$E\{N(t)^{[1]}|F_t\} = 0$ and $E\{N(t)^{[2]}|F_t\} = 0$. It is possible to show that $N(t)^{[1]} = \pi N(t)$ and $N(t)^{[2]} = SN(t)$, where π is the left eigenvector of $\alpha^2\Gamma$ corresponding to the zero eigenvalue and S the block matrix in the representation $T^{-1} = \begin{bmatrix} \pi \\ \cdots \\ S \end{bmatrix}$, with ($\dim S = (n-1) \times n$) (see [21]).

Consequently, there exists a symmetric positive definite matrix R such that the following Lyapunov equation is satisfied

$$\Gamma^{*T}R + R\Gamma^{*} = -Q \tag{13}$$

where Q is a symmetric positive definite matrix. Define the following Lyapunov functions $s(t) = E\{\left(\tilde{g}(t)^{[1]}\right)^2\}$ and $v(t) = E\{\tilde{g}(t)^{[2]T}R\tilde{g}(t)^{[2]}\}$.

Remark 2. *The applied methodology of convergence analysis of stochastic consensus schemes based on the quadratic Lyapunov functions originates from [21]. This approach starts from the stability criteria of the underlying stochastic difference equations (11) and (12), and derives the convergence results w.p.1 and in the m.s.s. under appropriate assumptions. In this sense, the result of Theorem 2 givenbelow becomes even more important, having in mind its potential application to the convergence rate analysis of the stochastic consensus schemes in general (not yet existing in literature)*

Starting from the adopted assumptions, we immediately obtain

$$s(t+\tau) \leq s(t) + C_1\delta(t)^2\left(1 + s(t)\right) \tag{14}$$

where $C_1 > 0$.

On the other hand,

$$\begin{aligned} &v(t+\tau) \\ &= E\left\{\tilde{g}(t)^{[2]T}E\left\{(I + \delta(t)x(t)x(t-\tau)\Gamma^{*})^T R(I + \delta(t)x(t)x(t-\tau)\Gamma^{*})\tilde{g}(t)^{[2]}\middle|\mathcal{F}_t\right\}\right\} \\ &+ \delta(t)^2 E\{\| N(t) \|^2\} \end{aligned}$$

For the linear term in $\delta(t)$ we have the following inequality

$$\begin{aligned} E\left\{\tilde{g}(t)^{[2]T}E\{x(t)x(t-\tau)Q|\mathcal{F}_t\}\tilde{g}(t)^{[2]}\right\} &\geq c_1 \min_i \lambda_i(Q)\left(\max_i \lambda_i(Q)\right)^{-1} v(t) \\ &\geq c_3 v(t), \end{aligned}$$

$(0 < c_3 < \infty)$, and for the quadratic term

$$E\{\tilde{g}(t)^{[2]T}\Gamma^T E\{x(t)^2 x(t-\tau)^2|\mathcal{F}_t\}R\Gamma\tilde{g}(t)^{[2]}\} \leq c_4 v(t).$$

$(0 < c_4 < \infty)$. In such a way one obtains

$$v(t+\tau) \leq (1 - \delta(t)k_1)v(t) + k_2\delta(t)^2) \tag{15}$$

where k_1 and k_2 are positive constants [14].

Using the result of Huang and Manton [21], Theorem 11, (14) and (15) imply that $v(t)$ tends to zero w.p.1 and in m.s.s., that $\sup_t s(t) < \infty$ and that $\tilde{g}(t)^{[1]}$ tends to a finite random variable $\tilde{g}_\infty^{[1]}$. According to [21], it follows that $\tilde{g}(t)$ tends w.p.1 and in the m.s.s. $\begin{bmatrix}\tilde{g}_\infty^{[1]} & 0\end{bmatrix}^T$, where $E\{\tilde{g}_\infty^{[1]2}\} < \infty$. Using the proof of Theorem 11 in [21], one obtains that $g(t)$ tends to $\tilde{g}_\infty^{[1]}\mathbf{1}$ w.p.1 and in m.s.s. □

The following theorem is devoted to a modification of the general results from [17], Chapter 3, making them applicable to the formulation of the convergence rate of the gain correction algorithm.

Theorem 2 (Convergence rate of the quadratic Lyapunov functions). *Let*

$$V(t+1) \leq [1-\delta(t)A]V(t) + \delta(t)^2 B \tag{16}$$

where $V(t) \geq 0$, $A > 0$, $B > 0$, *and* $\delta(t)$ *satisfies (A.2), with*

(A.7)$\frac{\delta(t)-\delta(t+1)}{\delta(t)\delta(t+1)} \rightarrow_{t\to\infty} \varepsilon > 0$

(A.8)$\sum_t \delta\,(t)^{2(1-d)} < \infty$ *for some* $d > 0$.

Then $V(t) = o(\delta(t)^{2d})$.

Remark 3. *The result given above has its roots in the approach to the analysis of the convergence w.p.1 of stochastic approximation schemes from [17], Theorem 3.1.1.. The above theorem is focused on quadratic Lyapunov function in order to achieve full compatibility with Theorem 1.*

Proof: From (16) we directly obtain

$$\frac{V(t+1)}{\delta(t+1)^{2d}} \leq \left(\frac{\delta(t)}{\delta(t+1)}\right)^{2d} \left\{[1-\delta(t)A]\frac{V(t)}{\delta(t)^{2d}} + \delta(t)^{2(1-d)}B\right\} \tag{17}$$

As

$$\left(\frac{\delta(t)}{\delta(t+1)}\right)^{2d} = 1 + 2d\frac{\delta(t)-\delta(t+1)}{\delta(t+1)} + O\left(\left(\frac{\delta(t)-\delta(t+1)}{\delta(t+1)}\right)^2\right) \tag{18}$$

one obtains

$$W(t+1) \leq [1-\delta(t)(A-2d\varepsilon)]W(t) + \delta(t)^{2(1-d)}B \tag{19}$$

where $W(t) = V(t)/\delta(t)^{2d}$, having in mind that

$$O\left(\left(\frac{\delta(t)-\delta(t+1)}{\delta(t+1)}\right)^2\right)(\delta(t)^{-1}-A) \underset{t\to\infty}{\rightarrow} 0$$

$$2d\frac{\delta(t)-\delta(t+1)}{\delta(t+1)}(\delta(t)^{-1}-A) \underset{t\to\infty}{\rightarrow} = -(A-2d\varepsilon).$$

Therefore, according to [17], $W(t)$ tends to zero. Thus, the result follows. □

Theorem 3 (Convergence rate of the gain correction algorithm). *Let assumptions (A.1)-(A.8) be satisfied. Then, the gain correction algorithm (6) converges to consensus at the rate*

$$\| \tilde{g}(t)^{[2]} \| = o(\delta(t)^d) \tag{20}$$

Proof: Direct application of Theorem 2 gives directly that $\tilde{g}(t)^{[2]}/\delta(t)^d$ converges to zero in the m.s.s. (compare (19) and (15)). In order to prove convergence w.p.1,

one has simply to take into account that $N(t)$ is a martingale difference process with bounded $E\{\| N(t \|^2\}$. Application of Theorem B.6.1 from [17] leads to the conclusion that $\sum_{t=1}^{\infty} \delta(t)^{1-d} N(t) < \infty$ if $\sum_{t=1}^{\infty} \delta(t)^{2(1-d)} < \infty$, as it is assumed by Theorem 2. Therefore, we can apply Theorem 3.1.1. from [17] to (6) and conclude that $\| \tilde{g}(t)^{[2]} \| = o(\delta(t)^d)$ w.p.1.□

Remark 4. *The practical meaning of the above theorems becomes clearer in the case when* $\delta(t) = 1/t^{\mu}$, $\mu \in (0.5,1]$. *Sufficient conditions for convergence of* $W(t)$ *are then* $A - 2d\varepsilon > 0$ *and* $1/t^{2\beta(1-d)} < \infty$, *implied by* $d < 1 - \frac{1}{2\mu}$. *For* $\mu = 1$ *(which is the maximal value) one obtains* $d < 0.5$. *The convergence rate decreases to 0.5 when* μ *decreases; for* $\mu = 0.5$ *convergence cannot be guaranteed. In the case when* $\delta(t) = 1/t^{\mu}$ *we have* $\| \tilde{g}(t)^{[2]} \| = o(1/t^{\mu d})$.

4 Offset Correction

The basic local algorithm for offset correction results from the general idea from [14, 15]

$$\hat{b}_i(t+1) = \hat{b}_i(t) + \delta(t) \sum_{j \in \mathcal{N}_i^{in}} \gamma_{ij}\, \varepsilon_{ij}(t) \tag{21}$$

where $\varepsilon_{ij}(t) = \hat{a}_j(t) y_j(t) + \hat{b}_j(t) + \eta_{ij}(t) - \hat{a}_i(t) y_i(t) - \hat{b}_i(t)$, $\eta_{ij}(t)$ representing the communication noise when sending $\hat{z}_j(t) = \hat{a}_j(t) y_j(t) + \hat{b}_j(t)$ from node j to node i;oneassumes $\hat{b}_i(0) = 0$ [14, 15]. The corrected offsets$f_i(t) = \beta_i \hat{a}_i(t) + \hat{b}_i(t)$ are given by

$$f_i(t+1) = f_i(t) + \delta(t) \sum_{j \in \mathcal{N}_i^{in}} \gamma_{ij} \left[f_j(t) - f_i(t) \right] + \delta(t)[x(t) +$$

$$\alpha_i \beta_i \Delta x(t) \Delta x(t-\tau)] \sum_{j \in N_i^{in}} \gamma_{ij} \left[g_j(t) - g_i(t) \right] + \delta(t)[n_i(t) + \nu_i(t)] \tag{22}$$

where $n_i(t)$ is defined above and $\nu_i(t) = \hat{a}_j(t) \zeta_j(t) - \hat{a}_i(t) \zeta_i(t) + \eta_{ij}(t)$.

The overall global model at the network levelis given by

$$f(t+1) = (I + \delta(t)\Gamma) f(t) + \delta(t) G(t) \hat{g}(t) + \delta(t) M(t) \tag{23}$$

where$f(t) = [f_1(t) \dots f_n(t)]^T$, $G(t) = [x(t) I + \Delta x(t) \Delta x(t-\tau) \alpha \beta] \Gamma$, $\beta = \mathrm{diag}\{\beta_1 \cdots \beta_n\}$, $M(t) = [m_1(t) \cdots m_n(t)]^T$, $m_i(t) = n_i(t) + \nu_i(t)$, $i = 1, \dots, n$.

Theorem 4 (Convergence of the offset correction algorithm). *Let the assumptions of Theorems 1 and 3 be satisfied, and let*

(A.9) $\sum_k \delta(k) o(\delta(k)^d) < \infty$.

Then, $f(t)$ *generated by (23) converges to* $w'\mathbf{1}$, *where* w' *is a random variable* ($E\{(w')^2\} < \infty$).

Remark 5. *The above theorem shows an important fact that convergence of the corrected offsets to consensus cannot be always obtained. It is important, however, that this fact does not represent a serious limitation for practice.*

*Proof:*As in Theorem 1, at the first step we introduce a transformation $T' = [\mathbf{1} \;\; T'_{n\times(n-1)}]$, where $\mathrm{span}(T'_{n\times(n-1)}) = \mathrm{span}(\Gamma)$, i.e.

$$(T')^{-1}\Gamma T' = \begin{bmatrix} 0 & \vdots & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & \vdots & \Gamma' \end{bmatrix}$$

where Γ' is Hurwitz [21]. After applying transformation T' to (23), we obtain

$$\tilde{f}(t+1)^{[1]} = \tilde{f}(t)^{[1]} + \delta(t)\left[G(t)^{[1]}\tilde{g}(t)^{[2]} + M(t)^{[1]}\right] \tag{24}$$

$$\tilde{f}(t+1)^{[2]} = (I + \delta(t)\Gamma')\tilde{f}(t)^{[2]} + \delta(t)\left[G(t)^{[2]}\tilde{g}(t)^{[2]} + M(t)^{[2]}\right]. \tag{25}$$

where $G(t)^{[1]}$ and $G(t)^{[2]}$ are $1\times(n-1)$ and $(n-1)\times(n-1)$ matrices obtained from

$$(T')^{-1}\Gamma T \quad = \begin{bmatrix} 0 & \vdots & G(t)^{[1]} \\ 0_{(n-1)\times 1} & \vdots & G(t)^{[2]} \end{bmatrix}$$

a) We shall first analyze (25) using the methodology derived from [17], Chapter 3. It is possible to start from Theorem 3.1.1 in [17] and to verify that: a) $G(t)^{[2]}\tilde{g}(t)^{[2]} \to 0$ as $t \to \infty$ and b) $E\{\| M(t)^{[2]} \|^2\} < \infty$, so that $\sum_t \delta(t)^2 E\{\| M(t)^{[2]} \|^2\} < \infty$, implying further that $\sum_t \delta(t) M(t)^{[2]} < \infty$ (according to Theorem B.6.1 in [17]). By direct inspection we realize that the assumptions of Theorem 3.1.1 are satisfied and that $\tilde{f}(t)^{[2]}$ tends to zero w.p.1.

b) Analysis of (24) does not follow strictly the lines of the general arguments from [17]. However, a direct insight shows that

$$\tilde{f}(t+1)^{[1]} = \tilde{f}(0)^{[1]} + \sum_k \delta(k) G(k)^{[1]}\tilde{g}(k)^{[2]} + \sum_k \delta(k) M(k)^{[1]}, \tag{26}$$

For the second term we have

$$\| \sum_k \delta(k) G(k)^{[1]}\tilde{g}(k)^{[2]} \| \leq k_3 \sum_k \delta(k) o(\delta(k)^d),$$

$k_3 > 0$. Boundedness of this term is ensured by (A.7). For the third term we conclude that $E\{\| M(k)^{[1]} \|^2\}$ is bound and (A.2) holds; it follows, using again Theorem B.6.1 from [17], that $\| \sum_k \delta(k) M(k)^{[1]} \| < \infty$. As a consequence, there exists a random variable $\tilde{f}_\infty^{[1]}$ (as in the case of Theorem 1) such that $\lim_{t\to\infty} = \tilde{f}_\infty^{[1]}$ w.p.1.

Thus, the result follows.□

Remark 6. *It is clear that in the standard case when $\delta(t) = 1/t^\mu$, $0.5 < \mu \leq 1$, (A.9) becomes $\frac{1}{t^{\mu+d}} < \infty$; this inequality can be easily fulfilled in practice for a*

given value of d, by choosing μ close enough to 1. This is an example of how the above analysis can be applied to predict the algorithm behaviour in practice. However, one should bear in mind that all the above results are based on sufficient conditions, depending to a great extent on the chosen methodology of analysis and operating conditions.

Convergence rate of the offset correction algorithm will be formulated using the second recursion (25) for the transformed variables, having in mind the existence of an additional exogenous term resulting from the gain correction recursion. Therefore, we shall apply Theorem 3.1.1 from [17] in its original form, dealing with the convergence w.p.1.

Theorem 5 (Convergence rate of the offset correction algorithm). *Let assumptions (A.1) - (A.9) be satisfied. Then ,* $\left\|\tilde{f}(t)^{[2]}\right\| = o(\delta(t)^d)$ *w.p.1.*

Proof: The methodology is based on a modification of the approach in Theorem 2. We first have for t large enough that

$$\frac{\tilde{f}(t+1)^{[2]}}{\delta(t+1)^d} = \left(\frac{\delta(t)}{\delta(t+1)}\right)^d \left\{\frac{1}{\delta(t)^d}(I+\delta(t)\Gamma')\tilde{f}(t)^{[2]} + \right.$$

$$\left. +\delta(t)\left[G(t)^{[2]}\tilde{g}(t)^{[2]} + M(t)^{[2]}\right]\right\} \tag{27}$$

After applying consecutively the steps analogous to those in the proof of Theorem 2, we obtain

$$h(t+1) = [I+\delta(t)\Gamma']h(t) + \delta(t)\left[G(t)^{[2]}\tilde{g}(t)^{[2]} + M(t)^{[2]}\right] \tag{28}$$

where $h(t) = \tilde{f}(t)^{[2]}/\delta(t)^d$. As$\Gamma'$ is Hurwitz and (A.8) holds, $E\{\| M(t)^{[2]} \|^2\} < \infty$ and the proposition of Theorem 3 holds. Therefore, we can apply Theorem 3.1.1 from [17] and directly conclude that $h(t)$ tends to zero w.p.1. Thus, the result follows.□

5 Calibration with Reference

A modification of the basic algorithm is obtained when one fixed sensor (e.g., sensor k) is preliminarily micro-calibrated using the fixed values $\hat{a}_k$ and $\hat{b}_k$; practical importance of such a modification has been pointed out in [14,15,16]. Formally, we have the relations $\hat{a}_k(t+1) = \hat{a}_k(t) = \hat{a}_k$ and $\hat{b}_k(t+1) = \hat{b}_k(t) = \hat{b}_k$ for some k (we say that the network is *pinned to* node k).Assuming that the sensor parameters α_k and β_k are known (e.g., as a consequence of the micro-calibration done preliminarily for sensor k), we also have $g_k(t+1) = g_k(t) = g_k = \alpha_k\hat{a}_k$ and $f_k(t+1) = f_k(t) = f_k = \beta_k\hat{a}_k + \hat{b}_k$. We will show below that under these conditions the proposed calibration algorithms (9) and (23) ensure

convergence of all the corrected gains and offsets exactly as desired to the given reference values g_k and f_k.

Assume without loss of generality that $k = 1$ and $g_k = g$. Then, (9) gives, after simple technical transformations,

$$r(t+\tau) = [I + \delta(t)\Delta x(t)\Delta x(t-\tau)\bar{\alpha}^2\bar{\Gamma}]r(t) + \delta(t)\bar{N}(t) \tag{29}$$

where $r(t)$ is an $(n-1)$-vector defined as $r(t) = \bar{g}(t) - g\mathbf{1}$, with $\bar{g}(t) = [g_2(t) \cdots g_n(t)]^T$, while $\bar{\alpha}^2$ and $\bar{\Gamma}$ are $(n-1) \times (n-1)$ matrices obtained from α^2 and Γ after deleting their first row and the first column; $(n-1)$-vector $\bar{N}(t) = [N_2(t) \cdots N_n(t)]^T$ is obtained from $N(t)$ after deleting its first component. Convergence properties of the recursion (29) can be easily analyzed by applying the methodology of Theorem 1, in relation with the recursion for $\tilde{g}(t)$[2] in (12). Namely, it is possible to show that the matrix $\bar{\alpha}^2\bar{\Gamma}$ is Hurwitz, making possible direct application of the procedure used in Theorem 1 for analysing the Lyapunov function $v(\cdot)$ (having in mind that $\bar{\Gamma}$ is quasi-diagonally dominant and therefore Hurwitz [28]). Consequently, we have:

Theorem 6 (Convergence to the reference gain). *Let the assumptions (A.1) - (A.6) be satisfied and let node 1 be a center node of the graph $\mathcal{G}$. Then $r(t)$ generated by (29) tends to zero w.p.1 and in the m.s.s. for any given value of the reference g.*

In the same way we have a modification of Theorem 3, dealing with the corresponding rate of convergence:

Theorem 7 (Convergence rate to the reference gain) *Let assumptions (A.1)-(A.8) be satisfied and let node 1 be a center node of the graph $\mathcal{G}$. Then, $r(t)$ generated by (29) converges to zero w.p.1 at the rate $\| r(t) \| = o(\delta(t)^d)$*

Similarly as in the case of gain correction in the last subsection, we can start from (23) and obtain for $f_1 = f$ the following recursion for offset correction

$$s(t+1) = [I + \delta(t)\bar{\Gamma}]s(t) + \delta(t)\bar{G}(t)s(t) + \delta(t)\bar{M}(t) \tag{30}$$

where $s(t)$ is an $(n-1)$-vector defined as $s(t) = \bar{f}(t) - f\mathbf{1}$, with $\bar{f}(t) = [f_2(t) \cdots f_n(t)]^T$, $\bar{G}(t)$ is an $(n-1) \times (n-1)$ matrix obtained from $G(t)$ by deleting its first row and its first column, while $\bar{M}(t)$ is an $(n-1)$-vector obtained from $M(t)$ bydeleting its first element. Theorem 4 can be directly applied in the following way:

Theorem 8 (Convergence to the reference offset). *Let assumptions (A.1)-(A.9) hold and let node 1 be a center node of the graph $\mathcal{G}$. Then, $s(t)$ tends to zero w.p.1. for any given value of the reference f.*

The proof follows from the proof of Theorem 4. Notice here only that the rate of convergence of (29) defined by Theorem 7 is not a prerequisite for obtaining the result of Theorem 8. Additional restrictions related to the behaviour of (24) in Theorem 4 do not exist in the case of calibration with reference.

Convergence rate for the corrected offset follows simply from Theorem 5.

Notice here only that the results of this section have great practical implications: a combination of the micro calibration procedure of a selected sensor and the distributed blind macro calibration algorithm applied to the network pinned to the same sensor seems to be a practically very efficient tool for macro calibration of large sensor networks. Some simulation results will be presented below.

6 Simulation Results

In spite of the fact that the paper has a predominantly theoretical character, some simulation results are presented in order to provide a reader with a feeling about noise immunity and achievable performance of the analyzed distributed macro-calibration algorithm in the case of stochastic disturbances. The role of the numerical results in the paper is, in fact, to provide an illustration of the functioning of the algorithm in the presence of communication and measurement noise, rather than to directly validate the main theoretical results related to the convergence rate. This is clear from the character of the main theoretical results themselves. Also, one should bear in mind that we are dealing in this paper with the *asymptotic convergence rate*, which is characterized by the asymptotic mean square value of the error, and cannot be directly observed through finite time transients w.r.t. the initial conditions (the form of these transients represents a completely different phenomenon).

In order to make the computational aspects of the algorithm more transparent, Algorithm 1 below is introduced, containing a corresponding pseudo-code which can be easily programmed using any language (the results given below are obtained using a Matlab program directly derived from Algorithm 1). The given pseudo code represents also a nucleus for any type of real-time implementation.

ALGORITHM 1: **Computation of the calibration parameters**

Initial values $\hat{a}_i(0), \hat{b}_i(0), Z_i(0)$, step sizes $\delta(t), t \geq 0$, weights γ_{ij}

Initialize the iteration counter $t \leftarrow 0$

Repeat

For all $\boldsymbol{i} \in \boldsymbol{\mathcal{N}}$ **do**

Observe measurements $\Delta y_i(t)$ and $y_i(t)$ according to (1) and (3)

Compute $\Delta \hat{z}_i(t)$ and $\hat{z}_i(t)$ using (5) and (21)

Send $\Delta \hat{z}_i(t)$ and $\hat{z}_i(t)$ to all the nodes $j \in \mathcal{N}_i^{out}$

Get data $\Delta \hat{z}_j(t) + \xi_{ij}(t)$ and $\hat{z}_j(t) + \eta_{ij}(t)$ from all $j \in \mathcal{N}_i^{in}$

Update $\hat{a}_i(t+\tau) \leftarrow \hat{a}_i(t) + \delta(t) \sum_{j \in \mathcal{N}_i^{in}} \gamma_{ij}\, \Delta\varepsilon_{i,j}^{\Delta}(t) Z_i(t)$ using (7)

Update $\hat{b}_i(t+1) \leftarrow \hat{b}_i(t) + \delta(t) \sum_{j \in \mathcal{N}_i^{in}} \gamma_{ij}\, \varepsilon_{ij}(t)$ using (21)

end for

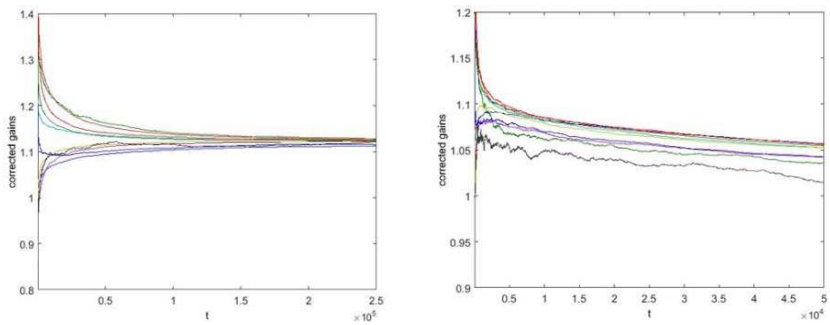
Update the iteration counter$t \leftarrow t+1$

Until convergence

It is obvious that the algorithm is extremely simple, requiring a small number of operations and small memory, and, thus, allowing implementation by using very cheap processing units.

Some basic properties of the proposed algorithm are illustrated by Monte Carlo simulations related to a sensor network with ten nodes, under the presence of both communication and measurement noise. Digraphs satisfying (A.1) and parameters γ_{ij} have been selected at random, as well as the local sensor parameters α_i and β_i about one and zero, respectively, with standard deviations 0.2. The signal $x(t)$ (supposed to be unknown) has been generated by a second order ARMA process, with standard deviation 1. It has been adopted that the step size is defined by the decreasing function $\delta(t) = k/t^{\mu}$, $\mu \in (0.5,1]$.

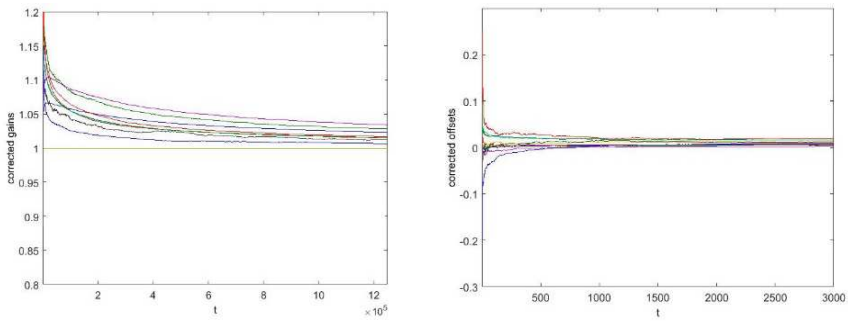
Figure 1 corresponds to the gain correction alone (offset is set to zero), with k=0.1 and $\mu = 0.6$, and noise standard deviation equal to 0.3. Figure 1a) is obtained by the algorithm with the instrumental variables defined above; the quality of the estimates is evidently high. It is obvious that the algorithm can handle the measurement noise efficiently. Figure 1b) illustrates the situation in which the instrumental variables are not applied (noisy gradient is applied directly); the corrected gains diverge, as a consequence of the correlated noise terms. Figure 2a) shows the behavior of the algorithm in the case when the network is pinned to node 1 ($g_1 = 1$ is taken as a reference). Convergence is visible, but the initial convergence rate is lower than in the case of the maximal number of degrees of freedom in adjusting calibration parameters (see also [14, 15, 16]).



a) b)

Figure 1

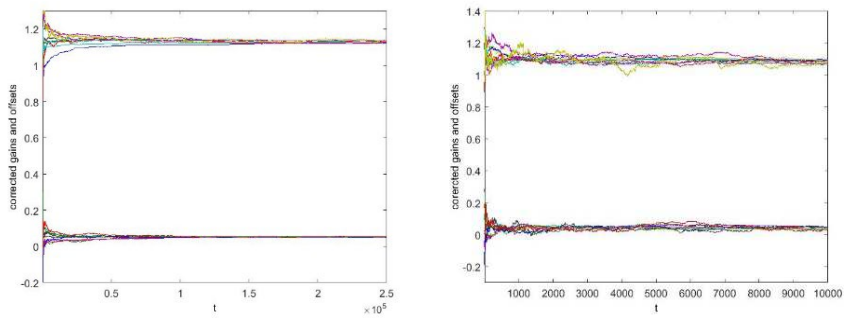
Corrected gains: a) algorithm with instrumental variables, b) gradient scheme



a) b)

Figure 2

a) corrected gains (node 1 taken as a reference), b) corrected offsets



a) b)

Figure 3

Simultaneously corrected gains and offsets: a) low level noise, b) high level noise

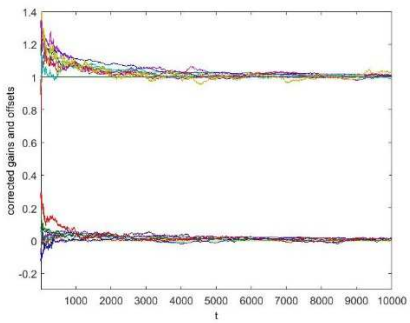


Figure 4

Simultaneously corrected gains and offsets: high-level noise, node 1 taken as reference

Figure 2b) corresponds to offset correction, when the corrected gain is fixed to one, under the same noise characteristics as in the case of Figure 1.

Simultaneous correction of both gains and offsets is theoretically in the focus of the whole paper. Figure 3a) corresponds to this situation, under the same noise characteristics as in the case of Figure 1. The estimation quality is high in spite of the noise presence. In Figure 3b) the noise standard deviation is increased ten times w.r.t. the conditions related to Figure 3a). The estimation quality is still completely acceptable, although the noise level surpasses all that can be expected in reality (having in mind the signal level). Figure 4 depicts the situation corresponding to Figure 3b), but when node 1 is taken as a reference. Notice that the estimates can be additionally smoothed by choosing parameters k and μ, but then the convergence rate may become much lower.

The highly acceptable character of the estimates in Figures 3 and 4 is a consequence of an achieved high asymptotic convergence rate in the theoretical sense.

Conclusion

In this paper, the distributed blind macro-calibration algorithm based on consensus proposed in [14, 15] is analysed in detail in the case of the presence of communication and measurement noise; the existing literature covers the convergence properties of this algorithm only partially, e.g. [15, 21].

The main original contribution of the paper is the formulation of the asymptotic convergence rate of the algorithm for gain correction, not yet existing in the literature. Convergence analysis of the gain correction parameters is done at the first step, showing that under very general conditions the *corrected gains become asymptotically equal w.p.1and in the m.s.s*. The expression for the *convergence rateof the gain correction algorithm* is derived using the general results of H.F. Chen in [17], adapted in an original way to the application to quadratic Lyapunov functions, The obtained formulation of the *convergence rate of the gain correction algorithm* enables obtaining a proof of convergence w.p.1 of the *offset correction algorithm*, including the formulation of the corresponding convergence rate. To our knowledge, such a result is not available in the literature.

The paper also contains an extension of the obtained general results to the case when the *observed sensor network contains a reference node*. An original proof is provided that the corrected gains and offsets converge in this case to the given reference values (under very general conditions).

A number of simulation results serve as an illustration of the typical behaviour of the proposed calibration algorithm under stochastic disturbances.

Further research could be directed to the generalization of the obtained results to more general sensor models involving nonlinearities; this is especially important in the case of the existence of a reference node, having in mind that the error due to nonlinearities may prevent ideal convergence to consensus. As stated above, it

would be a challenge to apply the methodology used in the paper to the convergence rate analysis of the calibration algorithms proposed earlier in [9-12].

References

[1] Greer, C., Burns, M., Wollman, D., and E. Griffor, E.: *Cyber-Physical Systems and Internet of Things*. NIST Special Publication, 2019

[2] Wu, F. J., Kao,Y. F. and Tseng, Y. C.: From wireless sensor networks towards cyber physical systems," Pervasive and Mobile Computing, Vol. 7, pp. 397-413, 2011

[3] Gharavi, H. and Kumar, S., Eds., Proceedings of the IEEE: Special Isssue on Sensor Networks and Applications, Vol. 91, August 2003

[4] Akyildiz, T. F., Su, W. and Cayirci, E.: Wireless sensor networks: a survey, Computer Networks, Vol. 38, pp. 393-422, 2002

[5] Barcelo-Ordinas, J. M., Doudou, M., Garcia-Vidal, J. and Badache, N.: Self-calibration methods for uncontrolled environments in sensor networks: A reference survey, Ad Hoc Networks, Vol. 88, pp. 142-159, 2019

[6] Barcelo-Ordinas, J. M, Chanet, J.- P., Hou, K.-M. and Garcia-Vidal, J.: A survey of wireless sensor technologies applied to precision agriculture, Precision Agriculture, Vol. 13, pp. 801-808, 2013

[7] Whitehouse, K. and CullerD.: Macro-calibration in sensor/actuator networks,Mobile Network Applications, Vol. 8, pp. 463-472, 2003

[8] Balzano, L. and Nowak, R.: Blind calibration in sensor networks, in Proc. Intern. Conf. Inf. Proc. in Sensor Networks, April 2007

[9] Stanković, M. S., Stanković, S. S. and Johansson, K. J.: Distributed blind calibration in lossy sensor networks via output synchronization, IEEE Trans. Aut. Control, Vol. 60, pp. 3257-3262, 2015

[10] Stanković, M. S., Stanković, S. S. and Johansson, K. J.: Distributed calibration of sensor networks, in Proc. 20th Mediterranean Conference on Control and Automation, 2012

[11] Stanković, M. S., Stanković, S. S. and Johansson, K. J.: Asynchronous distributed blind calibration of sensor networks under noisy measurements, IEEE Transactions on Control of Network Systems, Vol. 5, No. 1, pp. 571-582, 2018

[12] Stanković, M. S., Stanković, S. S., Johansson, K. J., Beko, M. and Camarinha-Matos, L. M.: On consensus-based distributed blind calibration of sensor networks, Sensors, Vol. 18, No. 11, 2018

[13] Schenato, L. and Fiorentin, F.: Average timesynch: a consensus-based protocol for time synchronization in wireless sensor networks, Automatica, Vol. 47, No. 9, pp. 1878-1886, 2011

[14] Stanković, M. and Antić, D. S.: Distributed consensus-based calibration of networked control systems, Facta Universitatis, Aut. Control Robotics, Vol. 18, pp. 95-111, 2019

[15] Stanković, M.: Distributed asynchronous consensus-based algorithm for blind calibration of sensor networks with autonomous gain correction, IET Control Theory Appl., Vol. 12, pp. 2287-2293, 2018

[16] Stanković, M. and Antić, D. S.: Distributed non-linear robust consensusbased sensor calibration for networked control systems, IET Control Theory Appl., Vol. 14, pp. 1200-1208, 2020

[17] Chen, H. F.: Stochastic Approximation and its Applications. Dordrecht, the Netherlands: Kluwer Academic, 2002

[18] Huang, M. and Manton, J. H.: Stochastic approximation for consensus seeking: mean square and almost sure convergence, in *Proc. 46th IEEE Conf. Decision and Control*, 2007, pp. 306-311

[19] Godsil, C. and Royle, G.: Algebraic Graph Theory. New York: Springer Verlag, 2001

[20] Ljung, L.: System Identification - theory for the user. Englewood Cliffs, N. J.: Prentice Hall International, 1989

[21] Huang, M. and Manton, J. H.: Stochastic consensus seeking with noisy and directed inter-agent communications: fixed and randomly varying topologies, IEEE Trans. Autom. Control, Vol. 55, pp. 235-241, 2010

[22] Lin, Z.: Coupled dynamic systems: from structure towards stability and stabilizability, Ph.D. dissertation, University of Toronto, 2006

[23] Saridis, G.: Stochastic approximation methods for identification and control - a survey, IEEE Trans. Autom. Control, Vol. 19, pp. 798-809, 1974

[24] Polyak, B. T. and Tsypkin, Y. Z.: Pseudo gradient algorithms of adaptation and learning, Automation and Remote Control, No. 3, pp. 45-68, 1973

[25] Polyak, B. T.: Convergence and rate of convergence of iterative stochastic algorithms, I General case, Automation and Remote Control, No. 12, pp. 83-94, 1976

[26] Borkar, V. S.: Stochastic Approximation: a Dynamical Systems Viewpoint. Springer, 2009

[27] Horn, R. A. and Johnson, C. A.: Matrix Analysis. Cambridge Univ. Press, 1985

[28] Siljak, D.: Large Scale Dynamic Systems: Stability and Structure. North Holland, 1978

Examination of Movement in Young Gymnasts

Klaudia Kaszala[1,2,3], Henriette Steiner-Komoróczki*[3], Gyorgy Fekete[3]

[1]Hungarian Gymnastics Federation, Budapest, Hungary
[2]Hungarian University of Sports Science, Department of Health Sciences and Sports Medicine, Budapest, Hungary
[3]Institute for Cyber-Physical Systems, John von Neumann Faculty of Informatics, Óbuda University, Budapest, Hungary
e-mails: klaudia.kaszala@physcon.uni-obuda.hu; steiner.henriette@nik.uni-obuda.hu; fekete.gyorgy@nik.uni-obuda.hu
*Corresponding author

Abstract: Objectives To improve gymnasts' performance, it is essential to assess their movements and build an effective training plan based on the results. Several aspects were considered in the analysis, including flexibility, dynamics, accuracy, and balance, which are greatly important in gymnastics. We studied the young athletes during the movement analysis, intending to help them achieve a more successful gymnastics career after recording and analysing their data. Design in this study, we analyse the movement analysis of junior gymnasts, which contributes to better competitive performance and personalised training. A movement and gait analysis of Hungarian youth and adolescent national team athletes was carried out, which confirmed that the method and training plan used to design Method Participants appropriately are 14 youth and adolescent national team gymnasts aged 13 to 15 years. Equipment is Goezhe gait Lab video-based computer-assisted motion capture system. Thus, the first measurement position was a smooth gait test. In contrast, the second step was a movement imitating a beam gait, i.e. the subject walked to the end of a one-centimetre line in the middle of the calibrated space. Exams detected and analysed the full-body model at all points, but the knee flexion and extension are the most exciting and essential for ANOVA and Mann Whitney tests. Results of the length of the step cycle increases with training, both on the left and the right side and ordinary and line walking. As a result of training, the range of knee motion increased significantly, and the maximum flexion force increased significantly. Conclusion the presented results showed the efficiency of the classic training methods in strengthening the muscles responsible for correct low extremity posture. The findings suggest a valuable method for posture improvement in young athletes. Further studies with larger sample sizes, age-range and sports backgrounds are required to clarify the efficacy of good training for flexion problems originating from incorrect posture caused by low muscle tone and strength.

Keywords: Biomechanics gymnasts; gait test; balance test; training plan

1 Introduction

The good performance of gymnasts is needed in the athletic race; it is important to assess their movements and build an effective training plan based on the results. [1, 2]

Several aspects were considered in the analysis, including flexibility, dynamics, accuracy, and balance, which are greatly important in gymnastics. We studied the young athletes during the movement analysis, intending to help them achieve a more successful gymnastics career after recording and analysing their data. Therefore, it is worth looking at the history of gymnastics: it was a sport for highly muscular athletes in the past, but now it is clear that this muscle mass is harming the results. The development of muscular strength, acceleration, endurance, and static strength is most effective between the ages of 13 and 15, so we tested flexibility and dynamism in our subjects. We only need to look at the most critical work of the quadriceps hamstring muscle – when kicking off the ground and at the end of exercises, when jumping off the apparatus and in the steadily extended leg posture. [4, 5]

1.1 Literary Research – No Literature

In Hungary, this is the first movement analysis of gymnasts using a movement laboratory, and no similar analysis has been found in the Hungarian or foreign literature. We discovered general musculo-articular and biomechanical descriptions but nothing specific. Gymnastics has numerous elements, so it would have been unrealistic to assess from the point of view of joint movement which joint is working in which position, set, and phase. As no such study has been done, we are starting from ground zero and analysing the gait of junior gymnasts, focusing on joint mobility and the building blocks of the sport: accuracy and flexibility. This data will reveal who is capable of what and how they can improve to perform better in a particular area. As mentioned earlier, 13-15 ages are crucial, so we chose this age group as there is still room to develop the flexibility, dynamism, and acceleration that are considered the basis of explosiveness. [7]

1.2 The Ritual: Sports, Movements, Drugs

Before analysing physical activity, it is worth clarifying the concept of sport, and perhaps most importantly, drawing from the European Charter of Sport: 'sport is defined as any physical activity, whether performed occasionally or in an organised manner, which aims to improve physical well-being and mental fitness, to develop social relationships or to achieve competitive results at various levels'.

There are also two branches of sport, one called participatory or recreational sport and the other spectacular sport. A brief explanation of these is essential because sports can be classified into both categories. In the case of a spectator sport, as the word itself suggests, the spectacle of the sporting event and the emotions that go with it, such as cheering and excitement, are the product to be sold, an experience designed to entertain the athletes themselves, participants. On the other hand, recreational sporting activity is carried out by the amateur competitor himself and is, therefore, a physical activity, with the sport itself as a direct experience. [8]

Furthermore, why does gymnastics fall into both categories? The explanation is simple, on the one hand (recreational sport), there is school physical education and gymnastics played as a hobby, which brings health benefits to the person playing the sport. On the other hand, gymnastics as a professional sport is at the centre, where fans can enjoy the exercises on display, so much so that it has now become the fourth most-watched sport at the Olympics games.

From the perspective of gymnastics, acrobatics is one of the essential starting points, as it is a crucial part of the two fundamental apparatuses, the floor and the vault (the two apparatuses found in both men and women). From a historical point of view, it is interesting to note that archaeological evidence has been found from antiquity that acrobatics had already appeared in Egypt. In the 16th and 17th Centuries, gymnastics originated, and the village fair or circus was considered the forerunner of the competitions. It required an organisation similar to that of a modern-day tournament. Both terms active participants and spectators had a community-building effect. Club gymnastics, like national associations, began to take off in 1800. Unlike today's competitions, there were no judges in the early days. The competitions, combined with other sports and the shows, were a famous festival, with the winners announced by a subjective jury. From a Hungarian point of view, 28 June 1885 is worth highlighting, as gymnastics was then publicly introduced as a sport. A regular and free gymnastics demonstration followed a parade of adults and students in Budapest. The winner had awarded a wreath, and this event became a national holiday. On 29 June, the Hungarian national association MOTESZ, the Hungarian Gymnastics Clubs Association, was formed with 15 clubs. Hungarian athletes had already competed at the 1896 Athens Olympic Games in keeping with historical fidelity. We have 15 gold, 11 silver, and 14 bronze medals in the gymnastics competition at the five Olympic Games. [9]

There are two specific events in today's men's and women's events, floor and vault. In addition, the women have half-bar and beam, while the men have ring, vault, bar, and stretching, so there are four and six events, respectively.

It is an indisputable fact – and all the experts agree that gymnastics is a primary sport that should be part of every young child's life since all sports feed on this movement material and its comprehensive elements help everyone. Several analyses have shown that gymnastics is studied from the point of view of motor

strength, i.e. speed and relative strength, joint mobility, flexibility, balance, and accurate spatial and temporal perception of movement. These are the fundamental skills affected, alongside psychological and aesthetic characteristics.

One of the essential components of gymnastics is the floor exercise. Furthermore, these exercises have been around since before the dawn of time. Modern ground exercises have been calculated since the 1950s – initially, they were called artificial ground exercises. The dimension of the floor is 12 meters by 12 meters and is made from unique materials and layers to provide more bounce, called a spring floor. Ensures that the athletes' joints are not damaged when they land after a jump. Medium where strength, balance, and acrobatic skills count alongside flexibility. Rotation, jumping, handstands, and circling are all part of the routine, along with additional strength elements, and for women, there are dance elements in the routine. The demonstration must be done at a given time, with music for the ladies. The first Hungarian Olympic champion in the event is Ágnes Keleti, who will turn 100 in 2021.

Moreover, what type of training is needed to become a good ground player while constantly practising elements and combinations on the ground and honing techniques? The literature includes natural jumping and static and dynamic balance exercises. In terms of free-form exercises, the focus is on exercises on stools, benches, tools, or racks. Another sport, athletics, also helps develop a good groundworker, whether long jumps from a standing or running position or short sprints.

The beam is exclusive to women's gymnastics and is a unique apparatus, as it is not related to any male apparatus. As early as the 1936 Olympics, competitors were allowed to compete here, but only women. Interestingly, it can also play a role in men's training at a younger age, improving balance, a fundamental skill in gymnastics. It reached its final form in 1950, 10 cm wide, 5 m long, made of wood or plastic. The 10 cm surface acts as a flattened floor, allowing some elements to be performed on it, also used for floor exercises. So that some acrobatic and gymnastic elements are combined in a single exercise, for this reason, the competitors first assemble and prepare their routines on the floor and then adapt them to the beam. Several basic skills are needed to excel in the sport, the main one being balanced, but courage is also important, as are other motor skills typical of gymnastics.

Balance is the ability (vestibular ability) that allows you to keep your body in the desired position or movement while changing postures and movements. The result of multifaceted practice is acquiring a specific balancing skill. The beam itself, by the way, stands on a 20 cm mat. There is also a given time interval, 70-90 seconds. On the floor, Agnes Keleti won the Olympic gold medal in Melbourne in 1956. Among today's gymnasts, the European champion in high bar and vault, Zsófia Kovács, is one of the best, with her first place at the World Cup.

The next event – found in both men and women – is the vault, which has a history stretching back thousands of years. In the old days, the man called the horse vault, as the men performed it in the long jump, while the women performed it on a horse without a crossed cauda equina. They used a buck (a log attached to a stand) as early as the 1800s. However, it was not used in Europe until the second half of the 1900s.

Today's dimensions are 1x1 m surface and exceptional design. We can use different heights for women (1.25 m) and men (1.35 m). The competitor did the run-up for the vault, stepped onto a spring-loaded bumper, then touched the apparatus with both hands, rotated longitudinally and transversely in the air, and landed on the ground with both feet. The gymnast also twists and turns before touching the apparatus in some jumps. We also have an Olympic champion in this event, Henrietta Ónodi, from 1988. The Olympic champion Zsófia Kovács, mentioned above, became the European champion in 2020.

In this case, a good running technique is essential, and this is interesting for our study because the most optimal canter technique is one in which the footrests are parallel to the ground. Motor skills can also be developed with other exercises such as stair-stepping and knee bending. The most necessary for developing specific acceleration power is the serga, with fast kicking from the ankles and knees.

2 Method

In the Ferencváros Gymnastics Hall of Ferencváros, the movements of 14 youth and adolescent national team gymnasts aged 13 to 15 years were the subject of our study, and the measurements were taken before and after the training sessions. As mentioned before, we also examined the balance positions and the gait on the beam for accuracy. There is no coincidence, as a balance position is one of the foundations of gymnastics. Hence, athletes who already boast national championship titles and international results are in great need of developing this skill.

We measured the athletes with the help of the Goethe Gait Lab motion capture laboratory. Our goal was to find a method that is not cumbersome, not expensive, not painful, easy to use, non-invasive, and produces measurable results that can be easily compared. Moreover, that was computer-controlled video-based motion tracking, with visible light as the essential tool to aid optical tracking. Four video cameras recorded the motion sequence under investigation. They are connected online during the test, controlled by a computer. We created an open terrain, as this was the best way to mimic the gymnasts' training and competition conditions, where nothing interferes with them, so we tried to work with as few cables and wires as possible to assess the competitors in a prominent field.

The most important part of the measurement is the calibration, for which a unique body is used, described below. After data acquisition comes transformation, i.e. converting the images from 2D to 3D, finally filtering and denoising the data.

Calibration is all about reducing the longitudinal distortion of the cameras, which is why we need 18 calibration points and an independent fixed point, as our primary goal is to achieve the highest possible accuracy.

We use a particular sphere grid developed by the Goethe Gait Lab, which gives high accuracy, practically less than 1 per cent, which means that we can test deviations in the millimetre range. Similarly, using 18 calibration points, the position of the chambers can be measured.

It is important to note that the fixed point is the absolute zero of the system. A salient point on the calibration body defines the zero point of the Descartes coordinate system concerning which the directions are given. X coordinate is taken as the (sagittal) direction of progress, i.e. the direction of motion. The coordinate is the vertical direction, i.e. the direction in which the foot is lifted, and the z-direction is the lateral direction (horizontal). It was necessary to create this type of coordinate axis with a left-hand twist to be paired with the designations used in medical examinations. It could be presented and interpreted by both trainers and the medical profession. [10, 11]

The basic test is the so-called Dempster's body model, which applies to the whole body and follows the body's movement. Naturally, the anthropometric characteristics of the subjects are measured beforehand, and the model has adapted accordingly. (Fig. 1)

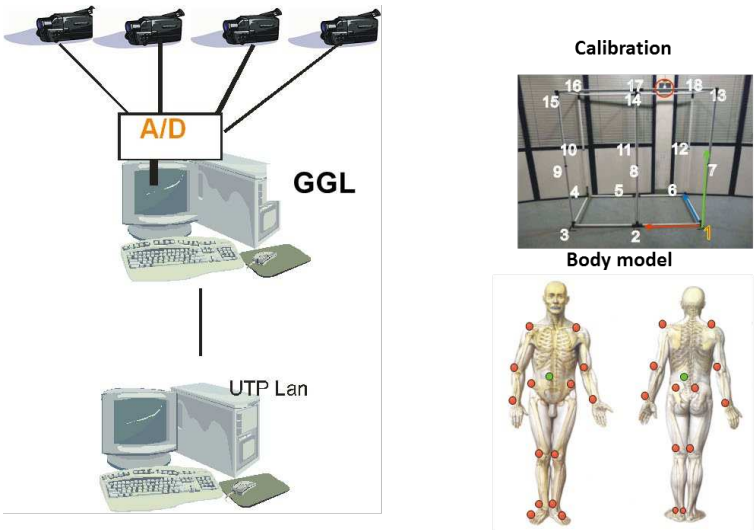


Figure 1
GGL system (hardware, calibration cube, body model: markerset)

The participants in the study wore black clothes, on which we applied reflective markers, i.e. we used the black and white technique. Both passive and active markers are helpful in this system, but we only used the passive ones, mainly because of the fast movements of the gymnasts. These markers need much lighting, and we used special spotlights for this. The markers were placed on the joints of the participants in the study using triangulation, i.e. we tried to determine the anatomical rotation point of the joint by placing three reflective points on each joint. In addition to the measured data, a unique database was used to record any additional data that were specific to the subject.

Two different measurement settings were used in the study. We were looking for a form of movement typical in gymnastics that could easily be compared with other groups' measurements. Thus, the first measurement position was a smooth gait test. In contrast, the second step was a movement from imitating a beam gait, i.e. the subject walked to the end of a one-centimetre line in the middle of the calibrated space.

In sports, movements are based on time outcomes, so the data were often not analysed using the distance-time function in the function analysis. We defined local minima (the start of the stride cycle during the gait) and maxima in the function analysis. They were examined and their relation to each other and their distance. In addition to this, statistical analysis was also carried out, using other methods than the basic statistical methods – ANOVA, Mann – Withney and regression.

In all cases, our null hypothesis was that the results of the two studies were not different (95% threshold set at $p =< 0.05$).

3 Results

The results obtained should be highlighted that our motion analysis is based on the step cycle for several reasons. On the one hand, this form of movement can be performed by everyone, regardless of whether they are an athlete or not, and thus the results obtained can be compared with other results in different variations. [12, 13] We have developed a test procedure that focuses on the factors of great importance in gymnastics. [14, 15] In the preparation of training plans, coaches emphasise flexibility and dynamism, so we have designed a training plan-oriented analysis in terms of the usefulness of the results. It will show to what extent this analysis helps trainers in their work and to what extent and how emphasis should be placed on developing these skills. This professional desire should also be taken into account in the approach from the engineering side. In the following is the characteristics of normal, healthy movement and the main parameters of our research. (Fig. 2, Fig. 3)

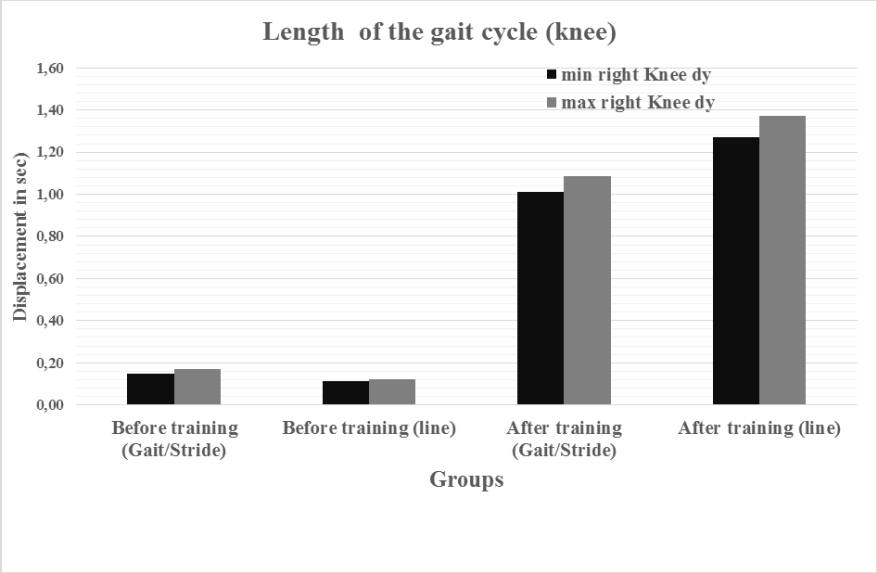


Figure 2

Lengths of the gait cycle (right and left knees in vertical plane)

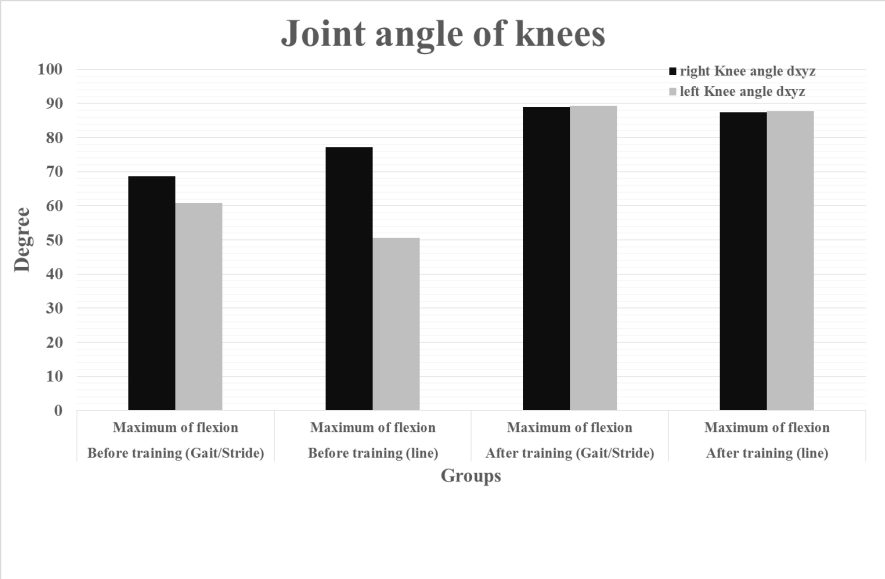


Figure 3

Joint angles of the knees (maximum flexion of right and left knees in 3D)

4 Discussion

The gait cycle is defined in terms of the path-time function, and the length of the step cycle is the distance from the heel strike of a leg to the next heel strike or the time it takes to reach that distance. [16, 17] According to the literature, the gait cycle consists of two phases: the support phase and the swing phase. The support phase is when the foot is on the ground, and the swing phase is when the foot moves forward. In our measurement, the aim is to make the right and left sides separable and accessible. Man has slightly modified the literature of physiotherapists and doctors, which means that we always consider the time and distance from the moment a heel strike is made to the moment the next heel strike is made.

If we consider the stride cycle to be 100 per cent, then in a healthy case, the proportion of the support phase is 60 per cent, and the proportion of the swing phase is 40 per cent. During the stride cycle, the centre of mass is also considered. It is also looked at in a healthy case, in which case it will give a sagittal, or x, angle of 45 degrees in the direction of gait. [18]

In addition to x, we also considered y-directional deflection, which represents explicitly the rate and height at which the heel or ankle rises off the ground. In this case, we can also analyse the movement of the centre of gravity, which shows that the centre of gravity follows the movement very well with its sinusoidal nature.

The z-directional deflection was also mentioned, which is the lateral displacement. In this case, it is shown how the limb follows this movement with lateral displacement.

We have given the step as a form of movement with three directional parameters. We can specify the position of the swing phase and the support phase in each direction and see characteristic deflections on the curve in each direction.

Since we are talking about gymnasts, we consider it extremely important for training planning to have as much information as possible about the dynamics of the movement. For this purpose, we have not only considered the step cycle as a function of time but also analysed the individual stages of the step cycle, i.e. the minimum and maximum of the curves and their distance. Here, too, a uniform system was developed, with the minimum being the ground grip and the maximum the highest point of the step cycle. The distance between the two was not defined in centimetres and time, as this gives an idea of the dynamics of the movement.

When analysing gait, it is essential to consider the movement of the joints. In this case, the movement and angular range of the knee joint is the most important for us. The literature shows that the normal knee angle is 60 degrees when walking horizontally and 80-90 degrees when walking upstairs. Passive knee flexion can reach 130-140 degrees with strong hip extension and 160 degrees with powerful movement. [18, 19, 20]

If we analyse the gymnasts' stride cycle, we first find an asymmetry between the right and left sides, which is also typical of the average population. If a gymnast wants to achieve a good result, this asymmetry must be reduced. Think of a floor exercise or a beam exercise.

It is also observed the following that the length of the step cycle increases with training, both on the left and the right side. ($p < 0.005$) (Fig. 2) This trend is also observed for normal and line walking. The analysis of the phases of the stride cycle continued and at this level proved a significant thesis for gymnastics. By examining the distance between the first (minimum and maximum) and the second (maximum and minimum) phases of the swing phase, it can be concluded that the swing phase is essentially shortened as the minimum-maximum distance increases. The initiating phase becomes more pronounced than before, i.e. the motion becomes more explosive.

Furthermore, explosiveness is a fundamental tenet of gymnastics, which a gymnast needs. It is fascinating that this type of shortening is much more pronounced when walking on the line (imitating a beam). The stride cycles are much shorter, which means that this type of walking will be much easier at the end of the training session. In summary, this is the most effective movement in training, which is particularly beneficial for gymnastics. As we have studied young athletes, it can also be considered very effective, as precision, flexibility and muscular explosiveness are required to perform the exercises in gymnastics.

As indicated above, when looking at joint angles, we are concerned with the knee joint for gymnasts. The angular extent is defined as follows: between the femur and tibia, as a three-dimensional vector. (Fig. 3.)

As a result of training, the range of motion increased significantly, and the maximum flexion force increased significantly. The state of maximum flexion during the step gives the maximum of the path-time function describing the flexion position and angular motion.

Conclusions

From the above results, it is clear that a well-executed training plan or training programme for young athletes increases flexibility and explosiveness and muscle strength. It can be seen from the increase in stride cycle both in natural gait and in line walking. The latter also tests the degree of balance retention by imitating movement on a beam. Robustness and flexibility are also necessary, and this was investigated by determining the maximum knee bending angle value (the fully flexed position). There was a significant increase in MKinden parameters compared to previous data. It indicates to us that the training plan and programme effectively improved the movement of the young athletes.

This research could only provide initial results because of the low number of subjects in the groups and the relatively small test movements. The present study had several limitations. The study consisted of a small sample with a similar

background, limiting the ability to apply the results to the general population. Second, the design of the study is not optimal. A large, double-blind, randomised, controlled trial is required to provide reliable evidence. There should have been a follow-up period after the research to study long-term effects.

The presented results showed the efficiency of the classic training methods in strengthening the muscles responsible for the correct low extremity posture. The findings suggest a valuable method for posture improvement in young athletes. Further studies with larger sample sizes, age-range and sports backgrounds are required to clarify the efficacy of good training for flexion problems originating from incorrect posture caused by low muscle tone and strength.

Our Practical Implications

The study raises many practical applications.

It has shown that there is scientific added value in using movement testing in the design of exercise plans.

The importance of analysing the movements of young athletes using movement analysis software

The results obtained can design better training plans to help athletes achieve their goals.

Movement analysis helps both athletes and coaches work more efficiently and safely in the short and long-term.

Acknowledgements

Our team sincerely thanks the White House Foundation for their expertise and assistance throughout all aspects of our study and their help in writing the manuscript. There has been no financial assistance with the project.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, Vol. A247, pp. 529-551, April 1955 (references)

[2] András, K. (2003): The relationship between sport and business - theoretical foundations. Budapest, Budapest University of Economics and Public Administration, No. 34, Workshop Paper

[3] Karácsony, I. (2014) Torna cikkek, könyvrészletek és tanulmányok szöveggyűjtemény, Magyar Torna Szövetség, Budapest pp. 9-24

[4] Karácsony, I (2014) Gymnastics: jumping, barrier, stretching, Hungarian Gymnastics Federation, Budapest pp. 9-83

[5] Honfi, L., Szatmári, L (2011) Theory and practice of sports gymnastics , Dialogue Campus

[6] Turó, A. (2018) The economic impact of sport and sport competitions, Budapest, Dissertation

[7] G. Dallas, A. Mavidis, C. Dallas, G. Dallas, A. Mavidis: Gender differences of high-level gymnasts on postural stability: the effect of ankle sprain injuries, Science of gymnastics journal, Vol. 9, No. 3, 2017

[8] Neulinger, Á. (2007): Social environment and sport consumption - Learned consumption requiring continuous reinforcement. Budapest: Budapesti Corvinus University Doctoral School of Business Administration

[9] Hungarian Gymnastics Federation Magyar Torna Szövetség- www.matsz.hu

[10] J. McClelland. Maes J. FellerK. Webster Single limb landing from different directions in young athletes after anterior cruciate ligament reconstruction Journal of Science and Medicine in Sport Vol. 20, 2017

[11] Laura Hutchison Rolf Scharfbillig Hayley UdenChris Bishop The effect of footwear and foot orthoses on transverse plane knee motion during running – A pilot study Journal of Science and Medicine in Sport, Vol. 18, Issue 6, pp. 748-752, Published online: 13 November 2014

[12] Daniel B. Tait Phillip Newman Nick B. Ball Wayne Spratford What did the ankle say to the knee? Estimating knee dynamics during landing — A systematic review and meta-analysis Journal of Science and Medicine in Sport Vol. 25, Issue 2, pp. 183-191

Published online: 14 August 2021 Astrid J. de VriesInge van den Akker-ScheekRon L. DiercksJohannes ZwerverHenk van der Worp The effect of a patellar strap on knee joint proprioception in healthy participants and athletes with patellar tendinopathy Journal of Science and Medicine in SportVol. 19Issue 4p278–282Published online: 14 May 2015

[13] Carmen E. Quatman Kevin R. Ford Gregory D. Myer Mark V. Paterno Timothy E. Hewett The effects of gender and pubertal status on generalised joint laxity in young athletes Journal of Science and Medicine in SportVol. 11Issue 3p257–263Published in issue: June 2008

[14] Kristin D. Morgan; Cyril J. Donnelly; Jeffrey A. Reinbolt Empirical Based Modeling for the Assessment of Dynamic Knee Stability: Implications for Anterior Cruciate Ligament Injury Risk2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) Year: 2018

[15] M. Ceccarelli; P. Bifulco; M. Bracale Study of the control strategy of the quadriceps muscles in anterior knee pain IEEE Transactions on Rehabilitation Engineering 2000 | Volume: 8, Issue: 3 |

[16] Caitlin N. Teague; Sinan Hersek; Hakan Töreyin; Mindy L. Millard-Stafford; Michael L. Jones; Géza F. Kogler; Michael N. Sawka; Omer T. Inan Novel Methods for Sensing Acoustical Emissions From the Knee for Wearable Joint Health Assessment IEEE Transactions on Biomedical EngineeringYear: 2016 | Volume: 63, Issue: 8 | Journal Article | Publisher: IEEE

[17] Erik E. Stone; Michael Butler; Aaron McRuer; Aaron Gray; Jeffrey Marks; Marjorie Skubic Evaluation of the Microsoft Kinect for screening ACL injury2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)Year: 2013 | Conference Paper | Publisher: IEEE

[18] Begum Yalcin; Hande Argunsah Bayram; Mehmed B. Bayram; Anil Isik Investigation of Biological Feedback Influence on Posture and Knee Biomechanics During Running 2017 21st National Biomedical Engineering Meeting (BIYOMUT) Year: 2017 | Conference Paper | Publisher: IEEE

[19] R. P. Dudhmande; A. M. Rajurkar; V. G. Kottawar Extraction of whole and torn meniscus in MRI images and detection of meniscal tears 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) Year: 2017 | Conference Paper | Publisher: IEEE

[20] Van Mow"Functional Tissue Engineering" 2006 International Conference of the IEEE Engineering in Medicine and Biology Society Year: 2006 | Conference Paper | Publisher: IEEE

[21] M. Guaitolini; L. Truppa; A. M. Sabatini; A. Mannini; C. Castagna Sport-induced fatigue detection in gait parameters using inertial sensors and support vector machines020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob) Year: 2020 | Conference Paper | Publisher: IEEEAuthors: Title, Proceedings of …, place and date of edition, pp.

Emotion Recognition Using Pretrained Deep Neural Networks

Marek Dobeš*, Natália Sabolová**

*CSPV SAV, Karpatská 5, 04001 Košice, Slovak Republic, dobes@saske.sk
**FF UPJŠ, Moyzesova 9, 04059 Košice, Slovak Republic, natalia.sabolova@student.upjs.sk

Abstract: We use pretrained deep neural networks additionally trained using transfer learning to recognise human emotions from photographs. Furthermore, we want to establish what face regions seem to provide the most information for the network. We find that the network performed the best jn the photographs with no occlusions (accuracy of 74.20%) followed by the photographs with occluded eye region (72.01%), faces with covered noses (70.21%) and faces with covered mouths (52.10%).

Keywords: emotion recognition; pretrained networks; deep learning

1 Introduction

Face recognition and facial expressions for emotion detection has been widely studied across different scientific disciplines (including social and behavioural sciences, neuroscience, cognitive psychology, etc.). For example, its profound application can be found in a clinical and neuropsychological research of prosopagnosia that is also called "face blindness" [1], alexithymia – impaired emotion recognition, often in association with autism [2], or social anxiety disorder [3]. Research in this field has been to a great extent influenced by technological growth [4]. Computational advances in programming, computer vision and machine learning allowed researchers to use novel techniques to be able to detect faces and facial emotions from still images or video sequences [5] [6]. There are attempts to build machine learning-based models to simulate cognitive functions produced by the human brain such as face perception and recognition [7]. However, it is essential to note that neurocomputational modeling approach does not completely reflect those phenomena, although it tries to better understand them by studying brain-inspired principles. The core of computer vision research focuses on designing intelligent computer systems that are similarly efficient and accurate (if not more), as our brains [8]. For instance, there is evidence for use of computer vision-based approaches for security purposes [9], such as access control or immigration verification. In forensic settings it is criminal identification or lie detection systems,

where facial micro-expression recognition [10] is processed. Furthermore, computational analysis of facial information presents great potential in robotics and internet of things [11] [12] and as it was mentioned at the beginning, healthcare – for computational diagnosis and assessment of mental or facial diseases and/or machine-assisted interventions [13], etc.

The process of facial expression recognition (FER) can be for a neurotypical human easy and automatic transfer of information about the person that is being observed [14] [15]. There are three main categories of models in the literature to quantify affective facial behaviours: categorical model – according to Ekman and Friesen [16] six prototypical facial emotions: happy, sad, anger, fear, surprise, disgust with a neutral composition of facial muscles are prominent non-verbal communication tools. The second is the dimensional model, where affective dimensions like valence and arousal are central parts of an emotional continuum [17]. The third is Facial Action Coding System (FACS) that reflects facial muscle movement specifics, where action units - AUs emotion-specified facial expressions and their combinations represent various emotions. The computational approach highlights the complexity and challenges of the FER process. Computer model requires large datasets that are divided into training data (reference data for learning), data for testing, and for validation. Firstly, a face needs to be identified and distinguished from the background by pre-processing [18]. Then features extraction and classification can be performed [19]. Depending on how the features are extracted, several extraction techniques can be recognized, such as the holistic approach, one of the most common - appearance-based approach (use of filters) and geometry-based method (shapes and locations of important facial features including eyes, mouth, and nose), colour-based technique, and template-based technique, etc., with its numerous hybrid variations [20].

What makes FER challenging is the nature of databases, annotation and labels objectivity/subjectivity, and computational efficiency [21]. Variability – interpersonal and intrapersonal is important if one aims to work with a computational model in real-life conditions [22]. While laboratory-based controlled images might present consistently taken frontal posed mugshots that are easier for computer to recognise, generalisation and effectiveness of training dataset can be better achieved with more robust, spontaneous real-world scene images from in-the-wild datasets [23]. For instance, differently posed faces (changed angle and lighting of facial features), including various ethnicities and age of people may also present difficulties in accuracy of the FER. Therefore, evolution in datasets was needed and there has been a simultaneous progress in methods and particularly algorithms that have been used: from the handcrafted and shallow learning algorithms to deep learning techniques that are more capable of overcoming these challenges [5].

This article specifically focuses on occlusions, one of the challenges that impair the recognition. It is nothing unfamiliar to speak about facial occlusions during the covid-19 pandemics, when wearing masks (lower part of the faces including mouth

and nose are covered) has become a regular part of our life that may have besides its health benefits social and personal impact on individual's ability to communicate [24]. Also, machine learning can be used to analyse the emotions of people wearing head-mounted displays while playing computer games [25]. There are already datasets consisting of images with facemasks occlusions (MaskedFace-Net, Correctly Masked Face Dataset, Incorrectly Masked Face Dataset) [26]. The present paper focuses on how computational models deal with similar occlusions by using artificially covered areas of faces (mouth, nose area and eyes). Interestingly, a recent study investigated the effect of sunglasses and face masks on perceiver ability to recognize facial expression [27]. They found a support for diagnostic potential of mouth region, with the largest impact of the masks when emotions happy, surprise and disgust prevailed. The eye region, typical for expressions angry and fear was influenced by both, sunglasses, and masks while the identification of fear was the most affected by masks. Eyes and mouth are particularly known diagnostic region for emotions sad and neutral expressions, however, in that study, recognition of neutral expression was better when eyes were covered (assuming overrated importance of eye region information). On the other hand, sad expressions were affected only by the sunglasses (suggesting a critical role of the eye region). In accordance with the literature on occlusion aware facial expression recognition using convolutional neural network (CNN), CNN with attention mechanism (ACNN) inspired by the intuition, automatically identifies the blocked facial patches and diverts attention to the distinctive as well as unobstructed regions in facial image [28]. In addition, evidence from the previous literature listed recognition rates for near frontal faces: 72.6% with eyes occluded, 69.8% for the recognition with mouth-occluded region and with no occlusions 74.28% recognition rate of facial expressions, which points out the significant role of mouth region [29]. Moreover, when the nose area was covered, Vyas and Hablani [30] reported a similar trend in their results: without occlusion 94.82% recognition rate, following eyes occlusion 93.10%, nose 91.37% and mouth in occlusion reached 89.67%.

2 Methods

We used photographs of human faces from the Face Expression Recognition database from Kaggle. Photographs were divided into three categories – happy emotional expression (8989), angry emotional expression (4953 photographs) and neutral emotional expression (6198 photographs). All the photographs were gray scale with the size 48 x 48 pixels.

For later processing we created three sub-categories from each category. We covered with white noise either eyes, nose and mouth of every photograph. We did this using computer processing rather than human processing to avoid possible bias in further learning. For every photograph we created a rectangle of

white noise with the size of 10 x 48 pixels and inserted it into the photograph at the position of the tenth pixel (eyes), 20th pixel (nose) and 30th pixel (mouth). In some cases the rectangle did not fit precisely, but on the average the desired areas of the face were covered. Thus we obtained nine subcategories of photographs. Examples of photographs are displayed in the Figure 1.

Figure 1
Examples of input photographs

To save time and improve efficiency we used transfer learning to create a set of deep neural networks for emotion recognition. We used googlenet network [31]. It is one of the standard pretrained networks used for image recognition and is readily available for application. We removed the last fully connected layer and softmax/output layer. We replaced them with a blank fully connected layer and a softmax/output layer to match the number of classes that we use. We froze the original googlenet layers and trained the network on our data of photographs of faces with emotional expressions. We randomly split our sample so that 70% of the data were used for training and the remaining data for validation. We introduced random augmentation of inputs and trained the network. After training we achieved the validation accuracy of 74.2% with which the network recognised individual emotions.

To investigate on what parts of face does the network concentrate, we applied the grad-cam method [32]. The algorithm will analyse what parts of the picture is the network most sensitive to. In Figure 2 we provide an example of the attentional heat map overlayed on a photograph.

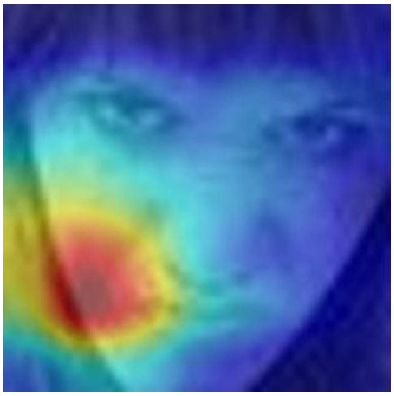


Figure 2
An example of the grad-cam method overlayed on a photograph with angry expression

3 Results

To see how the network will perform on the set of photographs with covered parts of faces, we repeated the transfer learning process with the datasets with covered face parts and achieved the validation accuracy of 72.01 for faces with covered eyes, 70.21 for faces with covered noses and 52.10 for faces with covered mouths.

To analyse how does a network focus when a part of a face is occluded we applied the grad-cam method to the networks trained on occluded data. The results are summarised in Figure 3.

We may notice the trend that presents itself over all three emotional expressions. If the area of eyes and the area of mouth is occluded the network focuses on this area. If the area of nose is occluded, the focus of the network shifts upwards to the area of eyes. This may explain why the network that was trained in the photographs with occluded noses had significantly higher accuracy than the network that was trained on the region of mouth. However, it does not explain the relatively high accuracy of the network that has been trained in the photographs with occluded eyes.

Conclusions

Our findings are in accord with previous research by Goodarzi et al. [29] and Vyas and Hablani [30]. We found that the deep neural network performed best on faces without occlusions and decreased its performance as eyes, nose and mouth were covered.

The later analysis using the grad-cam algorithm may answer why the network that used the photographs with occluded mouth region performed poorly as it focused on the occluded face region that provided little information. However, we were not

able to identify why the network trained on the photographs with occluded eyes performed relatively well, although it also focused on the occluded region.

It would be interesting for further research to establish why the eye region seems to provide more information for the computer network even if the substantial part of it is covered. Also, possible differences between genders could be investigated Bilalpur et al. [33].

We understand that our study has its limitations, such as the use of specific database of training photographs or use of only one pre-trained network. However, we assume that both database and network are relatively robust and should not significantly alter the outcomes should they be trained on different data.

Further research could shed more light on more similarities and differences between how deep neural networks analyse the real-world data [34] and how do the brains of mammals perform similar tasks [35].

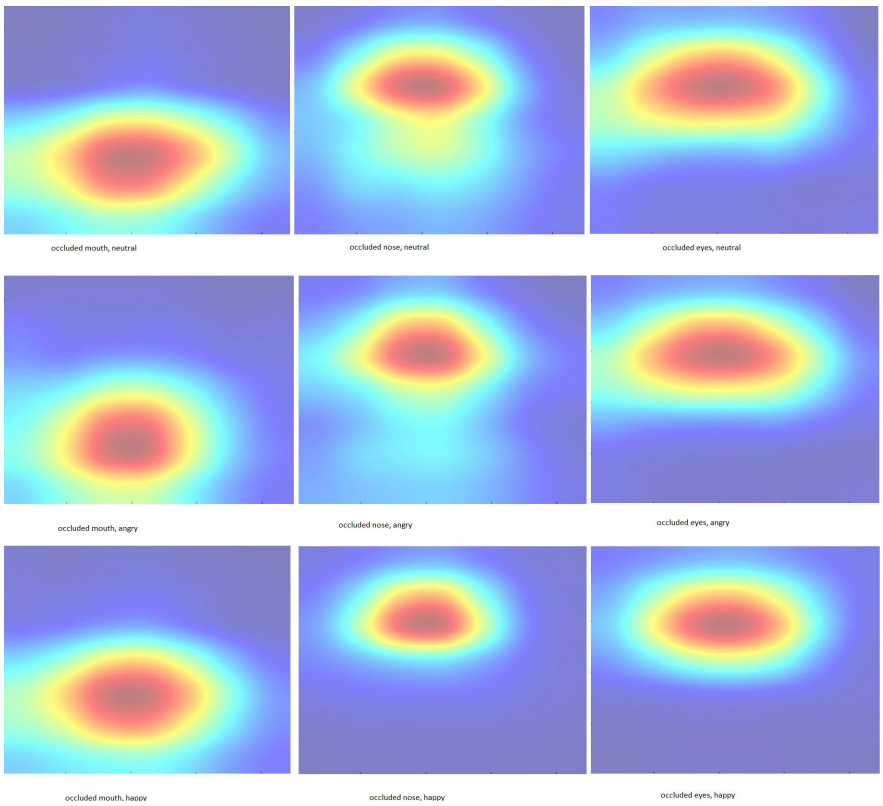


Figure 3
Grad-cam algorithm averaged over 1000 randomly selected samples. Below each sub-figure it is included what part of face is covered and what emotional expression is analysed. Red means the biggest focus, blue the lowest focus.

Acknowledgement

This work was supported by VEGA 2/0043/17 grant.

References

[1] Robotham, R. J., & Starrfelt, R. (2018) Tests of whole upright face processing in prosopagnosia: A literature review. *Neuropsychologia, 121,*106-121, doi:10.1016/j.neuropsychologia.2018.10.018

[2] Cuve, H. C., Castiello, S., Shiferaw, B., Ichijo, E., Catmur, C., & Bird, G. (2021) Alexithymia explains atypical spatio-temporal dynamics of eye gaze in autism. *Cognition, 212,* doi.org/10.31234/osf.io/gzubp

[3] Claudino, R. G.e., de Lima, L. K. S., de Assis, E. D. B., & Torro, N. (2019) Facial expressions and eye tracking in individuals with social anxiety disorder: A systematic review. Psicologia: Reflexão e Crítica, 32, 9, doi.org/10.1186/s41155-019-0121-8

[4] Dodig-Crnkovic, G. (2013) The development of models of computation with advances in technology and natural sciences. *The 6th AISB Symposium on Computing and Philosophy: The Scandal of Computation - What is Computation?* The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

[5] Li, S., & Deng, W. (2020) Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computin.* doi: 10.1109/TAFFC.2020.2981446

[6] Mehendale, N. (2020) Facial emotion recognition using convolutional neural networks (FERC) *SN Applied Sciences, 2*(3) doi:10.1007/s42452-020-2234-1

[7] Martinez, A. M. (2017) Computational models of face perception. *Current Directions in Psychological Science, 26*(3) 263-269, doi:10.1177/0963721417698535

[8] Jonathan, Lim, A., P., Paoline, Kusuma, G. P., & Zahra, A. (2018) Facial Emotion Recognition Using Computer Vision. *IEEE 2018 Indonesian Association for Pattern Recognition International Conference - Jakarta, Indonesia*, 46-50, doi:10.1109/INAPR.2018.8626999

[9] Saad, F. S. A., Lee, M. V., & Basah, S. (2019) Face recognition using machine vision for Security system. *IOP Conference Series: Materials Science and Engineering, 705*, doi:10.1088/1757-899X/705/1/012005

[10] Tamuli, R., Samantara, S., Sarkar, S., & Adhikari, S. (2020) Lie detection using facial micro-expresions. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 4,* (11), 25-30, Retrieved from: https://ijsrcseit.com/paper/CSEIT203115.pdf

[11] Berns, K., & Zafar, Z. (2018) Emotion based human-robot interaction. *MATEC Web of Conferences, 161*, 01001, doi:10.1051/matecconf/201816101001

[12] Balogh, Z., Magdin, M., & Molnár, G. (2019) Motion Detection and Face Recognition using Raspberry Pi, as a Part of, the Internet of Things. *Acta Polytechnica Hungarica*, *16(3)*, 167-185

[13] Leo, M., Carcagni, P., Mazzeo, P. L., Spagnolo, P., Cazzato, D., & Distante, C. (2020) Analysis of facial information for healthcare applications: A survey on computer vision-based approaches. *Information, 11*(3), 128, doi:10.3390/info11030128

[14] Bruce, V., & Young, A. (1986) Understanding face recognition. *British Journal of Psychiatry, 77*, 305-327, doi:10.1111/j.2044-8295.1986.tb02199.x

[15] Darwin, C. & Prodger, P. (1998) *The Expression of the Emotions in Man and Animals*. Oxford University Press

[16] Ekman, P., & Friesen, W. V. (1978) *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press

[17] Russell, J. A. (1980) A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178, doi: 10.1037/h0077714

[18] Revina, I. M., & Emmanuel, W. R. S. (2021) A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences, 3*(6), 619-628, doi: 10.1016/j.jksuci.2018.09.002

[19] Mlakar, U., & Potočnik, B. (2015) Automated facial expression recognition based on histograms of oriented gradient feature vector differences. *Signal, Image and Video Processing, 9* (1), 245-253, doi: 10.1007/s11760-015-0810-4

[20] Fuad, M. T. H., Fime, A. A., Sikder, D., Iftee, M. A. R., Rabbi, J., Al-rakhami, M. S., Gumae, A., Sen, O., Fuad, M., & Islam, M. N. (2021) *Recent advances in deep learning techniques for face recognition*. PsyArXiv. doi: 10.13140/RG.2.2.13444.86402

[21] Martinez, B., & Valstar, M. F. (2016) Advances, challenges, and opportunities in automatic facial expression recognition. In M. Kawulok, E. M. Celebi, & B. Smolka (Eds.), *Advances in Face Detection and Facial Image Analysis*, (pp. 63-100) Springer. doi: 10.1007/978-3-319-25958-1_4

[22] Weber, R., Soladié, C., & Séguier, R. A. (2018) Survey on databases for facial expression analysis. *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 5,* 73-84 doi: 10.5220/0006553900730084

[23] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017) *AffectNet: A database for facial expression, valence, and arousal computing in the wild.* PsyArXiv. doi: 10.1109/TAFFC.2017.2740923

[24] Marini, M., Ansani, A., Paglieri, F., Caruana, F., & Viola, M. (2021) The impact of facemasks on emotion recognition, trust attribution and re-identification. *Scientific Report, 11*, 5577, doi: 10.1038/s41598-021-84806-5

[25] Yong, H., Lee, J., & Choi, J. (2019 March) Emotion recognition in gamers wearing head-mounted display. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 1251-1252) IEEE

[26] Cabani, A., Hammoudi, K., Benhabiles, H., & Melkemi, M. (2021) MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19. *Smart health (Amsterdam, Netherlands), 19*, 100144, doi: 10.1016/j.smhl.2020.100144

[27] Noyes, E., Davis, J. P., Petrov, N., Gray, K. L. H., & Ritchie, K. L. (2021) The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science, 8,* 201169, doi: 10.1098/rsos.201169

[28] Li, Y., Zeng, J., Shan, S., Chen, X. (2018) Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing,* 1–1. doi:10.1109/TIP.2018.2886767

[29] Goodarzi, F., Saripan, M. I., Marhaban, M. H., & Rokhani, F. Z. (2018) Real time facial expression recognition in the presence of rotation and partial occlusions. *Journal of Theoretical and Applied Information Technology, 96* (4), 854-864

[30] Vyas, A., & Hablani, R. (2014) Effect of different occlusion on facial expressions recognition. *International Journal of Engineering Research and Applications, 4* (10), 40-44

[31] Anand, R., Shanthi, T., Nithish, M. S., & Lakshman, S. (2020) Face recognition and classification using GoogleNET architecture. In Soft Computing for Problem Solving (pp. 261-269) Springer, Singapore

[32] Selvaraju, R. R. et al. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Proceedings of the IEEE international conference on computer vision (pp. 618-626)

[33] Bilalpur, M., Kia, S. M., Chua, T. S., & Subramanian, R. (2017, October) Discovering gender differences in facial emotion recognition via implicit behavioral cues. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 119-124) IEEE

[34] Majaj, N. J., & Pelli, D. G. (2018) Deep learning—Using machine learning to study biological vision. *Journal of vision*, *18*(13), 2-2

[35] Storrs, K. R., & Kriegeskorte, N. (2019) Deep learning for cognitive neuroscience. *arXiv preprint arXiv:1903.01458*

Comparison of Direct and Inverse Model-based Disturbance Observer for a Servo Drive System

Nándor Fink, Guo Zenan, Péter Szemes, Péter Korondi

Department of Mechatronics, Faculty of Engineering, University of Debrecen, Ótemető u. 2-4, 4028 Debrecen, Hungary
fink.nandor@eng.unideb.hu; guozenan@eng.unideb.hu; szemespeter@eng.unideb.hu; korondi.peter@eng.unideb.hu

Abstract: The well-known disturbance observer – which is based on the transfer function of the inverse model –, was published almost four decades ago. To practically implement the inverse of the transfer function a filter is added to it, in order to eliminate high frequency disturbance signals. A key step of the design of this Inverse Model Based Disturbance Observer (IMBDO) is the selection of the filter with appropriate parameters. This paper proposes a disturbance observer, which is based on direct model (DMBDO) and it can work without any additional filter. It simplifies the design and the implemented controller code. Discrete time implementations of IMBDO and DMBDO are compared by a simple internet-based servo system in a non-real-time control environment. The effect of non-equidistant sampling is examined.

Keywords: friction compensation; servo drive system; discontinuous control friction compensation

1 Introduction

In the field of motion control, the elimination of disturbance is a key task. Disturbance source could be the effect of external load variation, parameter changes due to aging and friction due to operational conditions. This paper focuses on the non-linear, time invariant effect of friction.

In most cases making the proper model of the disturbance is ineffective or impossible. The disturbance elimination method based on the estimation of the disturbance is widely used. In [1] [2] [3] an observer-based control was applied for rehabilitation purposes, in [4] a long-distance laser positioning was carried out using DOB (Disturbance Observer) method, in [5] [6] the complex and changing environment justified the choice of using a DOB based control, in [7] the lack of sensory data was bridged with a DOB control. It is also used for the motion control of drones [8], for telemanipulation systems [9] [10] [11] and for other disturbance

rejection purposes [12] [13] [14] [15]. A specific type of neural network called long short-term memory (LSTM) recurrent neural network to identify the system behavior in real-time [16]. A method that uses a specific type of filter called a discrete Kalman filter to estimate the behavior of the system while being robust to small disturbances [17]. A combination of two different estimation techniques, an Extended Kalman filter and a Takagi-Sugeno fuzzy observer, to estimate the behavior of a strip winding system [18].

The idea of the estimation of the disturbance has a long history, it was already used in 1983 [19]. A guide for the design of the disturbance observer was stated in [20] [21].

The motivation for using either the inverse or direct modeling approach in disturbance observers is to improve the performance of the control system in the presence of disturbance [41]. These two methods could estimate the disturbance signal accurately and substract it from the measured output to obtain a better estimate of the output. Direct model based disturbance observer is easier to model the system dynamics directly, rather than computing the inverse model. The direct model can provide predictions of the system response, which can eliminate the signal frequency and estimate the disturbance signal. Despite the DOB control being a robust method, it also has its limits [22] [23] [24] [25]. In [26] an improved method has been shown for disturbance rejection (based on DOB), because the to-be-controlled system had too much uncertainty for the traditional DOB. Other problems that exceeds the boundaries of the classical method are stated in [27] [28]. In [29], it describes a study on modeling of observation in cognitive process, it discusses different aspects of the observation with perception and attention. The multicomponent fixed-point iteration is a specific type of mathematical iterative in adaptive control of a univariate second-order system in [30]. A new method for modeling tower crane systems using tensor product-based model transformation is proposed in [31]. A soft computing-based disturbance observer is published in [32] and a sliding mode-based disturbance observer is proposed in [33]. In [34] experiments were performed with an electro-hydraulic actuator controlled by a neural network sliding mode control method, where an Extended State Observer (ESO) was used for noise estimation and rejection with impressive results. Applying DOB for stochastic systems is in [35], and applying for the fractional-order system is in [36]. The ADRC (Active Disturbance Rejection Control) method is examined in [37] [38], and an overview of DOB robust control was presented in [39]. In [40] a signal was decomposed to seven components with different frequency bands with dual-tree complex wavelet transform (DTCWT) method, and each component was noise reduced based on a singular value decomposition (SVD) method. Basic methods and applications of the disturbance observer-based control are presented in [41]. The comparison of PID and sliding mode controller in the teaching field and internet-based control is shown in [42] [43]. Compared with other works, this paper presents a method without a filter to the inverse disturbance observer can eliminate high-frequency interference signals. The motivation of this

paper is to propose a direct model without a filter, simplifying the design steps and the code of the control.

The structure of the paper will be the following. After the introduction, the second section summarizes the mathematical foundations of disturbance observers. The third section presents the experimental setup used in the measurement. The fourth section compares the measurement results. The robustness is discussed in this section too. Finally, the fifth section draws the most important conclusions of the article.

2 Brief Overview of the Mathematical Bases of Disturbance Estimators

2.1 Inverse Model-based Disturbance Observer (IMBDO)

The simplified motor model is described by (1) (2) [41] where the input and output are noted by u and y, respectively. Since the actual real values of the model parameters are not known, the nominal parameters T_n and G_n are used in the inverse model. The d is the disturbance signal at the input and it is estimated by $\hat{d}$.

$$y(s) = \frac{G_m}{1 + sT_M}\big(u(s) + d(s)\big) \tag{1}$$

$$d = \frac{1 + sT_m}{G_m} y - u \qquad \hat{d} = \frac{1 + sT_n}{G_n} y - u \tag{2}$$

Fig. 1 shows the principal operation of a very simple disturbance observer described by (2). The upper part is a simplified diagram of a motor modeled by a mechanical time constant T_m and a gain G_m, while the lower part is the disturbance estimator, which is the inverse of the motor model.

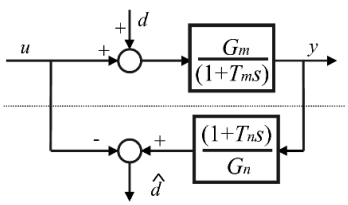


Figure 1
Simple Inverse Model Based Disturbance Observer, IMBDO

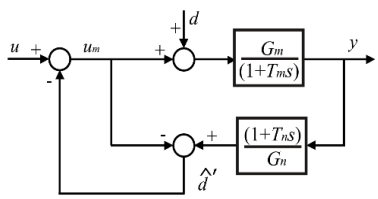


Figure 2
Simple Inverse Model Based Disturbance Observer with feedback

The differences between the nominal and the real system model are considered parameter perturbation, which can be formulated as follows.

$$T_m = T_n + \Delta T_n \qquad\qquad G_m = G_n + \Delta G_n \tag{1}$$

ΔT_n and ΔG_n are limited. It can be seen that the estimated disturbance signal, $\hat{d}$ (indicated in Fig. 1) contains not only the value of the real disturbance signal, but also carries information about the deviation of the parameters (3). The effect of the disturbance signal d can be compensated by feedback on the estimated disturbance signal $\hat{d}$ (see Fig. 2).

After simple calculations, it follows that from the input u to the output signal y the transfer function is $G_n/(1 + T_n s)$ in Figure 2, and y does not depend on the disturbance signal.

In practice, a pure differentiating term can be a serious problem as it amplifies the noise. The effect of noise amplification can be eliminated by the low-pass filter $W(s)$ in the feedback branch, as shown in Figure 3.

The cutoff frequency of $W(s)$ must be much higher than the reciprocal of the dominant time constant T_m of the servo system (so that the useful signal is not filtered out), in the meantime it must be small enough to suppress the measurement noise sufficiently. Of course, a low-pass filter $W(s)$ can be combined with the inverse of the system model and they should be realized together. (see Figure 4).

The absolute value of the transfer function of the disturbance signal to y can be sufficiently small in the lower frequency range if $W(s)$ is properly designed. This method can be generalized, but even then the problem remains that most controlled plants are integrative in nature, so the inverse will be derivative in nature.

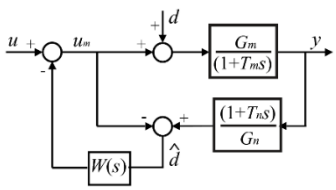


Figure 3
Simple Inverse Model Based Disturbance Observer with feedback and filter

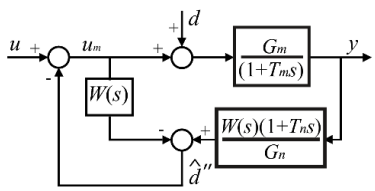


Figure 4
Simple Inverse Model Based Disturbance Observer with feedback and distributed filter

Possible time delay is an additional problem. In the late 1990s, a multitude of articles addressed how to design a $W(s)$ filter that provides optimal operation over a wide range. In summary, the biggest problem with this method is that it is based on the inverse model. The disturbance observer with inverse model has a strong connection in the real world application in [44] [45].

2.2 Direct Model-based Disturbance Observer (DMBDO)

Hereinafter, we again consider the same simple motor model (1) with the same parameter perturbation (3) as in the case of IMBDO. The principle is very simple, the difference between the output signals of the ideal and the perturbed model y_e is fed back to the perturbed model with a sufficiently high gain (Figure 5).

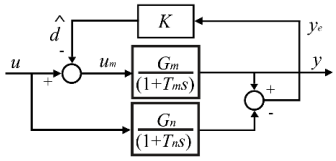


Figure 5
Direct Model Based Disturbance Observer

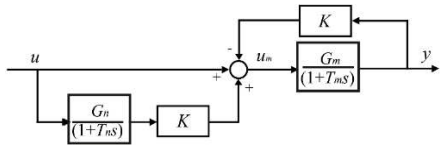


Figure 6
Redrawn direct model based disturbance observer

If K is large enough, then the difference between the two signals y_e must be smallso that the feedback signal $\hat{d}$ is not too large, which means that the feedback signal $\hat{d}$ contains information about the parameter uncertainties ΔT_n and ΔG_n. To calculate the transfer function from u to y, Figure 5 must be redrawn as shown in. According to Figure 6, the transfer function input u and output y using the l'Hospital rule:

$$\lim_{K\to\infty}\left(\left(1+\frac{KG_n}{1+sT_n}\right)\frac{\frac{G_m}{1+sT_m}}{1+K\frac{G_m}{1+sT_m}}\right)=\lim_{K\to\infty}\left(\frac{1+sT_n+G_nK}{1+sT_n}\,\frac{G_m}{sT_m+KG_m+1}\right)=$$

$$=\lim_{K\to\infty}\left(\frac{\frac{1+sT_n}{K}+G_n}{1+sT_n}\,\frac{G_m}{\frac{1+sT_m}{K}+G_m}\right)=\frac{G_n}{1+sT_n} \tag{2}$$

Similarly to Figure 1, add limited disturbance $d(t)$ to the input of the perturbed controlled plant (see Figure 7).

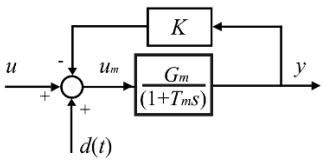


Figure 7
Disturbance rejection part of direct model based disturbance observer

Effect of disturbance $d(t)$ on the output:

$$\lim_{K\to\infty}\left(\frac{\frac{G_m}{1+sT_m}}{1+K\frac{G_m}{1+sT_m}}\right)=0 \tag{3}$$

This can be interpreted as meaning that the feedback signal $\hat{d}$ in Figure 6 carries information about both limited parameter uncertainty and limited input disturbance. The method can be generalized to multi-input, and multi-output systems as well.

2.3 Relay as an Approximation of Infinity Gain

Let us substitute gain K by a simple relay with a relay gain G_r.

$$\hat{d} = G_r sign(y_e) \tag{4}$$

The robustness of the disturbance rejection method is due to its large (ideally infinitely large) gain. This is ensured by a relay in the feedback branch. The gain can be defined by the ratio of the output and input signals. As the error tends to zero the gain tends to be infinite.

$$K = \lim_{y_e \to 0} \frac{G_r sign(y_e)}{y_e} = \infty \tag{5}$$

From the point of view of practical application, the biggest advantage of the relay control law (6) is that it can be easily implemented even with a simple analog operational amplifier, therefore a relatively high switching frequency can be achieved, and it can directly control power electronics semiconductor switching elements. Therefore, it can also be used where the maximum time constant of the controlled plant is relatively small (below ms) and where rapid control is required and the hardware used is not suitable for performing complex calculations between two samples. In contrast, the disadvantage is that it cannot provide for the high gain in the entire state space. There are two conflicting expectations about the value of G_r. The higher the value of G_r, the larger the range of the state space it can provide high gain, but at the same time it can result in a larger chattering around the ideal output signal. The disturbance rejection control method is widely used in the real world application, such as process control system, mechatronics systema and flight control system [41].

3 Experimental System

3.1 The Hardware Description

The system components are listed in Tab. 1 and a photo is presented in Figure 8. The data flow is shown in Figure 9.

Pulses from the incremental encoder that is attached to the permanent magnet DC motor (4) are sent to the computer via the interface circuit (2) and data acquisition card (1). The torque reference signal is transmitted from the computer to the motor controller (5) via the data acquisition card (1) and via the interface circuit (2).

The current signal of the DC motor (4) is sent to the motor controller (5) via the interface circuit (2) and PWM voltage signal returns to the servo motor (4). The source of the power is the ATX power supply (3). The disc (7) is connected to the DC motor shaft via a clutch as a load and the rotation of the disc can be monitored during the measurement using the webcam (6).

Table 1
Parts of the experimental system

Nr.	Element	Type
1	Data acquisition (Counter and DA card)	Advantech USB-4704
2	Interface circuit	ASIC
3	ATX Power supply	ATX 12V GPS-500A8
4	DC Motor with incremental encoder	maxon A-max 110075, MR, Type M, 512 CPT
5	Motor controller unit	maxon
6	Webcamera	USB 2.0 Logitech C525
7	Aluminum disc load	Custom made

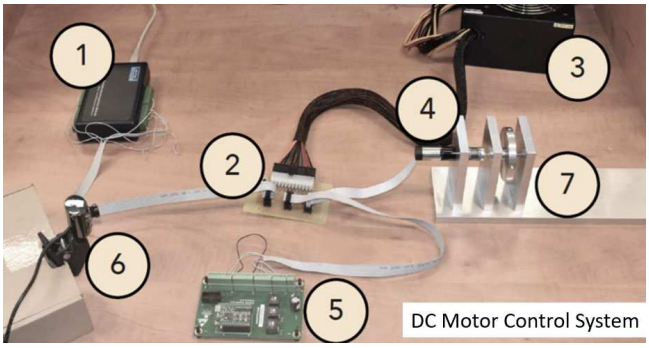


Figure 8
Photo of the experimental setup

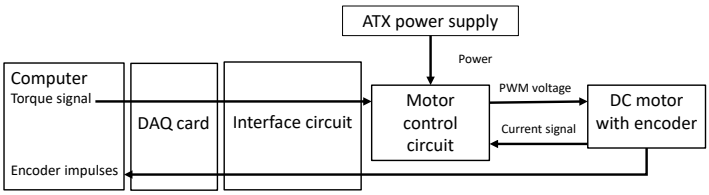


Figure 9
The signal and data flow

3.2 System Equations

3.2.1 Reference Model

The reference model is calculated online as a result of system identification measurement, which is part of the controller application. The discrete time $W_{dm}(z)$ transfer function is calculated first, where the input is the torque $M_m(z)$ and the output is the velocity $\omega(z)$.

$$W_{dm}(z) = \frac{\omega(z)}{M_m(z)} = \frac{0.06}{z - 0.6} T_s = 2\ ms, \tag{6}$$

where T_s is the sampling time of the experimental system. The continuous transfer function $W_{cm}(s)$ can be calculated from (8)

$$W_{cm}(s) = \frac{\omega(s)}{M_m(s)} = \frac{0.15}{0.004s + 1} \tag{7}$$

In contrast to, the simplified model (8) and (9), a detailed model of the servo system is used in the Simulink simulation as described in the following section. The disturbance signal d contains three effects: i.e. friction models, un-modelled dynamics and parameter perturbation.

3.2.2 Model of the Servo System

The goal of the simulation is to tune the parameters of the proposed disturbance observer, a simplified model is used which includes every challenging elements: un-modeled dynamics, the friction model, parameter perturbation and velocity saturation as shown in Figure 10.

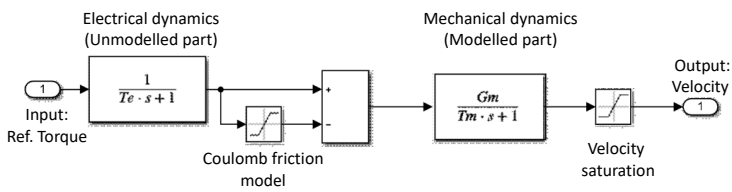


Figure 10
Simulink model of the motor

The input of this system model is the reference torque calculated by the disturbance observer. The torque is measured in perunit %. $T_e = 0.1\ ms$ represents electric dynamics (including the electric time constant of the motor and the current control loop), which is unmodelled from the point of view of the disturbance observer design. The output of unmodelled electric part is the electric torque of the motor. The torque effect of Coulomb friction is modeled by saturation. If the absolute value of the motor is smaller than $T_{rC} = 11\%$ then it is set to zero. The simulated ω_{sim} and the measured ω_{meas} velocities are compared in Figure 11.

The effect of Stribeck friction (what was approximated as Coulomb friction model in Fig. 10), is clearly seen as a time shifted signal in case of measured curve ω_{meas}. The static friction is bigger than the moving friction and there is hysteresis effect as well, which is why unlike ω_{sim}, the first quarter period of ω_{meas} cannot be mirrored to obtain its second quarter. The accurate modeling of the friction is out of the scope of this paper. This simplified (Coulomb) friction model is satisfactory from the point of view of parameter setting. A more detailed friction model of this system is published in [43]. $T_m = 23\ ms$ is the mechanical time constant and $G_m = 0.7$ is the gain. The parameter uncertainties are quite big $\Delta T_m = -19\ ms$ and $\Delta G_m = -0.55$. They are around 80%. The output velocity is saturated $\omega_{max} = 11\ rad/s$ as well. The time constants are read from the data sheet of the motor. The values of G_m, T_{rC} and the offset is identified by measurement shown in Figure 11. The Stribeck effect is ignored in this model since all differences between the real system and the reference model are considered disturbance which we try to reduce.

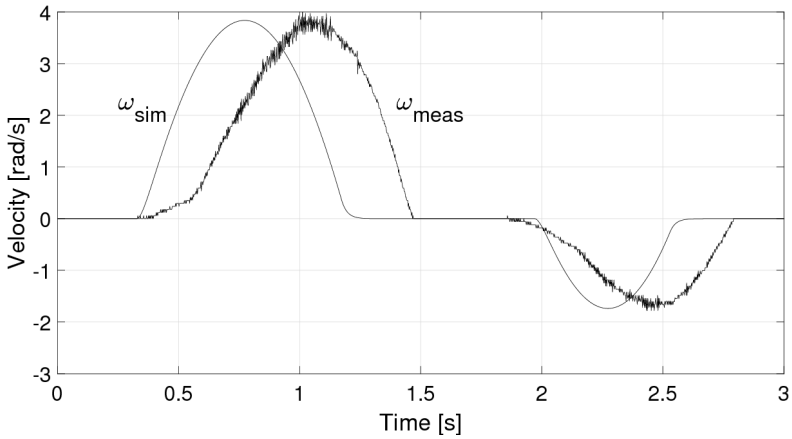


Figure 11
Comparison of simulated and measured velocity

The goal is to change the original dynamic of the servo system and force it to follow the reference model by disturbance rejection method. Let us use the following first-order reference model in the time domain.

$$J_r \dot{\omega}(t) = u_m(t) - B_r \omega(t) \tag{10}$$

where J_r is the inertia and B_r is the constant of the viscous friction of the reference model. u_m is the input electric torque. Parameters of the applied simplified model shown in Figure 10 are calculated from (10).

$$T_m = \frac{J_r}{B_r} \quad G_m = \frac{T_{rmax}}{100 J_r} \tag{11}$$

The simulation parameters are summarised in Table 2.

Table 2
Simulation parameters

Torque reference			
Amplitude=20 and 40		$T_p = 3$	$offset = 1.5$
Time constants			
$T_e = 0.0001$	$T_S = 0.002$	$T_m = 0.002$	$T_1 = T_2 = 0.002$
Gains			
$G_m = 0.7$		$G_r = 0.6$	
Saturations			
$\omega_{max} = 11$		$T_{rC} = 11$	
Parameter perturbation			
$\Delta T_m = -0.019$		$\Delta G_m = -0.55$	

3.2.3 Inverse Model-based Disturbance Estimation and Compensation

The theory of the inverse model based is typically developed in continuous time, as it is based on the mathematical model that describes the system's behavior over time. However, the implementation of the method is often done in discrete time practically, as most real-world systems are controllers using digital controllers that operate on discrete-time signals. The discrete-time version of the idea presented in Figure 4, is implemented. The Simulink model of the system with a disturbance observer is shown in Figure 12.

The sampling method is modeled by a zero-order holder. The lower part of the figure is the disturbance observer. The filters are applied. The first one for the modified input torque signal u_m is the discrete time version of a second order low pass filter given in (12).

$$W_f(s) = \frac{1}{(1 + sT_1)(1 + sT_2)} \tag{12}$$

The filter for the velocity signal is a multiplication of (12) and the inverse of the reference model (9).

$$W_f(s) = \frac{1}{(1 + sT_1)(1 + sT_2)} \tag{13}$$

$$W_{fx}(s) = \frac{0.004s + 1}{0.15(1 + sT_1)(1 + sT_2)} \tag{14}$$

A key step of the design is the selection T_1 and T_2. The time constants of the filter are set between the sampling time and the mechanical time constant.

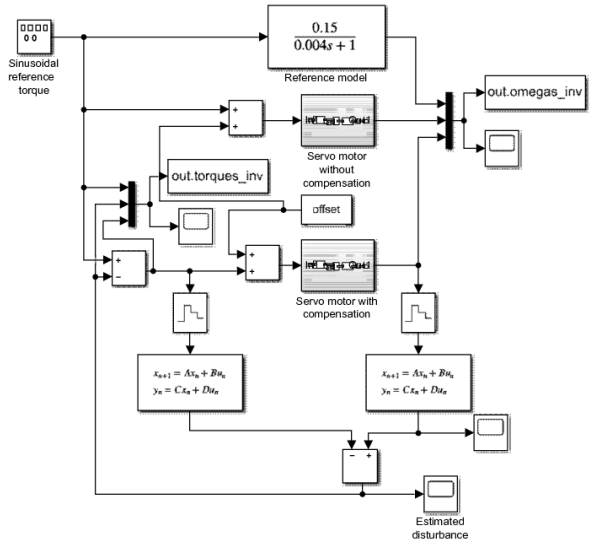


Figure 12
Simulink model of a common IMBDO

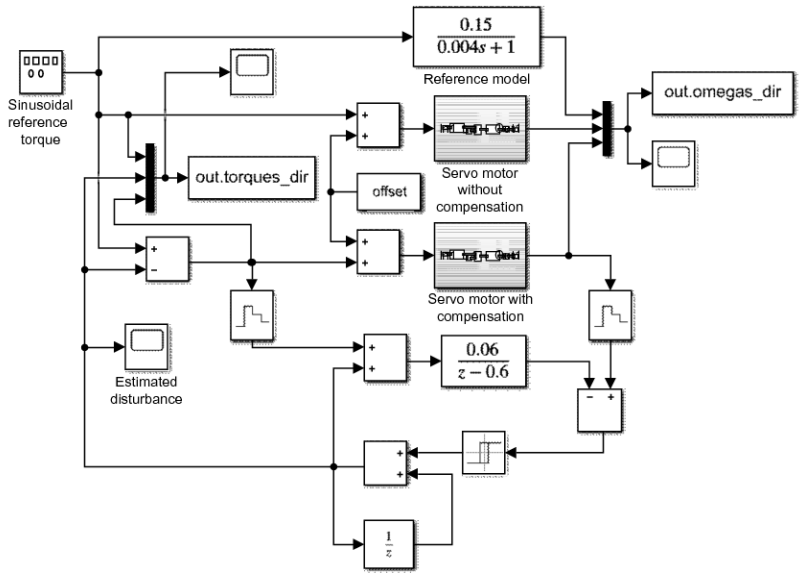


Figure 13
Simulink model of the proposed DMBDO

3.2.4 Direct Model-based Disturbance Estimation and Compensation

Since the estimated disturbance signal $\hat{d}$ is the output of the relay it has a high frequency component, which might cause chattering in the ω. On one hand, we need a small relay gain G_r as possible to reduce the chattering effect, on the other hand G_r must be big enough to cover the disturbance ($G_r > \hat{d}_{max}$). If $\hat{d}(t)$ is a continuous and limited function than the change of disturbance in one sampling period $\Delta\hat{d}(t)$ is also limited and only this change must be covered by the rely upon the discrete time application.

$$\hat{d}[k] = \hat{d}[k-1] + G_r sign(y_e) \tag{15}$$

3.3 Simulation Results

The main challenge is friction compensation. Two sets of measurement results are presented. The u inputs are sinusoidal in all cases in this paper. From the physical point of view, u is the torque reference signal, which is given in perunit percentage. First, the amplitude of u is 20%. The control loop of the armature current is realized by an operational amplifier, which has an uncompensated offset. The actual acceleration torque is shifted in the positive direction vertically. The offset of the operational amplifier causes asymmetrical non-compensated velocity ω_{nc} as shown in Figure 14, when the disturbance observer is switched off.

When the sinusoidal input torque u is smaller than the static friction torque, the motor cannot move ω_{nc} (due to Stribeck friction). The real motor movement starts around 0.1 s and the velocity increases faster than the ideal ω_{id} velocity. The motor stops again when the motion friction torque becomes bigger than the real motor torque.

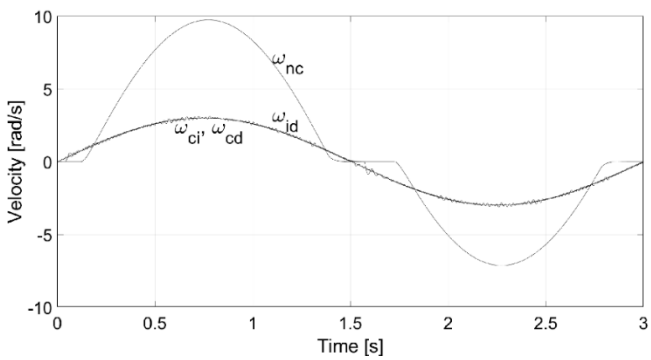


Figure 14
Velocities (u_{max}=20%)

The three velocities (calculated by the ideal model ω_{id}, by IMBDO ω_{ci} and by DMBDO ω_{cd}) are seems quite similar in Figure 14. To show the differences, the phenomenon around zero-crossing is magnified in Figure 15.

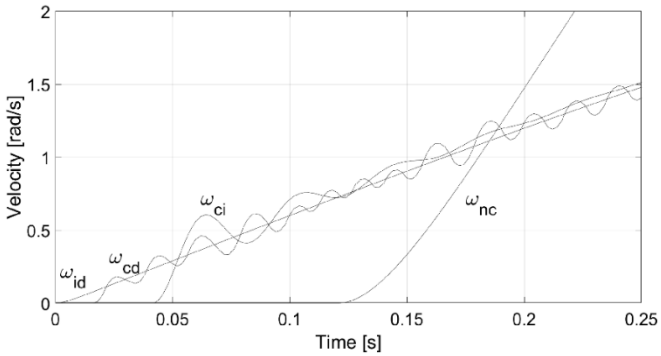


Figure 15
Velocities around zero-crossing (u_{max}=20%)

The response of DMBDO is faster but ω_{cd} has a continuous oscillation around ω_{id}. In case of IMBDO, ω_{ci} there is a kind of overshoot around zero-crossing and it has a very small constant error. The faster response of DMBDO is visible in Figure 16 and Figure 17 as well.

The Coulomb friction (as an approximation of Stribeck effect) changes its sign at zero-crossing of velocity signal so the disturbance observer must follow this change. The estimated disturbance of DMBDO $\hat{d}(t)_{cd}$ follows the Coulomb friction faster than $\hat{d}(t)_{ci}$ of IMBDO. Because of the offset the positive and negative half periods are asymmetrical. It explains the shape of the estimated disturbance signals $\hat{d}(t)_{ci}$ and $\hat{d}(t)_{cd}$ in Figure 16 and Figure 17.

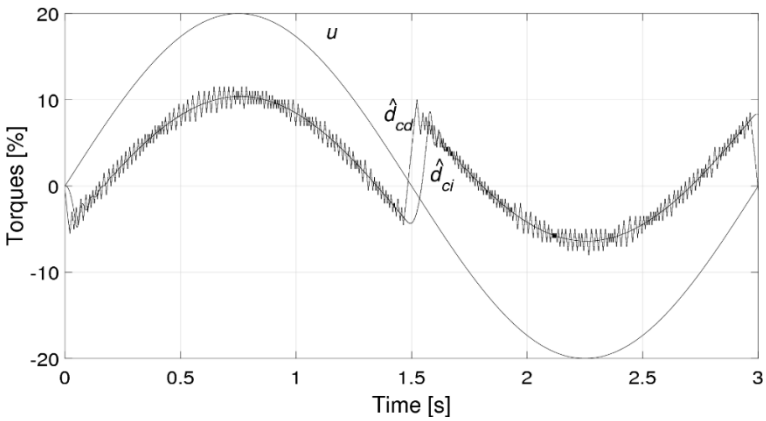


Figure 16
Torques (u_{max}=20%)

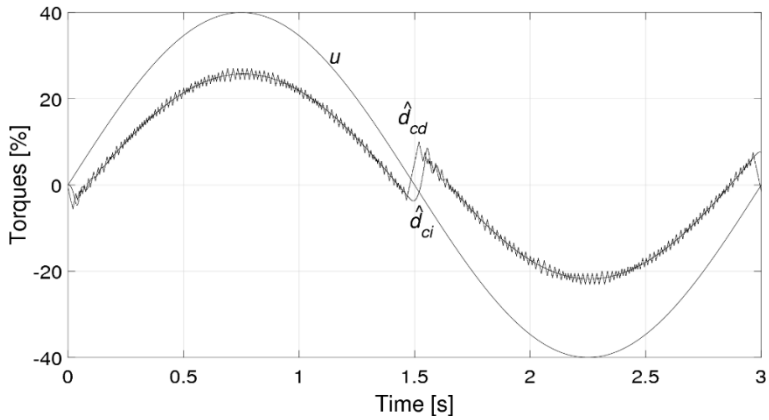


Figure 17
Torques (u_{max}=40%)

If the amplitude of the reference torque is 40% the non-compensated velocity ω_{nc} is saturated at the maximum velocity of the motor as shown in Figure 18. Because of the offset the saturation is longer in the positive half period than that in the negative half period.

4 Measurement Results

The experimental results are quite similar to that of simulation in the previous chapter. Since the real motor shaft velocity is calculated from the impulses of the incremental encoder it is quite noisy. First $u_m = 20\%$ is set. In most cases the measured noise is bigger than the difference between the two methods as shown in Figure 19 and Figure 20.

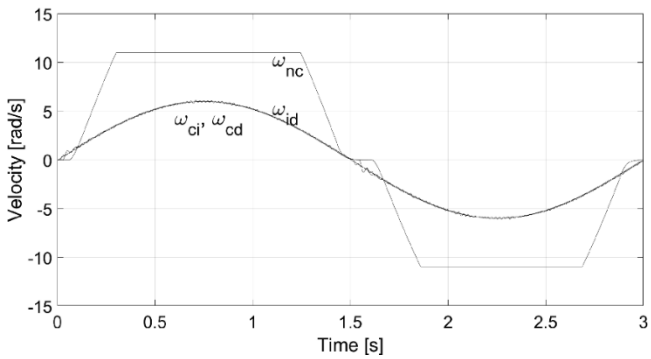


Figure 18
Velocities (u_{max}=40%)

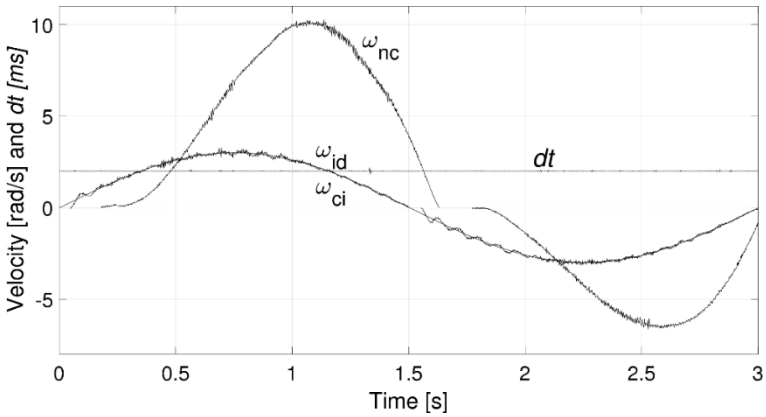


Figure 19
Velocities of IMBDO (u_{max}=20%)

The main challenge is the non-equidistant sampling. The applied operating system is a general purpose, not a real-time operating system, and accurate sampling cannot be ensured. Usually, the sampling delay is less than 0.2 ms, but sometimes the time difference between two samplings can be 20 times bigger than the intended $T_s = 2\ ms$. Figure 23 shows one of the worst periods (peaks in the figure) detected during the measurements. The distribution of length is shown in Figure 24.

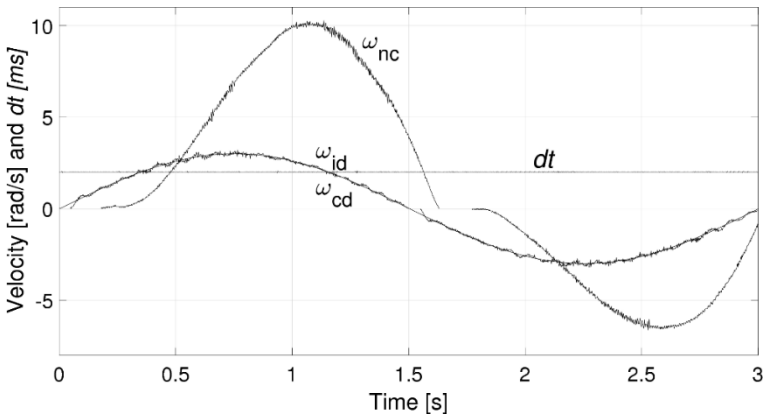


Figure 20
Velocities of DMBDO (u_{max}=20%)

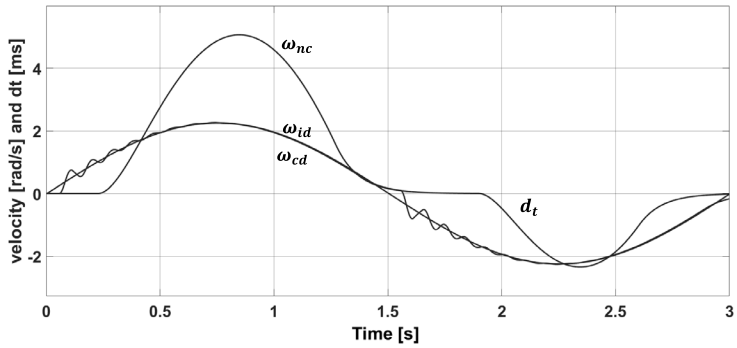


Figure 21
Velocities of IMBDO (u_{max} =15%)

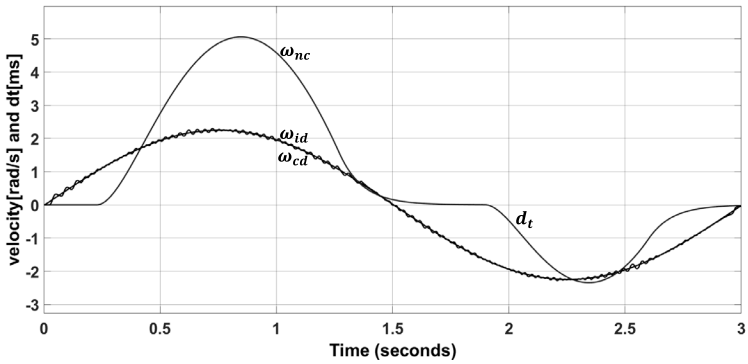


Figure 22
Velocities of DMBDO (u_{max} =15%)

Figure 21 and Figure 22 show another velocity comparison of IMBDO and DMBDO with the u_{max} =15%. It depicts that the DMBDO without the filter has nearly the same performance compared with the IMBDO. The Servo motor with compensation reduces the chattering with the u_{max} increasing. DMBDO has the same trajectory but better performance than IMBDO under the same parameters.

Figure 19 and Figure 21 depict that the velocity with the compensation could converge to the reference velocity in a finite time. The velocity response is steady, but when the velocity crosses the horizontal line it has an oscillation. Compared with the IMBDO, the velocity with compensation in DMBDO converges to the reference in a finite time, and the velocity in DMBDO has a more steady performance than the velocity in IMBDO (Figure 20 and Figure 22).

Time-variant sampling periods cause oscillation, which has a bigger amplitude in case of IMBDO, ω_{ci} than in case of DMBDO, ω_{cd}. The system operated for 5 minutes (100 periods) and the worst periods are selected in both cases. The results are shown in Figure 25 and Figure 26. As it was anticipated from Figure 15, the performance of DMBDO is a little bit better than that of IMBDO.

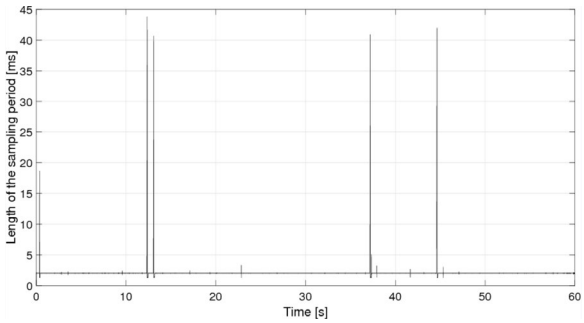


Figure 23
The length of the sample period

Finally, the case of $u_m = 40\%$ is measured as shown in Fig. 25 and Fig. 26. The Stribeck effect (as a time shifted signal) is clearly seen again. The zero-crossing is plotted in Figure 29. Similarly to Figure 15, the delay of the movement is shorter and the amplitude of oscillation is smaller in the case of DMBDO, ω_{cd}.

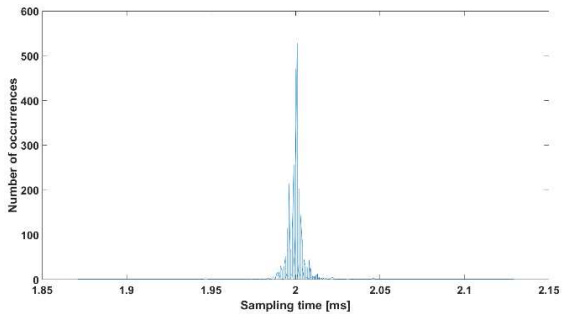


Figure 24
Distribution of length of the sampling periods

The theory of the direct disturbance observer and inverse disturbance observer is usually developed in continuous time because they are based on a mathematical model describing the behavior of the system over time. But the real implementation of the method is usually performed in discrete time because most real-world systems are controlled using digital controllers operating on discrete-time signals.

5 Control Code Robustness Analysis

IMBDO and DMBDO have a low sensitivity to the variations in the parameters of the controller plant and disturbance as a robust control method. Since the input signal u is determined by a relay it is pulse width modulated. The mean value of the torque u can be influenced by changing the duty ratio. The reason for the robustness is that the switching frequency of the torque reference signal is 500 Hz.

The response of the system to a change in load can be less than 2 ms. Let us compare the sampling time with the system time constant. According to the motor data sheet, the mechanical time constant is $T_m = 23\ ms$, which is an order of magnitude larger than the sampling time, so that the controller can react faster than the change would appear in the motion state. The other important question is how long it takes for the digitally calculated input signal u to convert to physical torque. The non-modeled dynamics of the armature coil and current control loop, which can be characterized by the electric time constant $T_e = 0.1\ ms$, plays a decisive role in this. It can be stated that the input signal used in our model changes at least one order of magnitude faster than the sampling time. Comparing Fig. 19 to Figure 20 and Figure 27 to Figure 28, it can be seen that there is no significant difference between the ideal ω_{id}, the compensated ω_{ci} and ω_{cd} velocity signals if the actual sampling periods are close to $T_s = 2ms$.

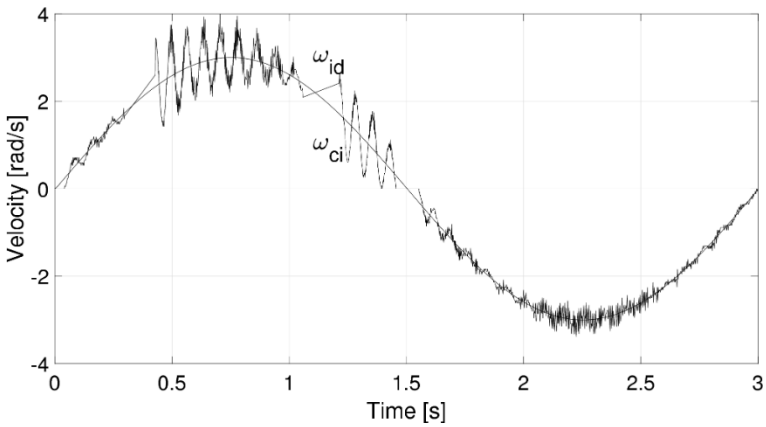


Figure 25
Velocities of IMBDO (u_{max}=20%)

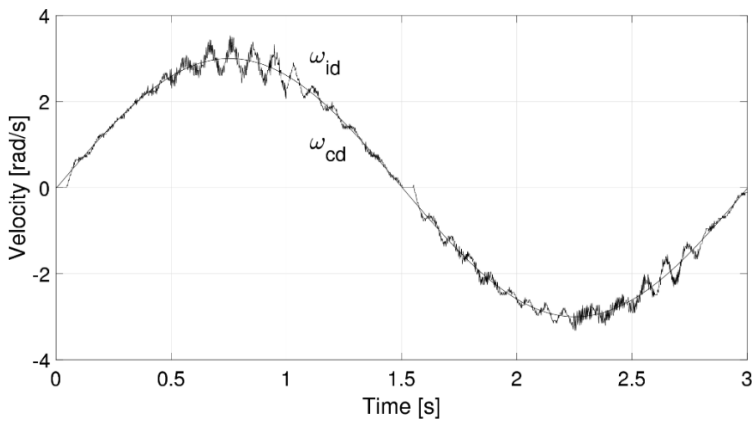


Figure 26
Velocities of DMBDO (u_{max}=20%)

Table 3
Comparison of time constants of the system

Times		
Actuator delay	***Sampling time***	***Dominant time constant***
$T_e = 0.1\ ms$	$T_s = 2\ ms$	$T_n = 23\ ms$

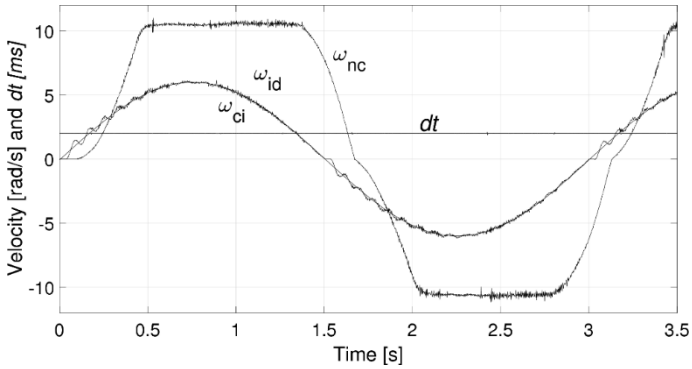


Figure 27
Velocities of IMBDO (u_{max}=40%)

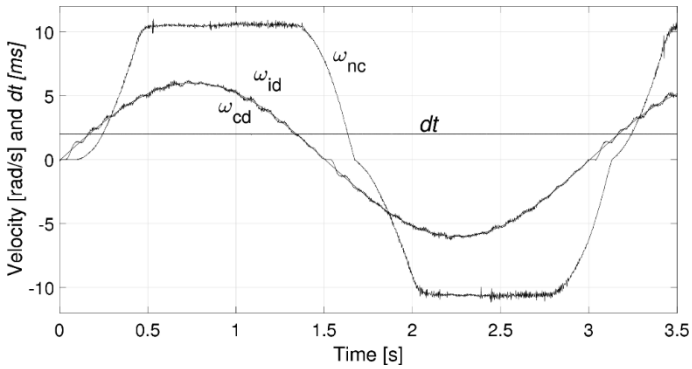


Figure 28
Velocities of DMBDO (u_{max}=40%)

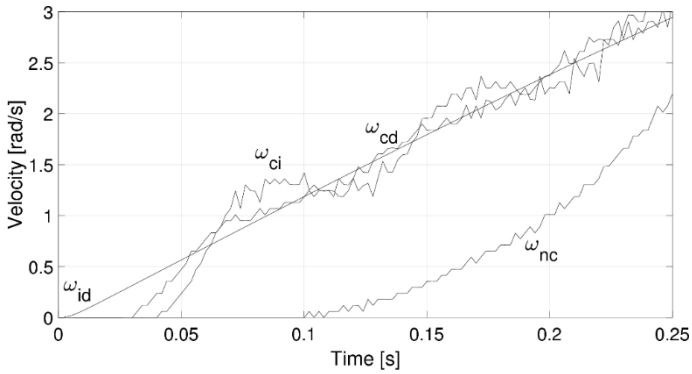


Figure 29
Zero-crossing of DMBDO and IMDBO (umax=40%)

Conclusion

A discrete time Inverse Model Based Disturbance Observer (IMBDO) and a Direct Model Based Disturbance Observer (DMBDO) are implemented as a simplified, non-real-time controller-based servo system. The performance of the two types of disturbance observers is compared by simulation and experiments. The DMBDO has some advantages to IMBDO in this particular, servo control application. The implementation of DMBDO is simpler since the filter is not needed, in case of model inversion. The design is easier and the resulting controller code is also simpler and more robust. Case of practical code implementation the execution time of the controller (also the computational complexity of the) algorithm is shorter, which can be important in case of low cost (i.e. microcontroller) applications. DMBDO can respond faster to the sudden change of the disturbance and it is less sensitive to non-equidistant sampling. It means it can be used in a non real-time environment, as especially in the education system.

Acknowledgement

The project TKP2020-NKA-04 has been implemented with the support provided by the National Research, Development and Innovation Fund of Hungary, financed under the 2020-4.1.1-TKP2020 funding scheme.

The authors wish to thank the support of the Hungarian Research Fund (OTKA K143595).

References

[1] Mohammadi A. and Dallali H. Disturbance observer applications in rehabilitation robotics: an overview., 2020, pp. 113-133, doi:10.1016/B978-0-12-817450-0.00005-5

[2] Saadatzi M. N., Poshtan J., and Saadatzi M. S. Application of mimo disturbance observer to control of an electric wheelchair using nsga-ii. Journal of Medical Signals and Sensors, 2011, 1(2) pp. 122-129

[3] Brahmi B., Driscoll M., El Bojairami I. K., Saad M., and Brahmi A. Novel adaptive impedance control for exoskeleton robot for rehabilitation using a nonlinear time-delay disturbance observer. ISA Transactions, 2021, pp. 108:381-392, doi:10.1016/j.isatra.2020.08.036

[4] Deng J., Xue W., Zho X., and Mao J. On disturbance rejection control for inertial stabilization of long-distance laser positioning with movable platform. MEASUREMENT AND CONTROL, 2020, 53(7-8) pp. 1203-1217, doi:10.1177/0020294020935492

[5] He L., Xiong C.-H., Zhang Q.-H., Chen W.-B., Fu C.-L., and Lee K.-M. A backpack minimizing the vertical acceleration of the load improves the economy of human walking. IEEE Trans Neural Syst Rehabil Eng., 2020, 28(9) pp. 1994-2004, doi:10.1109/TNSRE.2020.3011974

[6] Ming L. and Heping L. Disturbance observer-based variable structure control on the working attitude balance mechanism of underwater robot., 2009, 1 pp. 232-237, doi:10.1109/ICCET.2009.202

[7] Surapong N. and Mitsantisuk C. Position and force control of the scara robot based on disturbance observer. Procedia Computer Science, 2016, 86 pp. 116-119, doi:10.1016/j.procs.2016.05.029

[8] Sasaki K. and Yang Z.-J. Disturbance observer-based control of uavs with prescribed performance. INTERNATIONAL JOURNAL OF SYSTEMS SCIENCE, 2020, 51(5) pp. 939-957, doi:10.1080/00207721.2020.1746436

[9] Sun D., Liao Q., Stoyanov T., Kiselev A., and Loutfi A. Bilateral telerobotic system using type-2 fuzzy neural network based moving horizon estimation force observer for enhancement of environmental force compliance and human perception. Automatica, 2019, 106 pp. 358-373, doi:10.1016/j.automatica.2019.04.033

[10] Singh H., Panzirsch M., and Ryu J.-H. Preserving the physical coupling in teleoperation despite time delay through observerbased gradient control. IFAC-PapersOnLine, 2019, 52(18) pp. 25-30, doi:10.1016/j.ifacol.2019.12.201

[11] Coutinho F. and Cortes˜ao R. Online stiffness estimation for robotic tasks with force observers. Control Engineering Practice, 2014, 24 pp. 92-105, doi:10.1016/j.conengprac.2013.11.002

[12] Sakow M. Novel robust disturbance observer. ISA Transactions, 2020, 104 pp. 255-277, doi:10.1016/j.isatra.2020.05.023

[13] Tao Z., Shu W., Lei W., and Yang Z. Active disturbance rejection control of servo systems with friction., 2009, pp. 558-561, doi:10.1109/CASE.2009.119

[14] Ding K. and Zhu Q. Extended dissipative anti-disturbance control for delayed switched singular semi-markovian jump systems with multi-

disturbance via disturbance observer. Automatica, 2021, p. 128, doi:10.1016/j.automatica.2021.109556

[15] Wang L., Freeman C. T., Rogers E., and Young P. Disturbance observer based repetitive control system with non-minimal state space realization and anti-windup mechanism. IFAC-PapersOnLine, 2020, 53(2) pp. 1505-1510, doi:10.1016/j.ifacol.2020.12.2008

[16] Christudas, F, and Angeline V. D. "System identification using long short term memory recurrent neural networks for real time conical tank system." Rom. J. Inf. Sci. Technol 23 (2020): 57-77

[17] Volovyk, A. Yu, V. M. Kychak, and D. V. Havrilov. "Discrete Kalman Filter Invariant to Perturbations." Acta Polytechnica Hungarica 18, No. 10 (2021): 21-41

[18] Szedlak-Stinean, A. I., Precup, R. E., Petriu, E. M., Roman, R. C., Hedrea, E. L. and Bojan-Dragos, C. A., 2022, Extended Kalman filter and Takagi-Sugeno fuzzy observer for a strip winding system. Expert Systems with Applications, 208, p.118215

[19] Ohishi, K. and Ohnishi, K.and Miyachi, K. Torque-speed regulation of dc motor based on load torque estimation. Proc. IEEJ IPEC-TOKYO, 1983, 2 pp. 1209-1216

[20] Sariyildiz E. and Ohnishi K. A guide to design disturbance observer. J. Dyn. Sys., Meas., Control., 2013, 136(2) pp. 021011-1-021011-10, doi:10.1115/1.4025801

[21] Chen W.-H. Disturbance observer based control for nonlinear systems. IEEE/ASME TRANSACTIONS ON MECHATRONICS, 2004, 9(4) pp. 706-710, doi:10.1109/TMECH.2004.839034

[22] Sariyildiz E. and Ohnishi K. Analysis the robustness of control systems based on disturbance observer. International Journal of Control, 2013, 86(10) pp. 1733-1743, doi:10.1080/00207179.2013.795663

[23] Shim H. and Jo N. H. An almost necessary and sufficient condition for robust stability of closed-loop systems with disturbance observer. Automatica, 2009, 45(1) pp. 296-299, doi:10.1016/j.automatica.2008.10.009

[24] Choi Y., Yang K., Chung W. K., Kim H. R., and Suh I. H. On the robustness and performance of disturbance observers for second-order systems. IEEE TRANSACTIONS ON AUTOMATIC CONTROL, 2003, 48(2) pp. 315-320, doi:10.1109/TAC.2002.808491

[25] Oboe R. How disturbance observer changed my life., 2018, pp. 13-20, doi:10.1109/AMC.2019.8371055

[26] Qiu Z., Hu S., and Liang X. Disturbance observer based adaptive model predictive control for uncalibrated visual servoing in constrained

environments. ISA Transactions, 2020, pp. 106:40-50, doi:10.1016/j.isatra.2020.06.013

[27] Chen J., Sun R., and Zhu B. Disturbance observer-based control for small nonlinear uav systems with transient performance constraint. Aerospace Science and Technology, 2020, p.105, doi:10.1016/j.ast.2020.106028

[28] Wen X., Sun Y., Wang J., and Yao X. Rejection of timevarying frequency sinusoidal disturbance using refined observer for a class of uncertain systems. ISA Transactions, 2020, 100 pp. 136-144, doi:10.1016/j.isatra.2019.11.039

[29] Pozna, C. and Precup, R. E., 2012, Aspects concerning the observation process modelling in the framework of cognition processes. Acta Polytechnica Hungarica, 9(1), pp.203-223

[30] Redjimi, H. and Tar, J. K., 2021, Multiple Components Fixed Point Iteration in the Adaptive Control of Single Variable 2nd Order Systems. Acta Polytechnica Hungarica, 18(9), pp.69-84

[31] Hedrea, E. L., Precup, R. E., Roman, R. C. and Petriu, E. M., 2021, Tensor product-based model transformation approach to tower crane systems modeling. asian journal of control, 23(3), pp.1313-1323

[32] Raul-Cristian Roman, Radu-Emil Precup, and Emil Petriu. Hybrid data-driven fuzzy active disturbance rejection control for tower crane systems. EUROPEAN JOURNAL OF CONTROL, 2020, 58 pp. 373-387, doi:10.1016/j.ejcon.2020.08.001

[33] P. Korondi, K. D. Young, and H. Hashimoto. Sliding mode based disturbance observer for motion control., 1998, pp. 1926-1927

[34] Wending L., Guanglin S., Chun Z., Hongyu L., and Junyong F. RBF Neural Network Sliding Mode Control Method Based on Backstepping for an Electro-hydraulic Actuator. Strojniˇski vestnik - Journal of Mechanical Engineering, 2020, 66(12) pp. 697-708, doi:10.5545/sv-jme.2020.6866

[35] Wei X.-J.,Wu Z.-J., and Karimi H. R. Disturbance observer-based disturbance attenuation control for a class of stochastic systems. Automatica, 2016, 63 pp. 21-25, doi:10.1016/j.automatica.2015.10.019

[36] Liu R., She J., Zhu F., and Nie Z. Robust disturbance rejection for a fractional-order system based on equivalent-input-disturbance approach. Science China Information Sciences, 2018, 61 pp. 1-12, doi:10.1007/s11432-017-9368

[37] Han J. From pid to active disturbance rejection control. IEEE Trans. Ind. Electron., 2009, 56(3) pp. 900-906, doi:10.1109/TIE.2008.2011621

[38] Cortes-Romero J., Delgado-Aguilera E., and Jimenez-Triana A. Robust fractional active disturbance rejection control: A unified approach. ISA Transactions, 2020, 107 pp. 63-77, doi:10.1016/j.isatra.2020.08.003

[39] Sariyildiz E., Obor R., Ohnishi K., and Fellow L. Disturbance observer based robust control and its applications: 35th anniversary overview. IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, 2020, 67(3) pp. 2042-2053, doi:10.1109/TIE.2019.2903752

[40] Xihui C., Gang C., Ning L., Xinhui S., and Wei L. Research on a Noise Reduction Method Based on DTCWT, and the cyclic singular energy difference spectrum. Strojniˇski vestnik - Journal of Mechanical Engineering, 2020, 66(10) pp. 613-626, doi:10.5545/sv-jme.2020.6704

[41] Li, Shihua, Jun Yang, Wen-Hua Chen, and Xisong Chen. Disturbance observer-based control: methods and applications. CRC press, 2014

[42] Zenan, G., Szemes, P.T. and Korondi, P., 2022, October. Bond graph-based teaching method to enhance the synergy of mechatronics in Lab VIEW. In 2022 IEEE 9th International Conference on e-Learning in Industrial Electronics (ICELIE) (pp. 1-6) IEEE

[43] Róbert, M., Balázs, S., Dobák, D. and Tamás, S. P., 2022, September. Internet Based Control of a Servo Motor with a Sliding Mode Based Observer for Chattering Reduction. In 2022 IEEE 20th International Power Electronics and Motion Control Conference (PEMC) (pp. 209-215) IEEE

[44] Güvenç, B.A., Güvenç, L. and Karaman, S., 2010, Robust MIMO disturbance observer analysis and design with application to active car steering. International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal, 20(8), pp. 873-891

[45] Chan, S. P., 1995, A disturbance observer for robot manipulators with application to electronic components assembly. IEEE Transactions on Industrial Electronics, 42(5), pp. 487-493

[46] Gróf, P ; Takarics, B ; Petres, Z ; Korondi, P., Tensor product model type polytopic decomposition of a pneumatic system with friction phenomena taken into account., In: IEEE - IEEE (szerk.) 2010 IEEE 8th International Symposium on Applied Machine Intelligence and Informatics New York (NY), Amerikai Egyesült Államok: IEEE (2010) pp. 153-158, Paper: 5423747, 6 p. doi: 10.1109%2FSAMI.2010.5423747

Study Preferences in Higher Education

Imola Cseh Papp[1], Péter Karácsony[2,*], Tímea Juhász[3]

[1] Institute of Research on Adult Education and Knowledge Management, ELTE Eötvös Loránd University, Hungary, 1075 Budapest, Kazinczy u. 23-27, E-mail: papp.imola@ppk.elte.hu

[2] Keleti Károly Faculty of Business and Management, Óbuda University, Hungary, Tavaszmező u. 15-17, 1086 Budapest, E-mail: karacsony.peter@uni-obuda.hu

J. Selye University, Faculty of Economics and Informatics, Bratislavská cesta 3322, 94501 Komárno, Slovakia, E-mail: karacsonyp@ujs.sk

[3] Faculty of International Management and Business, Budapest Business School, Diósy L. 22-24, 1165 Budapest, Hungary, E-mail: juhasz.timea@uni-bge.hu

[*]Corresponding author

Abstract: It is a well-known fact that in modern societies educational systems are inseparably tied to the labour market. Individuals invest in their own economic potential through education and training; therefore, they increase their labour market value and their future earnings, according to human capital theory. The primary objective of this paper was to examine the decisions made by young people studying in Hungary on continuing their education and selecting the appropriate institution and degree programme, as well as their expectations for their degree. The examination was conducted by means of a questionnaire in seven Hungarian higher education institutions. The findings revealed that students' decisions on going on to higher education are largely influenced by the prospect of a better job in the future, as well as the likelihood of career and financial success.

Keywords: youth; higher education; labour market; employability; decisions; expectations

1 Introduction

Personal decisions on education are influenced by a variety of factors that can be, on the one hand, objective and not necessarily based on rational considerations, on the other hand. Future labour market opportunities, position in the social hierarchy, direct monetary expenditure and anticipated gains are considered when making objective decisions; for the latter, the importance of social environment and culture is taken into account [1].

It is important for young people to make the right job decisions in today's globalized world. They will have a better chance of coping with the changing labour market demands if they continue their education and obtain professional qualification. Choosing a profession is a lengthy and complex phase that is affected by a variety of factors [2]. The first stage involves secondary school selection in which age-related considerations such as the immediate environment, family and friends play a crucial role. However, it is the school that plays the crucial role in preparation. Ritoókné emphasizes [3] the value of experiential content, pointing out the basic distinction between being able to do something well and doing it for enjoyment, even though it is not necessary. This desire for experience differs from person to person and from profession to career, but it can be traced back especially well in the case of those who have entered their career through family or other compulsions rather than their own determination. In the literature it is Lange, who emphasises the importance of a rational judgment situation, which means that a person's professional experience is extensive, i.e., they have gained considerable theoretical and experiential knowledge [4]. It is also stressed that the autonomous personality is capable of self-reliance and accountability. They are characterized by strong self-confidence, process the information obtained and take into account the opinions of others, but make decision based on their own preferences.

Galasi measured labour market success solely in terms of earnings [5]. Hetesi and Kürtösi identified many other criteria as requirements for success, including finding a job, having a position that is ideal for the qualification, a short job search period, and a satisfactory wage [6].

The paper aims to address the question of what factors affect enrolment in higher education based on all of this. To answer this question, the authors apply an empirical approach based on the international literature.

2 Theoretical Overview

Motivational theories explain what influences a person's behaviour in different situations. Hofmesiter-Tóth considered the desire to learn to be a secondary need [7]. During their studies, people learn a career and earn a degree, which is a basic prerequisite for jobs and hence one of the requirements for establishing security [8].

Becker and Hecken, stated while examining secondary school students that stakeholders consider the scale of investment in education before deciding on further education, i.e., they aim to estimate how the cost/benefit ratio evolves in this situation [9]. A 2016 study confirmed the correlation between the labour market and an individual's interest in learning, finding that students' higher

education is driven by the acquisition of a wider range of new skills, knowledge and abilities [10]. Furthermore, they also discover that it helps them advance in their careers. The decision making skill model can also be used to define continuing education as a buying decision. According to Törőcsik [11], each sub-decision shows a change toward the child or parent's supremacy, as well as the presence of shared decisions on certain issues. The choice of the study programme and the admission strategy are dominated by the prospective student's opinion while the parent is competent in financial matters. Thus, parents leave room for individual ambitions and professional interests, but since they finance the studies, they have the final say on financial issues. Young people's choices are also largely determined by the opinions and values of their peers, and financial matters depend on the financial situation of the family. Joint decisions are most common in deciding on the settlement and housing. The family's financial status determines how far the student can go for higher education, but individual preferences are also allowed, and the family considers the best options together. The importance of a degree and potential job prospects are heavily influenced by parents' views [11].

Several studies have examined the factors that most influence students' decisions to study at one institution or another. These studies list various factors influencing the decision, such as the quality of the institutions and their social acceptance [12] [13] [14], the cost of education and tuition [15] [16], labour market opportunities [17], and the impacts of the family, friends and academic life [18]. The process of selecting an institution is a complex decision on the part of the student [19]. This choice can also affect a student’s future career, place of residence, and personal satisfaction [20]. Smith and Cavusgil, also confirm this statement, suggesting that this type of decision is in many cases unique in an individual’s life, which incurs different financial and psychological costs [21]. According to Chen [22], students select their prospective university based on a three-step process: (1) the student's personality (socio-economic background, personality traits, personal preferences, learning abilities, relational capital, and creative capital); (2) other decision makers (family members, friends, relatives, teachers, and employers); (3) external push-pull factors (personal driving forces, institutional characteristics). In their research, Galotti and Mark [23] found that students with better abilities considered significantly more aspects and applied to more institutions than their peers with more modest abilities.

Institutional selection criteria can be various. On the one hand, they can be interpreted in terms of student considerations, and on the other hand, in terms of institutional factors. Social and economic status, the educational attainment of the parents, budget, financial assistance, career, the opinion of parents, peers, and teachers are all variables that affect students. On the institutional level, the geographical situation, size, type of training, infrastructure, reputation, ranking and the quality of education are all factors to consider [24]. Niu and Tienda [25], confirm in their study that students consider different college/university

characteristics when choosing an institution. Drewes and Michael [26], draw attention to the importance of the geographical distance between the institution of higher education and the place of residence. Their survey of Canadian high school students showed that respondents preferred institutions close to home when choosing a university. Heine [27], in a study of student preferences, found that the reputation of the university, its facilities, but also the range of leisure activities and the atmosphere on campus can influence students' choices.

According to Kuráth and Sipos, in selecting a university, financial return factors, such as comparing education and living expenses to potential "returns," are becoming increasingly relevant. The attractiveness of the city of the educational institution is a competitive advantage: the site of the educational institution, the distance of the place of training from the student's place of residence and the labour market opportunities of the region. Student choice is influenced by which centre offers greater opportunities for placement and long-term employment. In other words, the economic effect appears strongly, which also influences the push or pull potential of a settlement or institution [28].

According to Frankowska et al. [29], the potential career prospects associated with the selected field of study play an important role in the student's decision. This is due to the fact that finishing a study programme helps them earn money after graduation, improve their social status, and achieve their goals. In their study, Heine and Willich [30] call attention to the risks of selecting a university and a study programme. Prospective higher education students are often underinformed about study programmes and prospects, as well as uncertain about their personal abilities, ambitions, and desires, according to their findings. As a result, these considerations can distort decisions on continuing education. In order to prevent this risk, it is becoming increasingly important to educate and prepare students, i.e., carry out career counselling activities.

In the mid-1990s, Powell [31], explains the four pillars of career planning. The first pillar is a person's self-esteem. The point is that the person is looking for the answer to the question, "Who am I?" (E.g., what am I capable of, what are my skills, abilities, what am I interested in, etc.)? The second pillar is to look for career opportunities. At this point the person reviews the opportunities offered by the environment (e.g., labour market, occupations, prospects, salaries, etc.). The third pillar of career planning is defining goals, that is, clarifying how you will select (e.g., clarifying goals, objectives in life, analysing career advantages, disadvantages, etc.). Finally, the fourth pillar is the action itself (learning, placement), which interacts closely with the individual's life goals, willingness to compromise, existing career goals, and their analysis. The driving force behind all this is nothing but decision. In Powell's theorem, awareness can be traced down in all four pillars.

According to KSH [32], 287.5 thousand students continued their studies in 62 higher education institutions in Hungary in the academic year 2020/21. Of these,

one in five were enrolled in economics and 13% in engineering. The number of foreign students has doubled in the last ten years.

Success on the job market and employability are inextricably linked to career planning. Watts [33], divided employability into three groups based on the role of time: immediate employment, immediate employability, and sustainable employability. He uses the term "immediate employment" to describe the percentage of people who are hired six months after graduation. Immediate employability examines whether a graduate student has the qualifications and expertise necessary for a graduate job without any further learning right after graduation. Sustainable employability explains that it is critical not only to get a first graduate job, but also to remain continuously employed, which necessitates continuous renewal. Professional opinions are divided on the attitude to work of the new generation entering the labour market: skillful, practical, have divided attention, highly materialistic, impatient. At the same time, they are overconfident and self-assured. They also have high expectations of their employer, striving for immediate success and results with the least possible investment of energy [34]. They clearly want to assert the "I am basically entitled to it" feeling in the labour market. A study by Juhász and Terjék [35] of Generation Y found that women in their sample appear to be more career-oriented than their male colleagues. Their impatience manifests itself in the fact that if their unrealistic ideas are not realised in a short period of time, they will leave the workplace without thinking and try to succeed elsewhere. Spending 3-4 years in a job is surprising. One in four young people are preparing to leave their jobs within a year [36]. Kiss [37], conducted an analysis of the data of the Graduate Career Tracking System, in which, examining the factors that influence the time of employment of recent graduates, it was concluded that good academic results and good language skills significantly improve the job search process.

3 Method of Research

Between 2019 and 2020, the authors surveyed students in Hungarian higher education on what aspects affected their decision to continue their studies, select their university and also how they see time spent in higher education contributes to their future labour market jobs, and whether or not their choice of profession was a well-established and appropriate decision. The following institutions participated in the research: University of Debrecen, University of Miskolc, Budapest University of Technology and Economics, Szent István University, University of Szeged, University of Pécs, University of Győr.

The main objective of the examinations was to involve students in the research in as wide a geographical and professional distribution as possible. The first stage of the survey focused on students and then involved employers to assess the extent to which they believe graduates meet labour market needs.

This paper presents some results of the questionnaire survey with students only.

Participants were asked to fill in an anonymous questionnaire, which was available online or on paper. Inquiries were made in person, on the internet, by email, and via social media.

When respondents completed the pilot questionnaire, all of the questions were correctly understood. Just one open question was included in the questionnaire, which was mostly made up of closed questions. The questions were typically based on nominal and metric variables, the latter on a 5-point Likert scale. The number of questionnaires that could be evaluated was 328. The questions of the questionnaire were divided into 4 areas as presented in Table 1.

Table 1
The structure of the questionnaire

1st group of questions	**2nd group of questions**	**3rd group of questions**	**4th group of questions**
General data	Career choice attitude to learning	Training	Employment and labour market knowledgeability
Name of the higher education institution, its geographical location where they study Gender, age and place of residence of the respondents Parents' educational attainment	Factors determining further education Factors determining the choice of a study programme Factors determining the choice of institution Expectations regarding the degree	Satisfaction with the quality of training Sources of additional information needed for further education	Employment objectives, expectations and realities Factors for a successful employment Knowledge of domestic labour market trends

The correlations that result from the responses to Question Group One, Two, and Four will be the subject of this article. The aim of the research is to look at students' motivations in relation to their desire to pursue their studies and the factors that influence their decision-making.

The research question is whether the students' decision to continue their studies at tertiary level is influenced by their desire to succeed in the labour market. This is tested along two hypotheses: (1) In further education, rational decision-making is crucial, and is mostly justified by the labour market situation and professional interests, but the family and friends also play a role. (2) As graduates, students expect primarily labour market and professional success.

The following statistical methods were used to test the hypotheses: single- and multivariate analyses, such as frequency, mean, standard deviation, factor analysis, ANOVA and SEM modelling. The programmes used for the evaluation

were SPSS Statistics 27 and SPSS AMOS 27. Since the questionnaire basically contained scale and categorical variables, the authors applied the methodological analyses mentioned above.

4 Results and Discussion

The examination includes an analysis of the responses of 328 individuals. The sample specification is presented in Table 2.

Table 2
Specification of the sample

	Total
Gender	34.1%Male 65.9% Female
Age	Average age: 23.16 years The youngest is 18, the oldest is 50.
Residence	41.4% Budapest 20.1% County seat 25% City, town 13.4% Village
Study programme	37.8% Arts 36.2% Business 18.4% Technical 4.6% Healthcare 3.1% Other

During the sample specification, the researchers inquired about the parents' educational attachment. Just 30.8 percent of the fathers had a secondary school certificate, while 45.8% had a college or university diploma. The proportions were higher among mothers: 49.7% had completed tertiary education and 38.8% a secondary school.

The above data was required to see to what extent parents' educational achievement affected their children's decision to continue their education. In addition, participants in the research were asked to identify additional motivating factors that influenced their decision to go on to higher education. Respondents had to decide on statements, on a five-point Likert scale, whether the given factor played a decisive role in this question completely (Value 5) or not at all (Value 1). Table 3 shows the mean and standard deviation of the answers to this question.

Table 3
Factors motivating participation in higher education (mean, standard deviation)

Statements	N Valid	Mean	Std. Deviation
My friends encouraged me.	328	2.53	1.347
My parents encouraged me.	328	3.92	2.115
My teachers encouraged me.	328	2.88	1.376
I was attracted by student life.	328	3.02	1.428
I could not find a job after school.	326	1.62	1.107
I did not feel like working.	328	1.93	1.346
I like studying.	328	3.17	1.244
I would like a degree.	327	4.33	1.046
I believe that I can find a job more easily with a degree.	328	4.14	1.083
I believe that I will have a higher salary with a degree.	328	4.16	0.996

The table shows that earning a degree is the primary aim for the respondents. However, there are financial and labour market / career building factors for a student to go on to higher education. Parents are a significant factor of the impacts of the immediate environment. The standard deviation value was very high, i.e., in this respect it cannot be said that the participants in the sample had a unanimous opinion. Higher salary, degree, and career reasons had the smallest standard deviation values indicating that respondents' opinions on these topics are nearly identical. It is interesting to note that the desire to learn (I like studying) and to learn new things was hardly motivating for the respondents. The question arises as to how much they can strive for good learning outcomes in the absence of this.

The ANOVA method allowed us to explore how parental education influences the above reasons. While analysing the educational attainment of fathers, it was found that a significant difference was found for the statement "I can find a job more easily with a degree" (F: 3.027 sig.: .030 $p<0.05$) and for the statement "I will have a higher salary with a degree" (F:5.48 sig.: .001 $p<0.05$). These findings were more strongly supported by students whose father had a degree than those whose father did not have tertiary education. The same result was found for mothers. When examining the statements regarding easier employment (F: 3.792 sig.: .011 $p<0.05$) and higher salary (F: 5.655 sig.: .001 $p<0.05$), a significant difference was confirmed according to the mother's educational attainment. In this case, too, the support for the statement was strongest for students whose mothers had a higher education degree. It is more likely that these students have already had an example to prove that those who have a degree can find employment more easily, for a higher salary, and in a better position than those who do not. The first hypothesis of the research was only partially confirmed.

For further investigation, the authors grouped the variables influencing participation in tertiary education into factors. Two variables: 'My parents encouraged me' and 'I was attracted by student life' were not suitable for factor construction. KMO and Barlett's test results were as follows. Chi-squared: 558.176 Sig: .001. Varimax rotation was used to construct the factors and the explained proportion was 64.65%. The factors, their names, components and Cronbach Alpha values are summarized in Table 4.

Table 4
Motivating factors for choosing higher education (p=0.05)

	Component		
	Career and financial considerations	No option	Emotional reasons
(KM1) I believe that I can find a job more easily with a degree.	0.901		
(KM2) I believe that I will have a higher salary with a degree.	0.839		
(KM3) I would like a degree.	0.768		
Cronbach Alpha	0.800		
(NO1) I did not feel like working.		0.802	
(NO2) I could not find a job after school.		0.782	
Cronbach Alpha		0.562	
(EM1) My parents encouraged me.			0.773
(EM2) My friends encouraged me.			0.673
(EM3) I like studying.			0.583
Cronbach Alpha			0.463

Three variables were identified by the authors. The first explains the career and financial factors (its components are labelled CF), the second explains the lack of option (labelled NO), and the third explains the emotional reasons for the choice (labelled E) (marked component: EM). The labels are used in the SEM model formation.

Based on the data, it seemed to prove that basically graduation, future financial appreciation, and labour market and career opportunities motivated students to continue their studies. The question arises as to what aspirations a person may have by obtaining a degree; and if these aspirations are in line with the reasons for going on to higher education. Respondents were also asked to rate their aspirations for a degree on a Likert scale of five, depending on the specificity of the statement. A score of five indicated completely characteristic, while a score of one indicated not characteristic at all. The results (mean and standard deviation) of the expectation of a degree are presented in Table 5.

Table 5
Expectations of a degree (mean, standard deviation)

Statements	N	Mean	Std. Deviation
	Valid		
Higher salary	327	4.09	1.068
Good job	327	4.28	0.884
Professional recognition	327	4.16	0.931
Managerial position	327	3.37	1.211
Interesting and responsible job, position	327	3.98	0.984
Good expertise in several areas	327	3.91	0.997
Protection against unemployment	327	3.80	1.224
Social recognition	327	3.55	1.150
Self-fulfilment	326	3.97	1.099
Family recognition	327	3.62	1.254
Job opportunity abroad	326	3.05	1.465

Students expect a good job opportunity, professional and financial appreciation from the degree. The beneficial effects of the degree on the labour market are also felt by the respondents, as expertise in several areas and protection against unemployment have also been highly appreciated. However, they do not think that obtaining a degree today would lead to greater social or family recognition. Nor did they think to be prevented from working abroad if they did not have a higher education degree. This is also supported by the data that when the respondents had to determine labour shortage, the occupations that required a secondary qualification were basically indicated: seamstress, carpenter, car mechanic, bricklayer, and only then were the occupations requiring a degree listed such as doctor, teacher, engineer, etc.

Collecting the variables into factors, it turned out that one variable was unsuitable for factor formation: self-fulfilment. The KMO and Barlett test results are as follows: KMO: .818 seems to be very good. Approx. Chi-square: 956.181, sig: .001. The factors were formed by varimax rotation, while the explained proportion was 61.485%. The factors, their names, components, and Cronbach Alpha values are summarized in Table 6.

Table 6
Factors of expectations of a degree (p=0.05)

Expectations	Component		
	Career	Financial security	Recognition by environment
Interesting and responsible job, position	0.800		
Professional recognition	0.746		
Good expertise in several areas	0.666		
Managerial position	0.517		
Cronbach Alpha	0.728		
Higher salary		0.837	
Good job		0.767	
Protection against unemployment		0.642	
Cronbach Alpha		0.710	
Family recognition			0.790
Job opportunity abroad			0.698
Social recognition			0.596
Cronbach Alpha			0.651

Based on the components the factors were named as Career, Financial security, Recognition by environment. Further research was directed at whether there were any ethnicity, age, or parental educational attachment differences in the given sample.

The ANOVA examination revealed that significant difference could only be explained when the career aspect was examined by gender (F: 6.868 sig.:.009 p0.05), and that professional career was more important for men as an expectation in this case than for women.

Analysing the parents' education by ANOVA method, significant difference could only be verified in the case of the fathers: examining the environmental assessment (F: 4.018 sig.: .008 p<0.05). If a student had a father with a lower level of education, the environmental respect that a degree can provide was more important to them than to students whose father had higher education.

The examinations by age confirmed that only in the case of career was there no significant correlation between age and the factor (Pearson Correlation:.110 sig.:.070 p>.0.05). Furthermore, the significant correlation was also proved by the fact that the younger a respondent, the more appreciated financial security by degree (Pearson Correlation: -.164 sig.:.006 p<0.05) and recognition by environment (Pearson Correlation: -.243 sig.:.001 p<0.05). The second hypothesis is to be accepted.

The questionnaire also asked respondents what factors they thought were necessary for a successful job. Listing a number of components, these were rated

by the students on a Likert scale of five, similar to the previous question. A score of five indicated completely and one indicated not at all, as shown in Table 7.

Table 7
Factors that support a successful job (mean, standard deviation)

Factors	N	Mean	Std. Deviation
	Valid		
A degree with good results	304	3.55	2.069
Good internship	304	4.05	0.934
Good connections	304	4.88	2.370
Good communication skills	304	4.48	0.744
Self-assurance	304	4.46	0.721
Adequate references	304	4.09	0.884
Language knowledge	304	4.30	0.941
Problem solving ability	304	4.20	0.855
Professional language knowledge	304	3.90	1.047

The students' opinion is thought-provoking: the qualification of the degree is now moved to the bottom of the priority list when it comes to job seeking. In employment respondents believe it is most important to know who has what relationship with whom and how they can assert themselves at an interview. These considerations take precedence over professional expertise. However, it is also evident from the viewpoints of the students that there was no consensus on the importance of a good degree, as the average was highest here.

The variables that induced a successful job were also grouped into factors by the researchers. The KMO and Barlett test: KMO: .745 is good. Approx. Chi-square: 444,357 sig: .001. The factors were formed by Varimax rotation and the proportion explained was 67.804%. The component values and the names of the factors are presented in Table 8.

Table 8
Factors inducing a good job

Reasons	Component			
	Soft skills	Professional knowledge	Reasons of the student	Good connections
(SK1) Good communication skills	0.837			
(SK2) Self assurance	0.802			
(SK3) Problem solving ability	0.619			
Cronbach alpha	0.692			
(P1) Professional language knowledge		0.821		
(P2) Language knowledge		0.816		
(P3) Adequate references		0.541		

Cronbach alpha		0.665		
(ED1) A degree with good results			0.808	
(ED2) Good internship			0.753	
Cronbach alpha			0.313	
CON Good connections				0.988

The authors formed four factors: the first symbolizes soft skills (its components are marked SS), the second, professional knowledge (marked: P), the third the reason related to the student (marked: SR), and finally the importance of good connections (Label: CON). The labels are used in the SEM model formation.

SPSS AMOS 27 offered the opportunity to perform additional analyses. It was investigated whether any correlation could be justified between the reasons for good jobs and opting for further education.

Among the factors that induce the choice of tertiary education, the following were included in the analysis: career and financial reasons (components labelled CF), no other option (labelled NO), and emotional reasons (component labelled E). In the case of successful labour market placement, two factors were included in the model analysis, which are basically linked to higher education as they can be acquired and developed there: soft skills (components labelled SS) and professional knowledge (components labelled P). The other two factors were not considered in this relation because they depended less on the educational opportunity and more on the student and the environment.

The authors used the SEM model they developed to evaluate thought associations. The association between one or more exogenous (independent) variables and one or more endogenous (dependent) variables may be examined using SEM (Structural Equation Modelling). Endogenous variables influence exogenous variables both directly and indirectly. The objective of the CFA (Confirmatory Factor Analysis) examination is to figure out how the examined and latent variables are related. The CFA should be used if the researcher already has knowledge of the latent variable structure.

The model of the factors and variables under study was created using SPSS AMOS 27 version. Five external models were constructed (measurement model) based on latent variables and indicators. The external models are composed of Career and financial reasons, No other option, Emotional reasons, Soft skills, Professional knowledge.

The Measurement model allows the exploration of the relationships between the indicators and the latent variables. The internal model (Structure model) allows the exploration of the relationships between the variables Career and financial reasons, No other options, Emotional reasons, i.e., the interrelationships between these variables. The back-and-forth arrows in Figure 1 represent the covariance or correlation between the variables, while the system of arrows depicts the influence of one variable on the other. The indicators were given letters that corresponded to

the letters in Tables 4 and 8. In the diagram, the error variables are represented by circles. These factors were not in the focus of the study but have an effect on the variables. The model is shown in Figure 1.

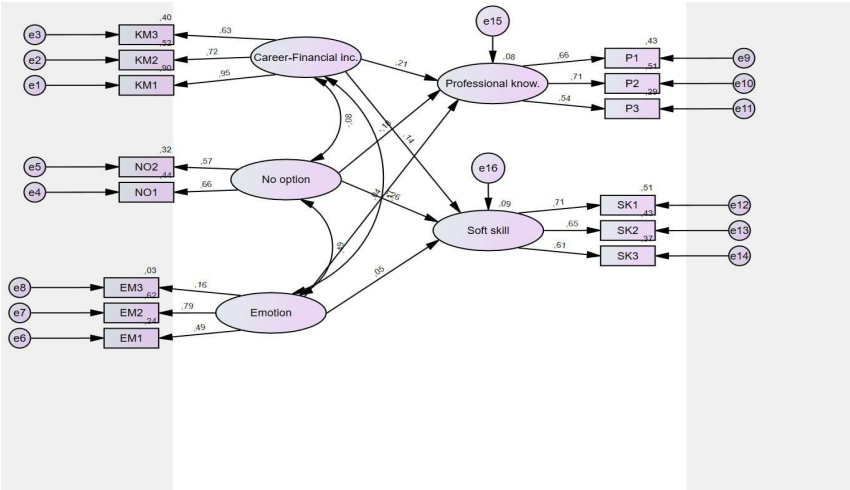


Figure 1

Relationships between the reasons for higher education and a successful job

KM= CF SK= SS

The following metrics should be used to assess the suitability of the model. The Chi-square was significant ((238) 220,782 df: 68 p: .00) in the first research metrics, confirming "absolute model fitness." However, this is insufficient to rule out the model's fit since the Chi-square significance on a sample size of 200 indicates a value of 0 more strongly. Other indicators need to be examined such as RMSEA (Root Mean Square Error Approximation) value: .085, which should typically be below 0.08, but this value is close to the limit. The third such indicator is the GFI (Goodness of Fit Index), which is acceptable for values above 0.9. For the present model, this value is .912, so it is adequate. In the context of Incremental Model Fit, four indices have been examined, AGFI, CFI, NFI, TLI, whose values above 0.9 are considered good. In the model, AGFI: .864, CFI: .836, NFI: .786, TLI: 780, these indices also have values close to the limit. In the case of Parsimonious Fit the value of Chi-squared/df is 3.376, which is good as it does not exceed the value limit of 5, so based on this, the model can be considered fit, as well.

The following findings are interesting when looking at the relationship between each element. Career and financial incentives have a favourable impact on professional knowledge, according to regression principles (.211). Students who entered higher education without any expectations continued to believe that developing soft skills was less beneficial to later success (-.262). Interestingly,

those who chose higher education on an emotional basis have not demonstrated in this regard that both soft skills and professional knowledge determine the way to success.

According to the model, a medium positive relationship value (.489) could be justified between "no other option" and emotional triggers, whereas there could be a relatively weak relationship between financial and career motives and emotional motivations proved by the model.

The fundamental hypothesis of the research, i.e., students' intention to go on to higher education is highly influenced by their ability to be successful in the labour market - has been justified in the light of the results.

According to the findings of this study, the responding students conclude that a degree has a value in the labour market because it opens up career prospects for young people and therefore, they expect outstanding professional and financial recognition from their studies. In recent decades, more and more research has emerged showing a strong link between education and average earnings [38] [39]. According to Pascarella and Terenzini [40], those who continue their education after secondary school are more likely to earn excellent salaries and a better career status than those who only have a secondary school certificate. In a scientific paper published in 1989, Murphy and Welch [41] examined the differences in incomes and discovered that one of the causes for the disparities was separate levels of education. This is also supported by the study published by Naguib et al. [42], according to which those with higher education earn on average better than those without tertiary education.

A number of research in international literature have emphasized the importance of family and friends in shaping learning decisions. Sellami [43] in his study highlighted the powerful influencing role of parents, stating that parents have a significant effect on their children's educational and career decisions. These findings are also echoed by Georgiou and Tourva [44], as well as Kobal Grum and Seničar [45]. However, it is important to note that the results obtained in the present study do not support the findings of these previous studies, as the students interviewed did not confirm a strong relationship between the desire to learn and the influence of parents and friends.

One of the most significant advantages of education is that it allows people to develop competences that can be valuable later in their careers [46] [47] [48]. Looking at different skills, the results showed that so-called soft skills are important for future professions [49]. But many studies show that relational capital is even more so. According to the respondents, having a network of connections can be extremely beneficial when looking for jobs after graduation. This result has been confirmed by several literature sources. According to a 2005 paper by Forret and Dougherty [50] there is a close relationship between relational capital and successful career building. In a 2000 study, Langford [51], concluded that an individual’s ability to make relationships is closely related to a successful

career. Furthermore, confident, determined action is also a key factor nowadays, so it can also play an important role in later career building, especially in areas where good aptitude is crucial [52]. Students interviewed in their study also considered self-confidence in career building as an important factor.

Conclusion

The objective of the present paper was to explore the motivations of university students in terms of their intention to continue their studies, as well as the factors that influence their decision-making. The findings support the attainment of favourable career opportunities, as well as personal and financial success.

The results of the study show that it is primarily obtaining a degree why graduates are motivated to participate in higher education, and that the possibility of a higher salary and finding employment ore easily with a higher education degree basically influence their decisions on whether to continue their studies in higher education after secondary school. The desire to learn (I like studying), to learn something new, hardly motivated the respondents. Parental influence is typically “only” indirect, and friends have little effect on further education, as well.

Students expect good job opportunities, as well as professional and financial recognition as a result of their degree. The latter two are more important the lower the father's education and the younger the individual interviewed. The beneficial effects of the degree on the labour market are highlighted by respondents; they expect applicable expertise and protection against unemployment.

As a result, it can be concluded that students assume that relationships and self-assurance are the key factors to finding a good job. Relatively stronger decision-relationships could be found among those whose main reason for continuing their education was a future career and livelihood, who were most likely to see professional knowledge as a guarantee of later success in finding a job.

Based on the above, the following solutions emerge:

- The transformation of pedagogical processes in institutions, with the aim of fostering the development of the child who selects the institution for employability.
- A stable, well-functioning economy, well-established ways connecting education and employment, a great variety of work experience opportunities, information resources, and a consultancy system are all important.

At the same time, given the dynamism of the labour market, students and all workers need to be prepared for a possible change of profession and be able to adapt to changes in the labour market.

A well-educated, trained, and confident society in the labour market could be created by improving and extending labour market knowledge, transforming people's mindsets and increasing their mobility, retraining, and the concept of "Lifelong Learning." In the future, the direction of the research could be to ask the

employers' side of the researchers which knowledge content should be included in the educational content and where the gaps are between the educational and labour market expectations.

References

[1] G. Halász: The education system. Technical Book Publishing House, Budapest, 2001

[2] P. Tóth: Learning strategies and styles in vocational education. Acta Polytechnica Hungarica, Vol. 9, No. 3, 2012, pp. 195-216

[3] M. Ritoókné: Personality development and career choice. Textbook Publisher, Budapest, 1986

[4] M. Ritoókné: Career development - career development counselling. ELTE, Eötvös Publisher, 2008

[5] P. Galasi: Have higher education diplomas really been devalued? Budapest Labour Economics Notebooks, Vol. 3, 2004, pp. 1-14

[6] E. Hetesi, Zs. Kürtösi and Zs. Hetesi: Who judges the performance of higher education services and how? A student satisfaction measurement models, empirical research results among active and graduate students. Management Science, Vol. 39, No. 6, 2008, pp. 2-17

[7] Á. Hofmesiter-Tóth: Consumer behaviour. Budapest, Aula Publisher, 2006

[8] M. R. Ryan and E. L. Deci: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist, Vol. 55, No. 1, 2000, pp. 60-68

[9] R. Becker and A. E. Hecken: Why are working-class children distracted from studying at universities? An empirical test of the "distraction thesis "by Müller and Pollak (2007) and its extension by Hillmert and Jacob (2003) Köln Journal of Sociology and Social Psychology, 2008, Vol. 60, No. 1, 2008, pp. 7-33

[10] M. J. Safranková and M. Sikyr: The study of university students' motivation. International Journal of Teaching and Education, Vol. 4, No. 4, 2016, pp. 48-59

[11] M. Törőcsik: Consumer behaviour of young people. In: Mária Törőcsik - Kuráth Gabriella (ed.). University marketing. Marketing in higher education. Pécs, University of Pécs, 2010, pp. 123-141

[12] R. Chapman: Non Simultaneous relative performance analysis: Meta-analysis from 80 collegechoice surveys with 55,276 respondents. Journal of Marketing for Higher Education, Vol. 4, No. 1/2, 1993, pp. 405-422

[13] R. Coccari and R. Javalgi: Analysis of students' needs in selecting a college or university in a changing environment. Journal of Marketing for Higher Education, Vol. 6, No. 2, 1995, pp. 27-39

[14] O. Fatoki: The determinants of the career choice of international students in South Africa. Mediterranean Journal of Social Sciences, Vol. 5, No. 23, 2014, p. 668

[15] M. Webb: Variables influencing graduate business students' college selections. C & U Feature, 1993, pp. 38-46

[16] L. Lin: What are student education and educational related needs?. Marketing and Research Today, Vol. 25, No. 3, 1997, pp. 199-212

[17] G. Soutar and J. P. Turner: Students' preferences for university: a conjoint analysis. The International Journal of Educational Management, Vol. 16, No. 1, 2002, pp. 40-45

[18] T. Shanka, V. Quintal and R. Taylor: Factors Influencing International Students' Choice of an Education Destination –A Correspondence Analysis. Journal of Marketing for Higher Education, Vol.15, No. 2, 2005, pp. 31-46

[19] M. Yost and S. Tucker: Tangible evidence in marketing a service: The value of a campus visit in choosing a college. Journal of Marketing for Higher Education, Vol. 6, No. 1, 1995, pp. 47-67

[20] P. Kotler and K. Fox: Strategic Marketing for Educational Institutions. (2nd ed.) New Jersey, Prentice Hall, 1995

[21] L. Smith and T. Cavusgil: Marketing planning for colleges and universities. Long Range Planning, Vol. 17, No. 6, 1994, pp. 104-117

[22] L. H. Chen: East-Asian Students' Choice of Canadian Graduate Schools. International Journal of Educational Advancement. Vol. 7, No. 4. 2007, pp. 271-306

[23] K. M. Galotti and M. C. Mark: How do high school students structure an important life decision? A short-term longitudinal study of the college decision-making process. Research in Higher Education. Vol. 35, No. 5, 1994, pp. 589-607

[24] C. D. Hachmeister: Indicator in focus: criteria of university choice and ranking use. 2007, accessed on March 2021, www.che.de/download/iib_hochschulwahl_und_rankingnutzung-pdf/

[25] S. X. Niu and M. Tienda: Choosing colleges: Identifying and modeling choice sets. Social Science Research, Vol. 37, No. 2, 2008, pp. 416-433

[26] T. Drewes and C. Michael: How do students choose a university?: an analysis of applications to universities in Ontario, Canada. Research in Higher Education, Vol. 47, No. 7, 2006, pp. 781-800

[27] C. Heine: First-year students in the old and new states: Reasons for university choice and evaluations of the university regions of West and East Germany. 2008, accessed on March 2021 http://docplayer.org/79735220-Studienanfaenger-in- den-alten-und-neuen-laendern-gruende-der-

hochschulwahl-und-bewertungen-der-hochschulregionen-west-und-ostdeutschland.html

[28] G. Kuráth and N. Sipos: Changes in domestic higher education application decisions. A regional effect. Space and Society, Vol. 33, No. 1, 2019, pp. 173-184

[29] A. Frankowska, A. Głowacka-Toba, R. Rasińska and E. Prussak: Students entering the labour market, their hopes, expectations and opportunities in the context of sustainable economic development. Journal of International Studies, Vol. 7, No. 3, 2015, pp. 209-222

[30] C. Heine and J. Willich: Information behavior and decision making in the choice of study and education. 2006, accessed on March 2021 https://www.uni-heidelberg.de/imperia/md/content/jrc/1503/fh_200603.pdf

[31] C. R. Powell: Career planning today: hire me. Kendall Hunt Publishing Company, 1995

[32] KSH: Higher Education 2020/https://www.ksh.hu/docs/hun/xftp/idoszaki/oktat/felsooktatas2021/index.html

[33] A. G. Watts: Career development learning and employability, Learning and Employability Series Two, The Higher Education Academy. 2006, accessed on March 2021, hca.ltsn.ac.uk/assets/documents/resources/resource database/id592_career_development_leaming_and_employability.pdf

[34] I. Csehné Papp: expectations and reality in theworld of work. TAYLOR Journal of Management and Organizational Sciences, Vol. 2, 2016, pp. 5-11

[35] Cs. Juhász and L. Terjék: Examining expectations for Generation Y. Publications of the Virtual Institute for Central European Research, Vol. 6, No. 25, 2014, pp. 221-226

[36] J. Martin and C. Schmidt: How to Keep Your Top Talent. Harward Business Review, Vol. 88, No. 5, 2010, pp. 54-61

[37] Zs. Kiss: Job search time: the indicator of employability. Quest Journals, Journal of Research in Business and Management, Vol. 2, No. 4, 2014, pp. 1-9

[38] C. Yue and Y. Liu: Impact of education on the income of different social groups. Frontiers of Education in China, Vol. 2, No. 2, 2007, pp. 191-200

[39] T. K. Bauer: Educational mismatch and wages: A panel analysis. Economics of Education Review, Vol. 21, No. 3, 2000, pp. 221-229

[40] E. Pascarella and P. Terenzini: How College Affects Students. JosseyBass, San Francisco, 1991

[41] K. Murphy and F. Welch: Wage Premiums for College Graduates. Educational Researcher, Vol. 18, No. 4, 1989, pp. 17-26

[42] C. Naguib, M. Baruffini and R. Maggi: Do wages and job satisfaction really depend on educational mismatch? Evidence from an international sample of master graduates. Education + Training, Vol. 61, No. 2, 2019, pp. 201-221

[43] A. L. Sellami: Parental Influence on Student Educational Expectations: Results from the 2012 Qatar Education Study. International Journal of Higher Education, Vol. 8, No. 4, 2019, p. 189

[44] S. N. Georgiou and A. Tourva: Parental attributions and parental involvement. Social Psychology of Education, Vol. 10, No. 4, 2007, pp. 473-482

[45] D. K. Grum and M. Seničar: Academic Motivation, Competitive Attitude and Self-esteem among University Students in Slovenia. Learning and Teaching, Vol. 7, No. 1, 2014, pp. 5-23

[46] K. Gál and L. Székedi: Digital Literacy at the Partium Christian University. In G. Hajzer, I. Mátyus, N. Merkovity, J. Molnár, P. B. Stumpf and O. Szabó Palócz (Eds.): Communicative Space—Political Space, Iurisperitus Publisher, 2000, pp. 59-70

[47] M. Garai-Fodor: The Impact of the Coronavirus on Competence, from a Generation-Specific Perspective Acta Polytechnica Hungarica. Vol. 19, No. 8, 2022, pp. 111-125

[48] E. Gogh, R. Racsko and A. Kovari: Experience of Self-Efficacy Leraning among Vocational Secondary School Students. Acta Politechnica Hungarica, Vol. 18. 2021, pp. 101-120

[49] T. Juhasz and G. Horváth-Csikós: The emergence of soft skill in agriculture. Probles and Perspectives in Management, 2021,19 pp. 453-466

[50] M. L. Forret and T. W. Dougherty: Networking behaviors and career outcomes: Differences for men and women? Journal of Organizational Behavior, Vol. 25, 2004, pp. 419-437

[51] P. H. Langford: Importance of relationship management for the career success of Australian managers. Australian Journal of Psychology, Vol. 52, 2000, pp. 163-168

[52] A. De Jong, K. De Ruyter and M. Wetzels: Linking Employee Confidence to Performance: A Study of Self-Managing Service Teams. Journal of the Academy of Marketing Science, Vol. 34, No. 4, 2006, pp. 576-587

Development of a Calculation Method for Determining the Permissible Dimensions of the Defect 11

Baglan Togizbayeva, Aliya Zabiyeva, Kyrmyzy Balabekova, Anuar Kenesbek

Departament of Transport, transport technic and technology, Faculty of Transport and energy, Eurasian National University name of L.N. Gumilyov
Kazhymukan street 13, 010000 Astana, Republic of Kazakhstan
e-mail: togizbayeva_bb@enu.kz, zabiyeva_ab@enu.kz, balabekova_kg@enu.kz, kenesbek_ab_1@enu.kz

Abstract: This article deals with the issue of derailment of the wheels in railway transport, due to the probabilities of the wheel rolling on the rail. We proposes a method for calculating the maximum allowable sizes of defects 11, at different angles of the wheel on the rail with different shapes and sizes of wear of the wheel and rail edges. This technique allows us to calculate the permissible dimensions of Defects 11, according to the conditions of rolling the wheel crest onto the rail with the most unfavorable combination of interacting parts.

Keywords: Finite element method; rail; wear; traction; train; crack; internal longitudinal crack (ILK)

1 Introduction

According to statistics, the reason for the largest part of the rolling surface defects of rails in Kazakhstan is metal shelling because the contact fatigue strength of the metal is not enough (defects 11.1-2) [1]. They are formed in the working part of the rail head when the metal under the influence of alternating train loads works in a zone of limited endurance. After skipping a certain tonnage, in places where there are non-metallic inclusions in the head, lying at a depth of 3-15 mm, there are dents and so-called «dark spots», indicating the presence of an internal longitudinal crack (ILC) in the rail head. According to the instructions [2], rails with a depth of more than 3 mm and «dark spots» with a load capacity of more than 10 million tons of gross km per year should be replaced as planned. However, with the sharp increase in the cost of materials for the upper structure of the track characteristic of modern conditions, the absence of a required contingent of track fitters and the limitation of issued «windows», there are sections of track on the road network on which up to

20% of rails have defects 11.1-2 and they are replaced with a continuous change of rails or during major repairs of the track. In agreement with planned procedure, according to the results of inspections of the track, only 5-6% of rails with defects 11.1-2 are removed. Therefore, a sufficiently large number of rails with contact fatigue shelling are working on the way.

The defects 11 formed on the working surface of the rails can be conditionally divided into several groups with differences in the signs of the depth of occurrence and the form of manifestation, which in turn affects the further operation of rails with these defects. The first group includes surface defects in the form of a grid of small cracks and small stains up to 3 mm deep, the "peeling" or "fish scales", which can spread throughout the entire rail. These stains do not affect the strength of the rail, but they interfere with the control of rails by standard means of track flaw detection, which significantly complicates the identification of other defects of the rail head. This reduces traffic safety and in some cases forces decisions on early replacement of rails. The second group includes dark and light spots on the side working edge of the rail and often accompanying metal splashes. These spots are formed above the internal longitudinal crack, which lies at a depth of, as a rule, 3-8 mm, and leads to the staining of the metal surface during further operation. The danger of these defects is explained by the fact that the zone of development of transverse fatigue cracks (defect 21.1-2) coincides with the zone of development of the internal longitudinal crack, i.e., the transverse crack develops in the line inclusion of alumina, which is the focus of the internal longitudinal crack. Since the longitudinal crack passes over the transverse one, it is difficult and most often impossible to determine the internal longitudinal crack by means of track flaw detection. However, after the split of the melted metal, the conditions of interaction of the "wheel-rail" system change and the absence of contact of the wheel with the lateral working edge of the rail at the point of the breakout excludes any further development of a transverse crack. The third group includes damage to the rails in the form of gouges on the side of the working face. Depending on the operating conditions and the depth of occurrence of non-metallic inclusions, they can be shallow (up to 4 mm) and small in length (up to 10 mm), arranged either singly or several pieces in a row, or of more significant dimensions (depth up to 14 mm, length up to 300 mm, and also possibly the merger of several dents), which, with the tightening of operating conditions and the general deterioration of the condition of the path, are becoming more common [4].

Of particular importance are the studies of the strength of rails with the shape of defects 11. This is due to the fact that when large sizes are reached in the cross section of the notch, the strength properties of the rail weaken. It is considered that when the cross-sectional area is reduced by more than 30%, it is necessary to replace the rail. Thus, it can be assumed that with certain sizes of gouges, there is a danger of breaking the rails under the train, and therefore such rails must be transferred to the category of acute defects. In addition, significant geometric dimensions in length and depth do not only worsen the local dynamic impact, but also significantly increase the likelihood of rolling the wheel crest onto the rail.

In this regard, the question arises of determining the criteria by which the replacement of rails with defects 11 should be carried out and the establishment of their boundary dimensions at which the operation of rails is safe.

Analysis of a huge number of rails being removed from the track due to various defects, rail breaks under trains, accidents and wrecks shows the presence of errors in the operation of flaw detection tools, the lack of reliable methods for predicting the survivability of rails and assessing their service life. Therefore, an important area of research on the fatigue strength of rails is undoubtedly the development of methods that allow predicting the service life of rails in transit.

2 Hypothesis and Methods

Methods should be based both on the processing of statistical data of the analysis of defects in rails and the construction on this basis of the function of the probable distribution of possible defects, and on the analysis of fatigue stresses and deformations to predict the durability of rails. These methods, as a rule, provide a good match of the calculated data with the operational ones.

Promising studies in the field of the rail problem are [3] [4]:

- Investigation of the causes of the origin and development of contact fatigue damage in rails using methods of fracture mechanics
- Further study of the effectiveness of grinding rails to remove nascent cracks on the rolling surface and reduce contact stresses
- Development of mathematical models of the mechanism of formation of contact-fatigue defects taking into account the volumetric deformation characteristics of rail steel
- Conducting comparative tests in operational conditions of rails restored by surfacing

A successful solution to the problem of ensuring the safety of train traffic on rails with defects 11 is unthinkable without a full study of the strength characteristics of rail steel.

To assess the structural strength of rail steel, it is necessary to take into account the properties that characterize the strength of steel under the conditions of the stress state that exists in the rail head experiencing the greatest specific pressures from the contact application of the load. To characterize the statistical strength, these properties must be determined at a static load position: for fatigue strength, at a long cyclic contact stress. At the same time, it is advisable to use strength parameters to assess the structural strength of steel, which could be used to calculate the load strength of the rail.

Some issues of the strength of rails with cracks are considered in [6] [8] [9].

The static bending strength of rails is characterized by two parameters: the load at which a residual deformation appears, and the load at which the rail is destroyed. At the same time, parameters such as the bending boom and the work of destruction are measured and calculated. The tests are carried out by applying a load to the middle of the span of the rail mounted on the supports. In, an analysis of the influence of the depth of the gouges formed on the rails of three metallurgical plants on the relative values of the fracture work, load and deflection boom during static bending of the rail is carried out. The depth of defects 11 ranged from 3 to 10 mm. The test results showed that with an increase in the depth of the gouges in this interval, the energy intensity, power intensity and plasticity decrease, and the energy intensity and plasticity decrease more (by a maximum of 90-95%) than the power intensity (by a maximum of 40-45%), but in no case, even with a depth of gouges equal to 10 mm, the level of characteristics was not achieved, in which brittle destruction of rails occurs. The length of the gouges practically did not affect the characteristics of the damaged rails.

The rapid increase in the force action from the wheels of the rolling stock on a particular section sharply raises the contact stresses and should be taken into account when assessing the level of the latter in the rails. However, in some cases, the dynamic impact of the wheel on the rail may threaten traffic safety due to a very dangerous brittle fracture, especially if there is an acute stress concentrator on the surface of the rail, which is defect 11 and the accompanying internal longitudinal crack. Therefore, the assessment of the brittle strength of rails with dents is of great interest.

The ability of the rail to perceive single and multiple shock loads is characterized by its total energy intensity, which is evaluated during tests under conditions of impact bending loading. The rail is laid head down on two supports spaced apart at a distance of 1 m and struck in the middle of a meter span with a 1,000 kg striker raised to a certain height. The energy intensity is characterized by the magnitude of the destruction work in joules.

Changes in the brittle strength of rails with gouges from 1 to 4 mm deep were studied by Kargapoltsev S.K. [7] for non-hardened rails P65. Studies have shown that at a temperature of -20°C, the brittle strength of rails decreases from 60-20 kJ to 10-15 kJ and does not depend much on the depth of the gouges. The presence of defects 11 with a depth of 1-2 mm reduced the copra strength at the same temperature of the bulk-heated rails by 6-11 times from 60 (P50) and 160 kJ (P65) to 8 (P50) and 15-25 (P65) kJ.

The next criterion for the structural strength of rails is flexural fatigue strength. Tests of rails under bending-cyclic loading are closest to the conditions of their operation on the way. Due to the fact that on the way the rail head is subject to an alternating cycle of stresses and tensile stresses in this case are the most dangerous, currently the method of testing rails under bending cyclic loading with the location of rail samples head down in the stretching zone is generally accepted. A pulsating

bending load is applied to a rail laid head down on supports located at a distance of 1 m from each other in the middle of the span with a cycle asymmetry of 0.1. A measure of the assessment of flexural fatigue strength is a limited endurance limit or the maximum load that the rail can withstand without destruction up to a given number ($2 \cdot 10^6$) cycles of its application.

Also, for rails with contact-fatigue defects, survivability is determined, i.e., the number of loading cycles from the beginning of the formation of a fatigue crack before the rail breaks brittle from this crack. 100,000 loading cycles are accepted as the norm of survivability for bulk-loaded rails. It is believed that during this time, the path between the two aisles of the flaw detector car is safely operated [10].

So, in the field of studying the structural strength and survivability of rails with defect 11, questions on the effect of gouges with a depth of more than 4 mm on the brittle strength and cold breakage of non-thermally strengthened and hardened rails have remained unexplored. The effect of gouges on the fatigue strength and survivability of rails is practically not covered. Only after considering these issues, it is possible to establish criteria that determine the safe movement of trains on rails with a defect 11.

2.1 Testing of Rails with Spoll under Cyclic Loading

The purpose of the tests was to determine the influence of vertical dynamic load on the survivability of rails with dents of various sizes, provided that the rail is located in the stretch zone.

For this purpose, 14 thermally hardened rails from different manufacturers were selected: Nizhnetagilsky and Azovstal, having a tonnage output from 190 to 592 million tons gross, full-profile samples with a length of 11.2 m with a defect 11 in the middle with a depth of 4 to 7 mm with a length of 60 up to 120 mm. The rails were removed from sections with a radius of 600 to 900 m with crushed stone ballast and wooden sleepers, on which rolling stock with axial loads of 270 kN was handled.

In order to determine the quality indicators of rail steel, its mechanical characteristics were studied and summarized in Table 1. It was found that the rails of all manufacturers, despite the difference in the missed tonnage, did not have a significant effect of operation on reducing the characteristics of the metal of hardened rails compared to the metal of rails in the initial state [11]. This confirms the conclusion [12] on the preservation of the service properties of the bulk metal of heavy types of rails.

The tests were carried out on a PDM-PU-200 (hydraulic pulsating machine) with a vertical pulsating load with an asymmetry coefficient of 0.1, a 1.2 m long sample with a notch in the middle was mounted on two supports with a design span of 1 m head down. The maximum load was 350 kN. The choice of such a load was due to

the fact that the endurance limit for new volume-hardened rails of the P65 type is 420-450 kN [13], and for the same rails with internal longitudinal crack – 280-320 kN. Therefore, it was decided to test the rails at a load average between the endurance limit for old-fashioned rails with a defect and new ones. The tests were carried out until the complete destruction of full-profile samples. The mechanical properties of the rails during the tensile and impact bending test must comply with the standards specified in Table 1, where $\sigma_н$-temporary resistance, σ_t- yield strength, $б_1$-relative elongation, ψ_1- relative narrowing, $a_н^{-40^0}$-impact strength.

Table 1
Mechanical characteristics of steel rails of type P65

Average operating of million tons gross	Cross sections	Mechanical characteristics					
		$\sigma_н$, MPa	σ_t, MPa	$б_1$, %	ψ_1 , %	$a_н^{-40^0}$ mJ/ m^2	hardness, HB
266	1	1340	1046	12	32	0.22	366
	2	1332	991	13	32	0.22	363
	3	1283	976	15	35	0.22	342
	4	1259	992	18	37	0.20	339
342	1	1388	1027	13	33	0.20	360
	2	1330	980	14	32	0.19	358
	3	1284	972	14	34	0.20	340
	4	1240	986	17	38	0.20	345
422	1	1315	990	14	35	0.19	368
	2	1326	967	15	37	0.20	376
	3	1283	980	15	32	0.20	358
	4	1240	986	16	35	0.21	340

The results of the experiment are shown in Table 2.

Table 2
Characteristics of samples of thermally hardened rails of type P65 with defects 11 and the results of cyclic tests

Sample number	Missed tonnage, million tons gross	Type of heat treatment	Defect size: depth/length, mm
1	190	Hardened	6/90
2	228		6/100
3	366		6/80
4	198		7/100
5	282		6/70
6	329		4/120
7	366		4/70
8	392		6/100

9	213		6/100
10	359		7/100
11	230		6/100
12	276		5/80
13	590		7/60
14	592		6/70

Analysis of the test results makes it possible to conclude that rails with gouges from 4 to 7 mm deep have sufficient survivability.

All samples, tonnage and type of heat treatment, have withstood the standard number of cycles to complete destruction, allowing them to be operated safely on the way.

Moreover, samples № 2, 7, 10, 11, 12 from Table 2 showed results 10-15 times higher than the standard. There was no direct correlation between the geometric dimensions of defects 11 and the survivability of rails. For example, sample No. 10 with a depth of a chink of 7 mm withstood 1178900 loading cycles before destruction, and sample No. 7 with a smaller defect depth (4 mm) accumulated 507700 cycles. The tests have shown that in the presence of a stress concentrator, such as a chink, the missed tonnage within the normative does not affect the survivability of the rails.

The study of the destroyed samples showed that the vast majority of fractures do not begin from the place of application of the load in the middle of the sample, where the dents had the greatest depth, but from the zones of the ends of the internal longitudinal crack located beyond the edges of the defect or from the edges of the dents themselves.

This indicates that microcracks present in the internal longitudinal crack propagation zone, as well as places of sharp changes in the geometry of the rail surface, are stress concentrators and additional studies of the stress state at these points are required.

The conducted studies reveal the picture of the stressed state of the metal of the rail head in the defect zone 11, the conditions of the transition of the internal longitudinal crack into the notch, show that with deep occurrence of longitudinal cracks, the cracks can reach significant sizes. If there is a geometric unevenness on the side face of the rail, which are dents, the probability of rolling the wheel onto the rail increases. It is necessary to calculate the trajectory of the wheel when rolling it onto the rail to determine the maximum permissible sizes of defects.

2.2 Calculating the Value of the Maximum Allowable Sizes of Defects 11

The works of H. Heyman [13] [14], F. Birman, [15] are devoted to the issues of wheel stability and improvement of the theory of calculations of wheel creeping onto the rail.

The conditions of the ultimate equilibrium of the wheelset are fulfilled at the moment when the left wheel on the rail is slightly raised and rests on it with its crest [17] [18]. The following forces and moments will be external to the rail threads:

1) Full dynamic vertical load P_1 transmitted at point D from the left wheel to rail A
2) The full dynamic load P_2 transmitted at the same moment from the right wheel to rail B
3) The moment M_1 acting on rail A; this moment arises due to part of P_1^P the dynamic force P_1 is applied to the neck of the axis, $M_1 = P_1^P\, a_1$
4) The same moment M_2 transmitted to rail B ; $M_2 = P_1^P a_2$
5) The dynamic frame force Y_p applied at a distance e from the point of contact of the left wheel with rail A. In this case, it is usually assumed that e is equal to $r_w + r_n$, where r_w is the radius of the wheel; r_n is the radius of the axis neck

The reactions of the rail threads are indicated; N_1 is the reaction of rail A normal to the U-U plane; N_2 is the vertical reaction of rail B; F_1 is the component of the tangential reaction of rail A acting in the U-U plane and assumed to be equal to $f_1 N_1$, where f_1 is the coefficient of friction of sliding the wheel crest along rail A in the plane U-U. The component of this reaction acting in the plane of the drawing, $F_1 = f_1 N_1 \cos\gamma_1$; F_1 is the component of the tangential reaction of rail B, assumed to be equal to f_2N_2, where f_2 is the coefficient of friction of sliding the wheel along rail B. The component of this reaction acting in the plane of the drawing, $F_2 = f_2N_2 \cos_{\gamma 2}$.[10].

The friction forces $f_1 N_1$ and f_2N_2 act in the opposite directions of sliding of the respective wheels. In the absence of longitudinal components of these slides, the angles γ_1 and γ_2 of the sliding direction with the longitudinal axis of the corresponding rigid base are zero; the larger these components, the larger the angles. In this calculation, the taper of the wheels is neglected (Figure 2).

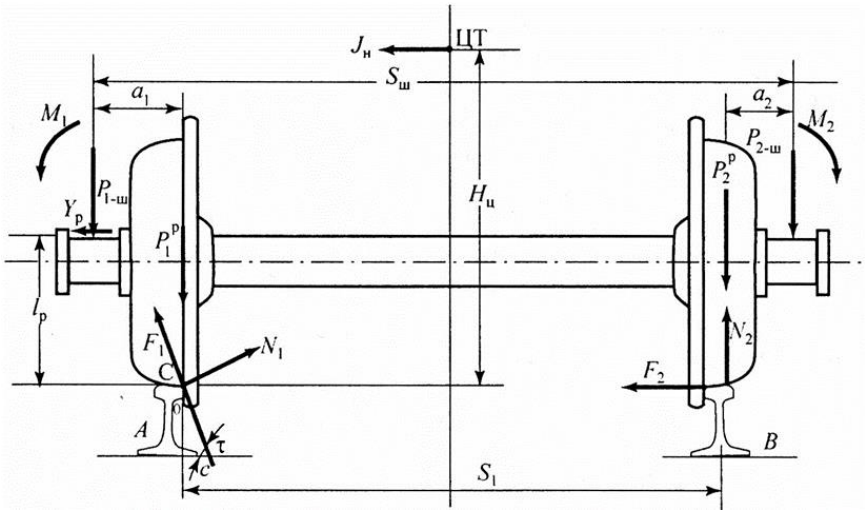


Figure 2
Diagram of forces acting on the wheel

Let us consider the conditions of the ultimate equilibrium of the wheelset at the moment when the wheel rests at point D on rail A with the rectilinear part of the ridge and tends to go down. Let's make up the equilibrium conditions of the wheelset in the form of projection equations on the U-U axis and on the axis perpendicular to it, and the equations of moments relative to the point D [10]:

a) $(P_1 + P_2 - N_2)\sin\tau = F + (Y_p + F_2)\cos\tau$

$b)$ $(P_1 + P_2 - N_2)\cos\tau + (Y_p + F_2)\sin\tau = N_1$ (1)

c) $M_1 + Y_p e = (P_2 - N_2)S_1 + M_2 + F_2 h_2$

The shoulder of force F_2 relative to point O is designated h_2 and is assumed positive if force F_2 passes below point D. From equation (1 c) is N_2, and from (1 b) is N_1. In order for the wheel to not roll the ridge onto the rail, the left part of equation (1 a) must be equal to or greater than the right part. The coefficient of stability against the creeping of the wheel crest on the rail

$$K = \frac{(P_1+P_2-N_2)\sin\tau}{F_1+(Y_1+F_2)\cos\tau} \quad (2)$$

At $K \geq 1$ there will be no rolling. It can be seen from equation (2) that by reducing the force f_1 or, other things being equal, the coefficient of friction f1, the coefficient of stability K can be significantly increased.

Therefore, lubrication of wheel ridges and side faces of rail heads (lubrication) is useful not only to reduce lateral wear of rails and wheel ridges, but also to increase traffic safety. Note once again that the values of P_1, P_2 and Y_p should be taken dynamically, the most unfavorable, but in really possible combinations.

When determining the stability coefficient K from equations (1b, c), the values of forces N_1 and N_2 were found, from equation (1 a) – the coefficient K.

To determine the possibility of rolling the wheel onto the rail, it is necessary to compare the vertical forces transmitted to the wheelset by the superstructure of the carriage and the frame force [109]. The safety condition under which the sliding of the wheel crest down is ensured will be the force ratio $H/Q \leq A$:

$$A = \frac{MB_1-(S_0-b_1)tg(\beta-\phi)}{S_0-(r_{\text{Ш}}-r)[tg(\beta-\phi)+\mu} - \frac{Q_2^+[\mu b_2(S_0-b_2)tg(\beta-\phi) \quad]}{Q_1^+\{S_0-(r_{\text{Ш}}-r)[tg(\beta+\phi+\mu) \quad] \quad \}} - \frac{S_0\{J_2\mu-J_1tg(\beta-\mu)+0.5 \quad \}mg[r-tg(\beta-\phi) \quad]}{Q_1\{(S_0-r_{\text{Ш}}+r)[tg(\beta-\phi)+\mu \quad]\}}, \tag{3}$$

Where,

m - The mass of the wheelset

J_1 and J_2 - Vertical inertia forces of the wheelset

S_0 - The distance between the rolling circle of the wheels of the wheelset

β - The angle formed by the crest of the wheel and the horizontal plane

b_1 and b_2 - The distance from the point of contact of the rolling wheel to the points of application of forces Q_1^+ and Q_2^+

μ - Coefficient of sliding friction

$\phi = arctg\mu$

Then the ratio $\frac{H}{Q} \div A = \eta$ will be the margin of stability at the entrance of the wheel to the rail.

To ensure traffic safety, it is necessary that the value of η be greater than 1. However, even if η is less than 1, it should not be concluded from this that a wheel derailment is inevitable. Not only the necessary, but also sufficient conditions must be met for the wheel to depart. During time t, the wheel should not shift by the value of the b_{ad}. If this condition is met, there will be a convergence. Depending on the length and depth of the gouge, the time of rolling the wheel onto the rail will be reduced and, consequently, the probability of a derailment will increase. Condition $y \leq b_{ad}$ is a sufficient condition to ensure traffic safety. In order to determine the cross-section of the wheel relative to the side surface of the rail, it is necessary to make a condition taking into account the components caused by rolling dw_y and sliding du_y [10].

$$dy = dw_y + du_y \tag{4}$$

Assuming that:

$$dt = \frac{1}{v} dx$$

where,

v – the constant speed of the wheel, we get:

$Y^{l} = w_{y}^{1} + u_{y}^{1}$

The value of u_y and the expression (4) is the relative transverse sliding of the wheel

$$U_y = \frac{1}{v}\frac{dU_y}{dt}. \quad (5)$$

Differentiating (4) by x we get

$$Y^{11} = u_{y}^{11} + w_{y}^{11} \quad (6)$$

The curvature of the trajectory traversed by the wheel during rolling is determined by the equality:

$$Y^{11} = \lambda^2 y + \psi^1 - d_0 \quad (7)$$

Where,

$\psi = \frac{1}{v}\frac{d\psi}{dt}$ – the relative angular slip

ψ – the angle of rotation of the wheel during angular sliding

λ and d_0 – variables determined by the relations

$$\lambda = \sqrt{\frac{tg\alpha + tg\beta}{2S_1 r}} \quad (8)$$

$$d_0 = \frac{h - y_0 tg\beta}{2S_1 r} \quad (9)$$

where,

α and β	are the angles formed by the banding part and the crest of the wheels in the horizontal plane
S_1	distance between wheelset circles
γ	the radius of the wheel in the riding circle
h_0	the amount of lifting of the oncoming wheel at the initial moment of contact of the ridge and the rail, counted from the position when both wheels touch the rails on the average riding circles
g_0	the transverse displacement of the wheel from the point of contact of the middle circle of the wheel with the rail before the onset of the ridge

Substituting (8) into (7), we get:

$$y^{11} = U_{н}^{11} - \lambda^2 y + \psi^1 - d_0. \quad (10)$$

Denoting f(x)= $U_y^{II} = \psi^I - d_0$ we obtain the following inhomogeneous differential equation that determines the trajectory of the wheelset when the wheel crest is rolled onto the head of a flat rail:

$$y^{11} + \lambda^2 y = f_0(x) \quad (11)$$

Using the Laplace transform, we move from equation (11) to the operational equation:

$$p^2 y - c_1 p - c_2 + \lambda_2 y = F(p) \quad (12)$$

where $y(p) \rightarrow y(x)$ $c_1 = y_0$ $c_2 = y_0^1(0)$

$$y(p^2 + \lambda^2) - (c_1 p - c_2) = F(p) \quad (13)$$

$$u_y = \frac{F(P)}{p^2+\lambda^2} + \frac{c_1 p + c_2}{p^2+\lambda^2}, \quad (14)$$

where $F(P)$ – function image $f(x)$, $F(P) \rightarrow F(x)$

$$\frac{a}{p^2+a^2} = \sin\lambda t \quad (15)$$

$$\frac{p}{p^2+a^2} = \cos\lambda t \quad (16)$$

We apply the folding formula:

$$F_1(P)F_2(P) \rightarrow f(q)f_2(1-q)dq \quad (17)$$

considering that:

$$Y = \frac{1}{p^2+\lambda^2} F(p) + \frac{c_1 p}{p^2+\lambda^2} + \frac{c_2}{p^2+\lambda^2}, \quad (18)$$

we get the expression:

$$Y = \int_0^t f_0(q)\sin\lambda(x-q)dq + c_1\cos\lambda x + c_2\sin\lambda x \quad (19)$$

The derivatives of the integration constants c_1 and c_2 in expression (19) can be determined based on the fact that at the moment when the wheel begins to run into the rail, the derivative of the function describing the trajectory of the wheelset is equal to the angle of approach χ_0:

$$y|y_{x=0} = 0 \cdot y^1|y_{x=0} = \chi_0 \quad (20)$$

using (20) we get $c_1=0$, $c_2 = \frac{\chi_0}{\lambda}$. Substituting into the expression (19) we have:

$$y = \frac{1}{\lambda}\left[\int_0^t f_0(q)\sin\lambda(x-q) + \chi_0\sin\lambda x\right] \leq b_{\partial on}. \quad (21)$$

This will be a sufficient condition that ensures that when the wheel enters the rail, there are no irregularities in the path and on the rail in the plan. If we take into account that there is a notch on one rail thread $m = m(x)$, then the initial equation (7) will have the form:

$$y^{11} + \lambda^2 y = u^{11} + \psi^1 - m^{11} - d_0 - \frac{mtg\alpha}{S_1 r} \tag{22}$$

To solve equation (22), we first determine the sliding forces caused by the forces in contact between the wheel and the rail when the wheel hits the rail. The resulting force y and Q will be the force:

$$P = \sqrt{y^2 + Q^2} \tag{23}$$

The angle between the direction of the force and the horizontal plane will be as follows:

$$\varphi = \text{arctg}\,\frac{Q}{y} \tag{24}$$

The force P can be decomposed into two components: the first component N is directed normal to the contact pad, and the second F_k is located in the plane of the contact pad, and the conditions are met for these forces:

$$N = P\sin(\beta + \phi_0);\ F_k = P\cos(\beta + \phi_0) \tag{25}$$

In addition, for slides, you can write:

$$u\ {}^{x}_{y} = u_k \cos\beta \tag{26}$$

As a first approximation, it can be assumed that the contact pad at the point of contact of the wheel and the rail is located in the Q plane passing through the x and k axes and inclined at an angle β in the horizontal plane.

The forces F_k and F_x act on the contact pad in the Q plane, and the forces N act on the normal. In the case under consideration, the longitudinal force of F_x can be neglected. Using the formulas given in [11], we determine from the known components of the tangent and normal forces relative to sliding in the direction of the K axis:

$$u_k = \frac{3\eta\mu N\psi}{\pi c l G} \tag{27}$$

Where,

η	The parameter of relative sliding in the direction of the K axis
μ	The coefficient of friction at rest
c and l	Respectively, the large and small semi–axes of the contact half-ellipse
ψ	The value depending on the ratio of the semi-axes of the contact ellipse
G	The shear modulus of elasticity

ψ is defined from the following expressions:

The complete elliptic integrals for $|k| \leq 1$ will have the form:

$$\left.\begin{aligned} B &= \int_0^{\pi/2} \frac{\cos^2\theta d\theta}{(1-k^x\cos\theta)^{1/2}} \\ C &= \int_0^{\pi/2} \frac{\sin^2\theta\cos^2\theta d\theta}{(1-k^x\cos^2\theta)^{1/2}} \\ D &= \int_0^{\pi/2} \frac{\sin^2\theta d\theta}{(1-k^x\sin\theta)^{1/2}} \end{aligned}\right\} \quad (28)$$

by

$$l \geq c \cdot k = \sqrt{1 - c^2/l} \quad (29)$$

$$\psi = (D - \delta C)\frac{c}{l} \quad (30)$$

The values of ψ at different *l/c* ratios are given in Table 3.

Table 3

The value of depending on the ratio of the small and large semi-axes of the ellipse

l/c	0/075	0/1	0/2	0/3	0/6	0/8	1/0
ψ	0.956	0.980	0.955	0.891	0.830	0.747	0.727

To determine the values of the semi-axes of the contact ellipse, one can use the Hertz theory [8]. The minor semi-axis of the contact ellipse is determined by the formula:

$$C = \sqrt{\frac{3(1-\delta)gL\rho N}{3\pi G}} \quad (31)$$

where G – Poisson's ratio.

$$g = \frac{c}{l} \quad (32)$$

where ρ – the characteristic of the dimensions of two elastic bodies (wheels and rails), the value of which is determined by the formula:

$$\frac{1}{\rho} = \frac{1}{4}\left(\frac{1}{R_x^+} + \frac{1}{R_x^-} + \frac{1}{R_k^+} + \frac{1}{R_k^-}\right) \quad (33)$$

Where $\frac{1}{R_x^+}$ and $\frac{1}{R_x^-}$ – the main radii of curvature of the rail at the point (x=0, K=0);

To determine the values of *g* and *L* included in the formula (31), it is necessary to first determine the value of γ :

$$\gamma = \arccos\left\{\frac{\rho}{4}\left[\left(\frac{1}{R_x^+} + \frac{1}{R_x^-}\right) - \left(\frac{1}{R_k^+} + \frac{1}{R_k^-}\right)\right]\right\} \quad (34)$$

and then we will use the data from Table 3 [6].

The magnitude of the small semi - axis of the contact ellipse can be written as:

$$C = \sqrt[3]{DN} \tag{35}$$

where

$$D = \frac{3(1-\theta)\rho L g}{2\pi G} \tag{36}$$

Substituting (25) – (35) into (26) we get:

$$U^{11} = A\left(N^{1/3}\frac{d\eta}{dx} + \frac{\eta}{3N^{2/3}} \cdot \frac{dN}{dx}\right) \tag{37}$$

$$A = \frac{3\mu\beta g\cos\beta}{\rho G D^{2/3}} \tag{38}$$

$$\eta = \tau\frac{F_k}{\sqrt{F_k^2 \quad {}^2 + F_x^2}};\ F=0,\ \eta = \tau \tag{39}$$

where τ – the value characterizing slippage, is determined from the functional relationship between the value of f and τ [8]

Knowing N and defining t and $\frac{dt}{df}$, we can find U^{11} by the formula:

$$U = AN\left(\frac{dt}{df} \cdot \frac{df}{dx} + \frac{d\tau}{dN} \cdot \frac{dN}{dx}\right) \tag{40}$$

In the case under consideration, we assume that the transverse and vertical forces have the form of a rectangular pulse with a long vibration length on the rail, having the form of a half-wave of a sine wave:

$$m = A\sin\varpi x\left(0 \le x \le \frac{\pi}{\omega}\right)\ \varpi = \frac{\pi}{L_B} \tag{41}$$

where,

A_1 The amplitude of the unevenness value

L_B The length of the puncture

Consider a section of track with a gouge on the rail, on which the wheel runs. Considering that $x \geq L_в$ is in, and the wheel runs at an angle of χ_0 and substituting (41) and (42) in (22), we get:

$$y = \frac{u^{11} + \chi_0}{\lambda}\sin\lambda x - \frac{d_0}{\lambda^2}(1 + \cos\lambda x) + \frac{A_1\varpi^2(\varpi\sin\lambda x - \lambda\sin\varpi x)}{\lambda(\varpi^2 - \lambda^2)} \text{ by } x \leq x_0 \tag{42}$$

It was noted in [11] that for ratio values ranging from 1.48 to 9.56, the stability of the wheel depends not only on the ratio of forces F_y and F_x, but also on the angle of attackχ_0 and the duration of action x_0. Thus, given the values of A_1, χ_0, the calculations of y wheel edges of 10 mm are performed. In this case A_1 is the projection of the depth of the notch H, measured at an angle of 45° to the side face of the rail head, which corresponds to the depth t. The data of the calculations χ_0- angle of atack, radian, x_0- the time of action duration,secunds, carried out on the computer are shown in Table 4.

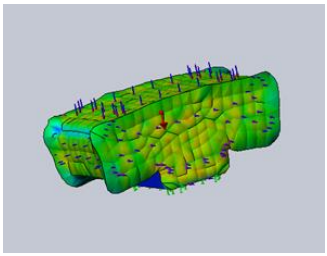
Table 4

Acceptable values of chip on the rail depending on the angle of occurrence

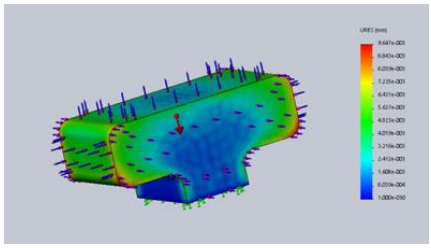
χ_0, rad.	x_0, sec.						
	4	**6**	**8**	**10**	**12**	**14**	**15**
0	700	600	520	510	490	450	430
0,006	580	520	430	430	420	400	390
0,0088	530	450	420	420	400	390	380
0,01	470	420	410	400	380	370	360
0,012	420	390	380	370	330	350	330
0,016	350	340	340	340	320	320	320
0,02	320	300	290	280	270	260	230

These values of geometric dimensions are the maximum permissible at which it is possible to roll the wheel crest onto the rail. The data given in Table 4 show the law of changing the permissible dimensions of the gouges from the conditions of entering the crew into the track – the permissible dimensions of defects 11 must be set at the maximum possible angles of the wheel on the rail. In general, the greatest angle of the wheel on the rail should occur in curves of small radius. However, in case of malfunctions of rolling stock and the movement of long-composite trains "herringbone" [10], the maximum angles of the run-in may be in the band of curves or straight sections of the track. Therefore, the maximum allowable dimensions of defects 11, set at the angles of occurrence $\chi_0 = 0.02$ rad, will be determined for all sections of the path regardless of curvature. Considering the most unfavorable case of interaction of the wheel-rail system, it is necessary to establish the maximum permissible dimensions of defects 11 of the conditions for rolling the wheel crest onto the rail: a length of 25 mm at a depth of 3 mm. If these dimensions are exceeded, the rails must be replaced as planned. Immediate replacement is impractical due to high costs and the random nature of the combination of factors that lead to the wheel creeping onto the rail, depending both on the condition of the rolling stock and its load, and on the condition of the rail track.

According to the method of calculating cracks, a mathematical experiment was carried out using the Solidworks program. The main characteristics of the study were temperature and load, the results are shown below (Figure 3)



a



b

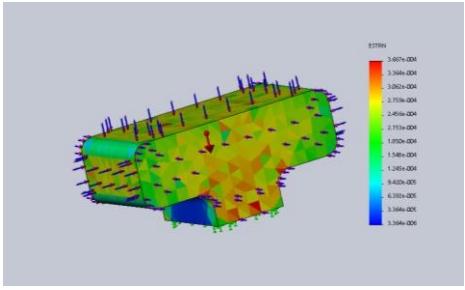
c 

Figure 3

Impact of temperature and load on the rail depending on the depth of the crack

a- depth of the crack 1mm,b- depth of the crack 2 mm, c- depth of the crack 3 mm and more

As we can see from the analysis on lines with a sufficiently high load, where pointed rolling of wheels is more common, the probability of rolling stock derailment increases in the presence of defects 11 with a depth of more than 3 mm and a length of more than 25 mm, which proves the developed calculation methodology.

Derailment of wheels is inevitable if two conditions are met: necessary when the coefficient of stability margin at the entrance of the wheel to the rail is less than 1 and sufficient when the wheel shifts by a value bдоп equal to the projection of the wheel edge on the horizontal axis.

The determination of the transverse displacement of the wheel is reduced to the compilation of an ordinary inhomogeneous differential equation, which includes components due to rolling and sliding. This allows considering a sufficient condition that ensures traffic safety when the wheel enters the rail, which does not have defects in the riding surface.

Conclusions

In this article we described methods of calculating the value of the maximum allowable sizes of defects 11 at different angles of the wheel running into the rail with different shapes and sizes of wear of the wheel and rail edges.

The presence of a spoll on the rail raises the likelihood of the wheel rolling onto the rail. To solve the problem of the maximum allowable size of the gouges, a theoretical study of the interaction of a wheel with various wear with a rail head having a Defect 11 was carried out. The analysis used previously conducted studies [10] with the introduction of a component into the calculations in which the dents are represented as half-waves of a sine wave.

The given calculation method makes it possible to establish the permissible dimensions of Defects 11 according to the conditions of rolling the wheel crest onto the rail with the most unfavorable combination of interacting parts. On lines with a sufficiently high load, where pointed rolling of wheels is more common, the probability of rolling stock derailment increases in the presence of Defects 11 with

a depth of more than 3 mm and a length of more than 25 mm. On road sections where there is no heavy train traffic, less stringent restrictions should be set on the size of the gouges (4 mm depth and 36 mm length). The appearance of defects of 11 large sizes should serve as a reason for the appointment of a planned replacement of the rail, however, it would be more economical to repair the rail in transit.

References

[1] Wu L., Wen Z., Jin X. Finite element analysis of wheel/rail frictional temperature during wheel complete slides on rail. Jixie Gongcheng Xuebao. 2018, T. 44. № 3, pp. 57-63

[2] Zh Arystanov, B. Togizbayeva,. A. Karazhanov,. Zh. Alipbayev,. K. Burkanova Investigation of the impact of operating conditions on the service life of locomotive wheelset tyres 2022 Journal of the Balkan Tribological Association T. 28, № 6, pp 882-896

[3] Sergeev A. I., Suvorov A. A. Improving the electrical characteristics of measuring devices based on rail-to-rail operational amplifiers. Measurement Techniques. 2018, T. 61. № 8, pp. 817-823

[4] Tadić N., Zimmermann H., Zogović M., Banjević M. A low-voltage complementary metal-oxide semiconductor adapter circuit suitable for input rail-to-rail operation. International Journal of Electronics. 2019, T. 97. № 11, C. 1283-1309

[5] Królicka A., Lesiuk G., Mech R., Radwański K., Kuziak R., Janik A., Zygmunt T. Comparison of fatigue crack growth rate: pearlitic rail versus bainitic rail. International Journal of Fatigue. 2021, T. 149. p. 106280

[6] Mao X., Shen G. A design method for rail profiles based on the geometric characteristics of wheel–rail contact. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit. 2018, T. 232. № 5, pp. 1255-1265

[7] Kargapoltsev S. K., Davadorg B. Algoritmic base for prognosis the rail residual resource of the lateral wear of rail. Innovation and Sustainability of Modern Railway. 2019, pp. 539-542

[8] Wang K., Liu P., Zhai W., Huang Ch., Chen Z., Gao J. Wheel/rail dynamic interaction due to excitation of rail corrugation in high-speed railway. Science China Technological Sciences. 2019, T. 58. № 2, pp. 226-235

[9] Ju B., Zhu H., Wang Z., Sun Z., Xu J. Rail contour matching method based on random vibrations of a rail inspection vehicle. Zhendong yu Chongji. 2019, T. 36, № 3, p. 65-69

[10] Shiyao Lu, Jingru Wang, Guoqing Jing, Weile Qiang, Majid Movahedi Rad: Rail Defect Classification with Deep Learning Method, DOI: 10.12700/APH.19.6.2022.6.16, pp. 225-241

[11] Ren M., Dong C., Qin M., Wang B., Han X. A novel slew-boosting circuit for rail-to-rail operational amplifier. Journal of Circuits, Systems, and Computers. 2021, T. 30, № 2, p. 215

[12] Maxim Arbuzov, Serhii Tokariev, Oleksii Hubar, Volodymyr Andrieiev, Volodymyr Suslov: Relaxation of the Elastic Clamp in Rail Fastenings, DOI: 10.12700/APH.19.3.2022.3.17, pp. 219-230

[13] Ljubo Marković, Ljiljana Milić Marković, Srđan Jović, Miloš Milovančević: Evaluation of Alternative Solutions of General Design of Railway Lines with Regards to Environmental Protection, DOI: 10.12700/APH.19.6.2022.6.8, pp. 184-214

[14] Vesna Jovanovć, Dragoslav Janošević, Dragan Marinković, Nikola Petrović, Jovan Pavlović: Railway Load Analysis During the Operation of an Excavator Resting on the Railway Track, 10.12700/APH.20.1.2023.20.6, pp. 79-93

[15] Aditjandra P. T., Zunder T. H., Islam D. M. Z., Palacin R. International Series in Operations Research and Management Science. 2019, T. 226, pp. 413-454

[16] Eller, B., Fischer, S. Tutorial on the emergence of local substructure failures in the railway track structure and their renewal with existing and new methodologies (2021) *Acta Technica Jaurinensis*, 14 (1), pp. 80-103

[17] Slobodan Rosić, Dušan Stamenković, Milan Banić, Miloš Simonović, Danijela Ristić-Durrant, Cristian Ulianov: Analysis of the Safety Level of Obstacle Detection in Autonomous Railway Vehicles, DOI: 10.12700/APH.19.3.2022.3.15, pp. 187-205

[18] Dmytro Kurhan, Mykola Kurhan, Nelya Hmelevska: Development of the High-Speed Running of Trains in Ukraine for Integration with the International Railway Network, DOI: 10.12700/APH.19.3.2022.3.16, pp. 207-218