Some Problems of Dynamic Contactless Charging of Electric Vehicles

Nikolay Dimitrov Mazharov[1], Stefan Milchev Hristov[1], Dimitar Andonov Dichev[2], Iliya Slavov Zhelezarov[2]

[1]Faculty of Electrical Engineering and Electronics, Technical University of Gabrovo, Hadzhi Dimitar Str. 4, 5300 Gabrovo, Bulgaria

[2]Faculty of Mechanical and Precision Engineering, Technical University of Gabrovo, Hadzhi Dimitar Str. 4, 5300 Gabrovo, Bulgaria

E-mails: madjarov@tugab.bg, stefan.milchev@elkosys.com, dichevd@abv.bg, izhel@tugab.bg

Abstract: This paper presents a survey of the main publications and general design requirements and problems of inductive power transfer systems for dynamic charging of electric vehicles (EVs). The main different roadbed geometries, based on a single transmitter track and segmented transmitter coil array have been discussed. Different case studies considering charging scenario, vehicle speeds, power levels and transmitting and receiving coils geometry have been conducted. Some problems about the design of charging station and EV side control system and energy management system have been analysed. A prototype of a charging station has been developed and built to supply inductive power transfer system which delivers 10-35 kW power at an air gap between transmitting and receiving parts of 75-100 mm and horizontal misalignment of ± 200 mm. The results have shown that the system can transfer the specified electrical power at efficiency of about 82-92% and that the inductive power transfer module and its dynamic matching during charging, exhibited high degree of stability under a misaligned (x-y-z) condition and battery state of charge.

Keywords: electric vehicle; contactless charging; transmitting and receiving coil; inductive power transfer; energy management system

1 Introduction

In the past decade, EVs have gained popularity due to concerns about environment pollution with greenhouse gases and a desire to move toward "greener" energy [1, 2, 4, 17]. The latest most popular plug-in technology for EVs has some disadvantages: the charging infrastructure (public charging stations) is vulnerable to weather conditions (rain, snow) and vandalism (stealing the cord, blocking the

outlet). The charging cable can represent a trip hazard, and due to the large amount of power being transferred, it also carries the risk of electrocution.

Contactless Charging (ClCh) technology for EVs improves upon EV convenience and related infrastructure as well as charging safety. EVs that are contactlessly charged are easy to use - the user simply parks or drives the EV into the charging area and allows it to charge. ClCh infrastructure can be built and sealed with no outlets, making it inherently safe from weather, vandalism, and electrocution hazards. One of the most important prerequisite for increasing the share of EVs, using ClCh technologies, is development of well-distributed infrastructure facilities which could allow more frequent charging and shorten the charging time required.

Concerning the contactless charging process (time, power, EV speed), several different techniques have been and currently are under investigation [2, 4, 7, 11, 12, 14, 15, 17]. The following categories could be identified, each having its benefits and disadvantages.

The first one is static charging, when the EV is stopped and parked [12]. Taking into account the duration of the charging time there are two possible solutions. Long duration of static charging time, typically several hours, when the EV is stopped in a garage, parking lot, bus or taxi terminals, etc. The second solution is fast static charging time less than one hour [15].Typically, these charging spots are public. In both cases the charging process is based on some parameters that are specific to the battery of the EV and the capacity of the charging infrastructure.

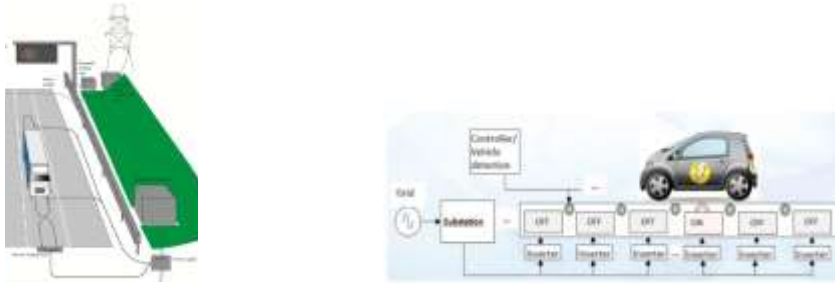
The second category uses charging scenario when the EV is on-route. Also, two solutions are available. The first, when EV stops for a short period of time at the traffic light, bus stop, etc., is called stationary charging [9]. A stationary charging system could be very suitable in urban environments where the exact locations of EV stops could be predicted. The charging time is from seconds to minutes and high power is transferred from the infrastructure to the EV. The studies show [3-6, 8, 17, 18] that when implementing stationary charging the EV's battery can have a considerably smaller volume. The second on-route charging solution uses scenario with movable EV [9, 11]. It is called dynamic charging. The idea is in public urban road infrastructure in some places, where there is speed limitations, for example before crossroads, traffic lights or eventually in highways to have a special road line and zone for charging of slow moving EVs. EV can be charged continuously while in motion and theoretically solve the EV battery problem with unlimited driving range. The vehicle may travel at constant or variable speed in a special lane that hosts the charging infrastructure. The advantage of dynamic charging is that this technology allows when EV passes over charging zone to add energy to batteries. As a result, the longer charging zone the lower battery's capacity is needed, respectively its weight and more important cost. Also, the energy of charge depends on the speed - the higher speed, the less average energy is transmitted.

This paper endeavours to review the available dynamic ClCh technologies for EVs developed by charging technology companies, by car manufacturers, by universities and research institutes. The paper presents the authors' results concerning analysis of power transfer in dynamic mode, overview architecture, primary and secondary side, communication problems on two levels and main features of energy management system. The developed dynamic inductive charging station (ChS) with power P=10-35 kW and frequency f=18-25 kHz is presented. The dynamic infrastructure containing four primary coils and different charging scenarios was tested at vertical air gap between transmitting and receiving coils - 75-100 mm and horizontal misalignment – up to ± 200 mm.

2 State of the Art Dynamic Charging

In dynamic charging infrastructure, vehicles are highly unlikely to be precisely aligned, so the currently agreed solution is the installation of visual signals would be put on the road to help the driver align the EV while driving. Charging a vehicle while it travels would mean that an EV user would not have to make stops to recharge during extended road trips. In fact, travels of hundreds of kilometres would be possible with dynamic charging to obtain portion or uninterrupted energy and to increase EV mileage. The analysis of the current situation in this area shows that the dynamic charging technology is still in R&D phase [4, 6, 7, 8, 9, 11, 14, 15, 23].

Dynamic EV charging approaches predominantly are based on inductive power transfer (IPT) technology and can be mainly categorized into two types based on transmitter array design - single transmitter track Fig. 1a) and separate segmented transmitter coil array Fig. 1b).



a) b)

Figure 1

Dynamic ClCh scenarios: a) single charging pad; b) string of charging pads

The first type consists of a substantially long transmitter track connected to a ChS. The receiver is noticeably smaller than the length of the track. The transmitter track can be from a few meters to several tens of meters long. There are some

drawbacks in this design - the electromagnetic field emitted within the uncoupled area must be suppressed to eliminate harmful exposure and the compensation capacitor should be distributed along the track to compensate large inductance and magnet coupling is fairly low because of the smaller transmitter area covered by the receiver coil resulting in lower efficiency.

Segmented coil array based designs have multiple coils connected to ChS. Transmitter track based systems are easier to control as the track is powered by a single source. Magnet coupling along the track is nearly constant when the EV moves along the track. On the other hand, segmented coil array eliminates field exposure and requirement for distributed compensation as happen in single transmitter track.

Some of the notable achievements in designing dynamic EV charging platforms can be identified as follows. UC Berkeley has conducted a test as a proof-of concept of a dynamic ClCh system for EV based on IPT in the late 70s' [10]. They transferred 60 kW of power through 7.6 cm distance to a passenger bus along 213 m long track. Due to limited semiconductor technologies, the operating frequency of Berkeley system was 400 Hz and only 60% efficiency.

KAIST On Line Electric Vehicle (OLEV) is the first on the market ready to support public transport by dynamic wireless high-power charging [11] - up to 180 kW for tramways and up to 100 kW for buses at 20 cm gap. The generated magnetic field around the EV is < 2.41 μT. The OLEV's technology is presented in Fig. 2.

Figure 2
Wireless solutions for EVs in South Korea

Conductix-Wampfler, a German company [12], has developed a dynamic charging module that works on the principle of a construction kit - depending on the EV (car or bus), a charge of 60, 120 or 180 kW is delivered wirelessly. The charging modules are delivered ready for service, meaning that once the preparations have been made below ground, they only have to be lowered into the shaft. It takes very little time before the inductive charging point is ready to be used by the final users – Fig. 3.

The Conductix-Wampfler's technology, in charging mode, the receiver coil on the bus is lowered to about 40 mm from the ground. This closeness to the charging plate allows the magnetic field to be focused in such a way that stray magnetic

fields remain almost entirely restricted to the immediate vicinity of the coil. Next to the vehicle, the field values are significantly below the thresholds prescribed by ICNIRP recommendations, i.e. generated magnetic field <<6.25 μT [12, 13].

a)

b)

Figure 3

Conductix-Wampfler ClCh technology: a) main components of ClCh station; b) installing a charging module at a bus station with 120 kW power

Bombardier's PRIMOVE system addresses both the static and the dynamic charging needs of buses, cars, and even light rail systems [14]. Currently, Bombardier's dynamic charging has only been applied to light rail systems, using single charging pads built into the track; however, it could be adapted for use with road vehicles. The system's roadside components for buses include: transmitting coils which provide the inductive magnetic field; shielding to prevent electromagnetic interference; a Vehicle Detection and Segment Control (VDSC) cable that identifies PRIMOVE vehicles above the system; a Supervisory Control and Data Acquisition interface which supplies information for system control and diagnostics; and inverter and power supply cables. The on-board equipment includes a power receiver system of pickup cables and compensation capacitors, inverter, a battery, and a VDSC antenna. Bombardier's PRIMOVE system is currently being used in public transportation by buses in Braunschweig, Germany.

There are some EU projects funded by FP7 program that investigated and developed contactless dynamic charging solutions [8, 9, 15]. The FastInCharge [8] solution appears as being ambitious and innovative regarding the fact that it is applicable both to static and dynamic charging. The output power of the FastInCharge technology is around 35 kW (P_{MAX} = 50 kW), which is one of the most powerful technologies that have been designed for mini busses. The air gap is from 75 mm to 100 mm. The system efficiency is close to 92% at zero misalignment between coils. The authors of this paper were responsible researchers and developers of IPT module and power electronics of two charging stations (fast static and dynamic) and both types of charging stations have been tested and validated in Douai, France [8].

However, dynamic charging introduces some other design challenges and problems. It is necessary to track the receiver coil position and switch the appropriate ChS, respectively coil, when the EV moves along the array. In addition, the distance between transmitting coils in horizontal direction needs to be carefully optimized. These coils cannot be kept too close to each other due to

two reasons. Firstly, negative mutual inductance between adjacent transmitter coils could generate negative current stress when several transmitting coils are supplied simultaneously. Secondly, design cost will be increased with many transmitting coils in a given length of the track. Connecting source converters to multiple coils is also a design issue. Several transmitter coils can be connected to a single power converter in parallel, or there can be one converter connected to each coil. In [8, 16, 18] a variant has been proposed in which several transmitting coils are powered by one ChS, consecutively by using electronic switches. On the other hand, the distance in vertical direction is also important. When the distance between transmitting and receiving coils is quite large, efficiency reduces quickly.

Additionally, a few more problems could be added in the implementation of dynamic EV charging systems such as strategies for dynamic IPT charging coil misalignment compensation, intelligent control on the transmitting charging side and receiving EV side levels, communication between these two levels and energy management of the whole dynamic charging process. Foreign-object detection and electromagnetic compatibility (EMC) issues are important safety requirements and have to be respected during development of every dynamic charging infrastructure [13]. Interoperability and standardization of dynamic contactless EV charging systems, including the proper terminology, are still opened issues, too.

3 Problems of Dynamic Charging

3.1 Analysis of Energy Transfer

The critical overview of the previous paragraph has shown, that for the purpose of the discussed dynamic infrastructure, the solution with separate segmented transmitter coil array was selected (Fig. 1b). The main advantage is that it is suitable for dynamic and also for static EV's charging, what was an important requirement specified in FiC project [8]. Therefore, there are some geometrical and electrical parameters limitations of the transmitting and receiving coils. For example, the receiving coil is integrated in the EV and because of that positional and dimensional limitations are available. On the other hand, the geometrical and electrical parameters of transmitting coils must coincide with receiving coil in order to meet safety exposure requirements. The overall view of the developed dynamic infrastructure is shown in Fig. 4, where n numbers of transmitting coils are supplied by one ChS.

When two EVs sequentially move at a certain distance in a row in order to avoid the overlapping of two receiving coils from the group of transmitting coils the following condition has to be satisfied:

$x_n - x_{EV} < x_{dist}$, where (1)

$x_n - x_1 = (n-1).(x_1 + x_3 + x_4)$ is the maximum distance between first and last coil supplied by one ChS; x_{EV} is EV length; x_{dist} is the distance between two EVs; n is the number of coils connected to one ChS; x_n is the distance from the beginning of the first coil to the end of the last coil, supplied from one ChS; x_1 is coil length; x_3+x_4 is the distance between transmitting coils.

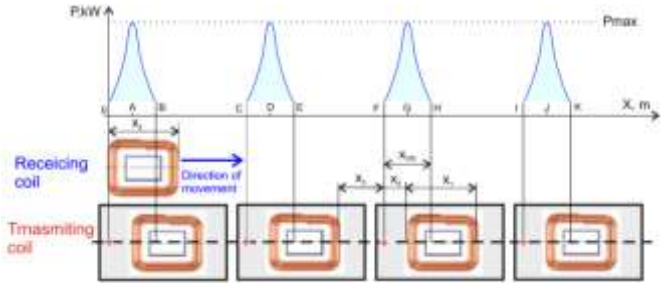


Figure 4
Dynamic infrastructure and energy transfer level

For determination of transmitted output power (P_{OUT}) was used the fundamental expression, given by [20].

$$P_{OUT} = \frac{P_{LOSS}.(k.Q)^2}{2.[1+\sqrt{1+(k.Q)^2}]} \quad , \tag{2}$$

where Q is equivalent IPT quality factor and P_{LOSS} are coils losses, which are divided between transmitting and receiving coils according to the ratio of both quality factors. We adopted (2) for the ratio of P_{LOSS} toward the output power P_{OUT} and it is equal to:

$$\lambda = \frac{P_{LOSS}}{P_{OUT}} = 2.\left(1+\sqrt{1+(kQ)^2}\right)/(kQ)^2 \, , \tag{3}$$

It is obvious, that to achieve better efficiency it is necessary that λ<<1. Based on (3) there was made an analysis whose results are presented in Fig. 5 at k = *0.05-0.5* and Q=*10-50*. The value of losses increases dramatically at $k < 0.1$ and $Q < 10$. The implemented analysis proves that for the reliable operation of an IPT module it is necessary that $k > 0.3$ and $Q > 20$, in other words $k.Q = 6 \div 10$. Therefore, (3) could be simplified using above values:

$$\lambda = \frac{P_{LOSS}}{P_{OUT}} \approx 2.\left(1+\sqrt{\beta_{IPT}{}^2}\right)/\beta_{IPT}{}^2 \approx \frac{2}{\beta_{IPT}} \tag{4}$$

where: $k.Q = \beta_{IPT}$, $\beta_{IPT} \gg 2$.

If the magnet coupling k has low value ($k < 0.3$) it is possible by optimizing Q (increasing the inductance) to keep the ratio $k.Q \gg 6$. Otherwise the IPT module will have bad cost efective indicators.

In low powered IPT systems (up to 500 W) is possible to transfer energy at bad magnetic coupling by optimizing the quality factor through correction of windings

inductance. A similar optimization process of Q for IPT electric vehicles chargers is economically unprofitable because of more litz wire and ferrites used and as a result more electrical losses. In some cases, it is appropriate to adjust the length and/or width of the coils.

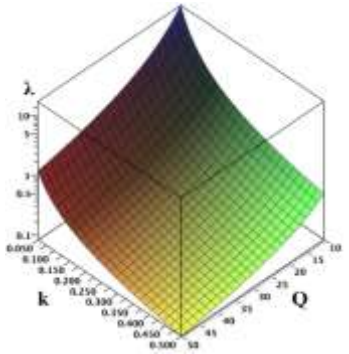


Figure 5
Coil losses for different k and Q

By changing the size of the transmitting coil, at the same turns number, it can be assumed that there is almost proportional change of inductance and its active resistance and thus quality factor remains constant. Therefore, the transferred power according to (2) is changed proportionally in accordance with the magnet coupling, that in a given geometry of the transmitting and receiving coils depends only on their horizontal and vertical misalignment.

According to previous work done in [2-5] and our study [16-18, 19] it was proved, that the optimal relationship between the geometrical dimensions of the IPT coils are $D/x_1<0.25$ (D is the vertical distance between the transmitting and receiving coils and x_1 is the length of the coil), which guarantee magnetic coupling k greater than 0.3 (see Fig. 5) and hence, good energy transfer and efficiency. When D/x_1 is close to 0.25, the efficiency of the wireless module is a maximum of 80%, while at $D/x_1<0.125$, the efficiency reaches 93% [16]. Indirectly, these requirements are used in determination of the distance between the transmitting coils at dynamic charging.

The summary of the above results is shown in Fig. 6. It presents the dependence of efficiency as a function of horizontal misalignment between the coils (0, 100, 200 mm) and a vertical distance between them of 100 mm.

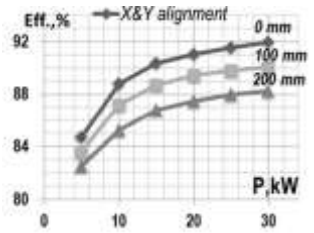


Figure 6
IPT efficiency Vs. output power and X&Y misalignment with 100 mm vertical air gap and power of 10-30 kW

The total max efficiency of the charging station from the mains to the battery at zero misalignment and 100 mm air gap is 90-92%. This efficiency is obtained by the charging station modules in the following way: (a) HF inverter - 97-98%; (b) IPT module 94-95%- primary 96-97% (copper app. 98%, ferrite core app. 98%), secondary 97-98% (copper app. 99%, ferrite core app. 98%); c) output rectifier - 98-99%.

In dynamic charging mode, movable car, misalignment Δx between transmitting and receiving coils has variable value and respectively magnetic coupling k. The dependence between these two parameters, for different transmitting coil dimensions has been investigated through computer simulations, using Ansoft Maxwell software tool (see Fig. 7). For all cases, the receiving coil dimensions are 800/700 mm and the gap between the coils is 100 mm, specified by concrete application [8].

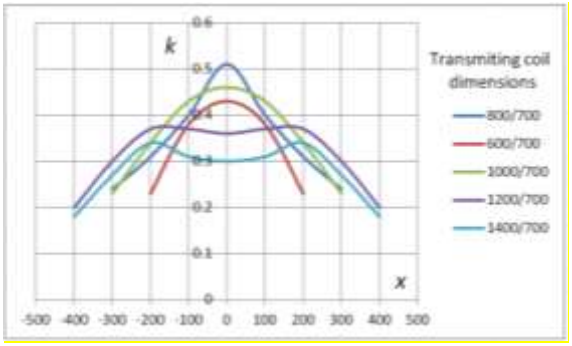


Figure 7
Magnetic coupling k, at 100 mm air gap vs. horizontal misalignment and different transmitting coil dimensions

It is seen that for the same dimensions of transmitting and receiving coils (800/700 mm) and zero horizontal misalignment (x=0), k obtains the maximum value, equal to 0.51 (dark blue curve). The coupling is over 0.3 when x is in the range of -200 mm to +200 mm. The biggest area (-300 mm to +300 mm) of efficient energy transfer is at primary transmitting coil with dimensions 1200/700 mm, where the

maximum coupling is 0.37. Therefore, the maximum instantaneous power will be 0.37 / 0.51 times smaller compared to the dimensions 800/700 mm.

In accordance with Fig. 4 and (2), the average transferred power (that is proportional to k) from the transmitting to the receiving coils is determined by the expression:

$$P = \frac{1}{x_{ON}+x_{OFF}}.\int_0^{x_{ON}} P(x)dx \quad , \tag{5}$$

where $x_{ON} = 2.x_4$ is the distance when overlapping of two coils is less than permissible misalignment and $k>0.3$ and $x_{OFF} = x_1 - x_4 + x_3$ is the not transferred power distance (see Fig. 4).

The determination of average power is carried out after approximation of $P(x)$, with standard geometric shapes [21, 22]:

-in case of a triangle with height P_{MAX} and basis x_{ON}:

$$P_1 = \frac{1}{x_{ON1}+x_{OFF1}}.\int_0^{x_{ON1}} P_1(x)dx = \frac{1}{x_{ON1}+x_{OFF1}}.\frac{P_{MAX1}.x_{ON1}}{2} \tag{6}$$

-in case of trapezium with bases x_{ON1} , $0.7x_{ON1}$ and height P_{MAX1} :

$$P_2 = \frac{1}{x_{ON2}+x_{OFF2}}.\int_0^{x_{ON2}} P_2(x)dx = \frac{1}{x_{ON2}+x_{OFF2}}.\frac{P_{MAX2}.1{,}7x_{ON2}}{2} \tag{7}$$

-in case of a parabola:

$$P_3 = \frac{1}{x_{ON3}+x_{OFF3}}.\int_0^{x_{ON3}} P_3(x)dx = \frac{1}{x_{ON3}+x_{OFF3}}.\int_0^{x_{ON3}} (-C_1.x^2 + C_2)dx =$$

$$= \frac{1}{x_{ON3}+x_{OFF3}}.\left(-\frac{C_1}{3}.{x_{ON3}}^3 + C_2.x_{ON3}\right) \quad , \tag{8}$$

where C_1 a coefficient, characterized the slope and C_2 tip parabola displacement.

The energy that is transmitted to the EV running over n number of transmitting coils at speed V is equal to

$$E = n.P.(t_{ON} + t_{OFF}) = n.P.(x_{ON} + x_{OFF})/V \quad , \tag{9}$$

where t_{on}, t_{off} is the time during which each transmitting coil is switched on / off. It is obvious that $t_{ON}=V/x_{ON}$, $t_{OFF}=V/t_{OFF}$.

Table 1 presents the results of average power value calculated by expressions (2)-(9) in accordance with the magnet coupling k - Fig. 7, the size of the transmitting coil and horizontal misalignment. The dimensions of the receiving coil (800/700 mm), the horizontal distance between each transmitting coils (800 mm) and the gap between the transmitting and receiving coils (100 mm) are unchanged.

From (9) and Table 1 it is obvious that the transferred energy value depends on the EV speed, number of transmitting coils and efficient power transfer distance – x_{ON}. Also, the results of this analysis could be used for calculating the maximum

EV speed in order to receive the necessary energy in a given dynamic infrastructure.

The length of the transmitting coil also determines the value of the transferred power. The highest instantaneous power is transferred at the same dimensions of the transmitting and receiving coils because the magnetic flux is closed symmetrically through their cores. When increasing the length of the transmitting coil to a certain value x_{ON} is increased, too. Instantaneous power value is smaller than the variant with the same coils dimensions and the average power increases up to maximum value. The next increase of the transmitting coil length decreases the area of efficient energy transfer and hence the average power value. The reason for this is that due to the considerable difference in dimensions of both coils the magnetic flux of the transmitting coil is closed in the area above it without reaching the receiving coil. This increases the leakage inductance and significantly reduces the magnet coupling. Additionally, it can be noted that this design has a significant area of the transmitting coil, which is not covered by the receiving coil and the electromagnetic field in this area will attack the nearest metal parts of EV or other equipment. As a result, actual electromagnetic standards are not satisfied [13].

Table 1

Transferred power value at different dynamic charging scenarios

Receiving coil dimensions - 800/700 mm Horizontal distance between each transmitting coils - 800 mm Vertical distance between the transmitting and receiving coils -100 mm Maximum power - 50 kW							
Transmitting coil dimensions, mm	Horizontal misalignment in direction X (Fig. 4), mm	Maximum value of magnetic coupling	x_{ON}, mm	x_{OFF}, mm	P_{MAX}, kW	P, kW average power	Number of transmitting coils for the distance of 100 m
600/700	±150	0.43	300	1100	42.2	4.4	≈71
800/700	±200	0.51	400	1200	50	6.25	≈62
1000/700	±250	0.46	500	1300	45.1	7.34	≈55
1200/700	±300	0.36	600	1400	35.27	8.9	50
1400/700	±250	0.38	500	1700	31.4	4.1	≈45

The last column of Table 1 presents the necessary number of transmitting coils which could be installed in the same length of dynamic charging zone – 100 m. The transmitting coil - 1200/700 mm required about 20% less coils number compared with 800/700 mm variant, which respectively reduce the cost of switching sensors and other communication equipment. Complete economic evaluation and final decision can be performed considering the total production costs of transmitting coils, protective boxes and installation costs.

3.2 Control of Transmitting Power

In dynamic charging with movable EV and respectively its receiving coil, an important development issue is how to realise switching (on/off) of transferred energy. Three possible solutions have been analysed and investigated:

a) after "start charging" all transmitting coils to be supplied with low HF power. When the receiving coil covers the proper transmitting coil, equivalent resistance decreases and the output HF generator current will increase. Monitoring these changes, the HF generator is switched on and specified value of energy is transferred. All other transmitting coils are still supplied by low HF power. This solution is not cost effective because each transmitting coil has to be supplied by separate HF generator;

b) monitoring of HF generator load when the receiving coil covers the proper transmitting coil. The equivalent impedance changes are used to start energy transfer only to this transmitting coil. There is a disadvantage – the impedance could be changed not only by receiving coil, but also by other metal parts from the car or other external metal parts;

c) using sensors for switching-on each transmitting coil, when requirements for correct re-covering with receiving coil are satisfied. Switching-off a transferred energy, is sensorless through measurement of the output HF generator current. When it falls below specified value the generator is switched-off.

Through analysis the third solution has been selected and realised as a more reliable and cost-effective. It is visualized in Fig. 1b) and Fig. 4. It consists of n number of transmitting coils supplied from one ChS and sensors for starting the charge algorithm (points O, C, F, I in Fig. 4). The dynamic charging scenario of EV is possible, if two preliminary conditions are fulfilled. The first one includes some organizational issues, relating to the identification, payment, etc. The second one is technological - correct positioning of the receiving to the transmitting coil in the charging area. Correct positioning means that all specified requirements for misalignment between two coils are satisfied. It is preferable, if the used sensor could have possibility to register correct positioning and only after that to switch-on the proper transmitting coil. Additionally, the selected sensor must operate without influence by the high frequency electromagnetic field between the two coils. By means of experimental studies it was proved that the greatest functional reliability is achieved by a magnetic sensor - type MGT 201 [25], which consists of two parts - active and passive.

The active sensor part is mounted on the front wall of the transmitting coil in the direction of EV movement (direction *X*) - Fig. 4. The distance x_4 between the sensor and this wall is subject of adjustment, and thus determine the moment of its activation, respectively starting of the charge, in accordance with the overlapping level of the transmitting and receiving coils, i.e. the maximum allowable misalignment in the EV movement direction.

The sensor's passive part, Fig. 8, includes two permanent magnets, magnetic core and a protective box and it is mounted to the rear wall of the receiving coil in the EV movement direction. The type of permanent magnets, geometrical dimensions and configuration of the magnetic core ensure accurate distribution of the magnetic field - Fig. 8, which is in accordance with specified correct conditions of horizontal and vertical misalignment between two coils. Only when these conditions have been implemented, magnet sensor switched-on and allows switching-on of the HF generator, in order to transfer electrical energy to the transmitting coil. As the displacement between the coils is larger, the output power is smaller. This means that when EV is moving and the receiving coil passes over the transmitting coil, the charging power at the beginning is minimum, it reaches a maximum at no misalignment and again goes to minimum value - Fig. 4.

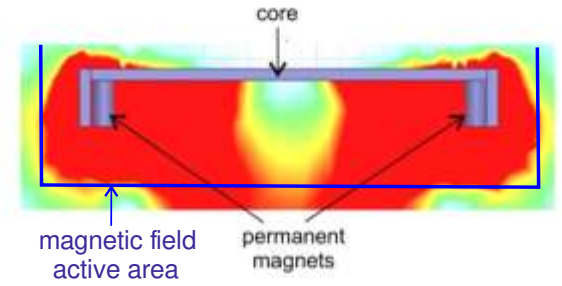


Figure 8
Magnet sensor - passive part

The second minimum (Fig. 4 - points B, E, H, K) is used to stop the charge by the controller, located in the ChS and switch-off the transmitting coil, i.e. switching-off is sensorless. Thus, the ChS is ready to supply the next transmitting coil. During the movement of EV, the same algorithm is repeated for each transmitting coil located in the charge area of dynamic charging infrastructure.

3.3 Dynamic Charging Architecture and Communication Levels

Fig. 9 presented overall-view architecture of developed dynamic charging technology for EV. It contains two main parts - ChS and EV side. All power electronic modules and relevant electrical circuits, for both sides, and EV battery pack are marked in grey. In ChS side, main electronic modules - AC/DC input rectifier and HF inverter and also in EV side- AC/DC module, are specially designed in order to fulfil specified electric parameters for dynamic charging as: maximum transferred electrical energy, operating frequency, efficiency, etc. These parameters are also input data for the design of ClCh module.

To organise functionality of power electronic modules, the necessary control and communication units were developed and also are presented in Fig. 9. In EV side there are several units: vehicle management unit (VMU), which operates as a master, electronic control unit in EV side (ECU EVS) and battery management system (BMS) operates as slave controllers. They are connected through CAN bus.

ECU EVS controller receives data from AC/DC module: output DC current and voltage, AC/DC module temperature and from the mechanical unit (not shown in Fig. 9), positioning down and up the receiving coil. BMS is integrated in a battery pack and has monitoring functions regarding state of charge of EV battery.

In ChS two controllers are available: ECU ChS and charging station management unit (ChS MU). The first one compares data on measured output AC/DC current and voltage with set-up charging current and voltage and defines the proper HF inverter control signals. Information about ChS energy consumption is stored in ChS MU unit.

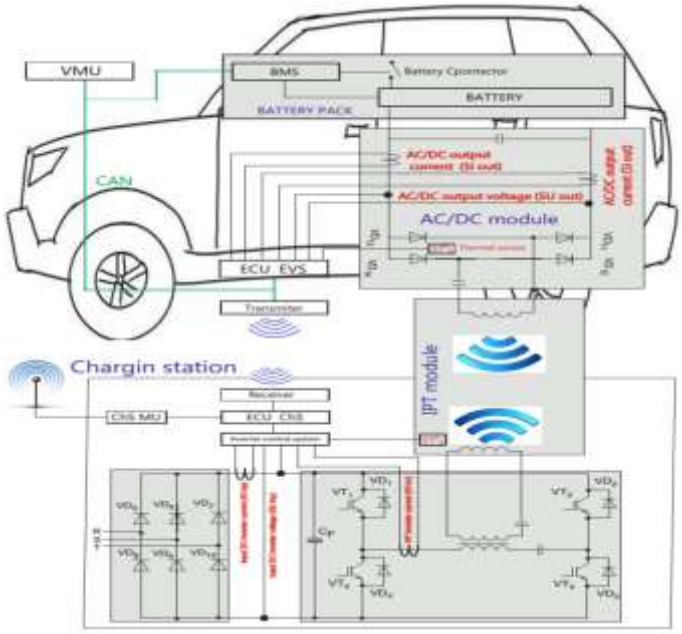


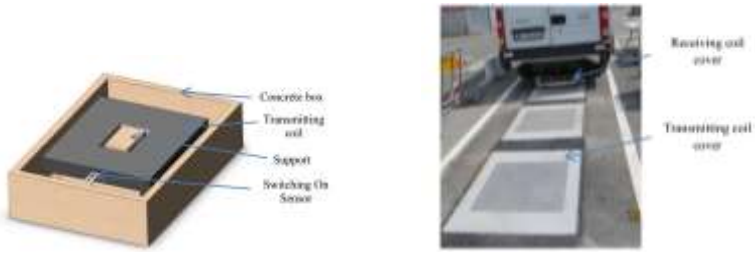
Figure 9

Contactless charging architecture – primary and secondary side

The communications of contactless dynamic charging EV technologies are organized and implemented on two levels. The low level is communication between ChS and EV charging units. It operates with technological and control data as: BMS defined set-up data (charging current and voltage), measured AC/DC output current and voltage, Start and Stop charging signals, etc. All data are transferred through Wi-Fi transmitter in EV side and Wi-Fi receiver in ChS.

This is one-directional communication. The higher level is between ChS and charging user (EV driver) with energy management system (EMS). It is organized and operates on regional level and serves a limited number of stations and users, respectively – Fig. 9.

Not only electrical, but also construction parameters are important. For example, electromagnetic field exposure, which must cover the relevant EMC standards, depends on IPT module shielding [13, 21]. Transmitting coils are part of road infrastructure, because they are built in the ground. Special construction measures are taken to guarantee reliable operation of transmitting coils - protection against external standard limited mechanical loads, against different environment conditions as: moisture, dust, extreme temperature, etc. To fulfil all these requirements, a special design of protective boxes for integration transmitting coils in a road was developed (Fig 10a)). Special attention was paid of the cover design and used materials - limited thickness of 40 mm, considering the specified gap between transmitting and receiving coils, max 100 mm, and to meet current standards for public roads [13], that means to withstand 12 tones external load. The material selected was a polymer concrete reinforced with fiber glasses (non-metal is allowed) – Fig. 10b).



a) b)

Figure 10

Dynamic charging architecture: a)protective concrete polymer box; b)real test of EV dynamic charging

The discussed dynamic architecture has been realized, tested and validated at power P=10-35 kW, frequency f=18-25 kHz, vertical air gap between transmitting and receiving coils -100 mm and horizontal misalignment - up to ±200 mm – Fig. 10 [8].

3.4 Energy Management

The electric power industry expects 400% growth in annual sales of EVs by 2023, which may substantially increase electricity usage and peak demand in high adoption areas. Understanding customer charging patterns can help utilities anticipate future infrastructure changes and to develop intelligence EV energy management system that will be needed to handle large vehicle charging loads, including contactless ChS [24]. Major findings could be grouped into two categories: charging behaviour and grid impacts.

Charging behaviours. The studies found [8] that the vast majority of in-home charging participants charged their vehicles overnight during off-peak periods. Where offered, time-based rates were successful in encouraging greater off-peak charging. Public ChS usage was low, but primarily took place during business hours and thus increased the overlap with typical peak periods. EV owners frequently used the (often free) public stations for short charging sessions to "top off their tanks."

Grid impacts. Length of charging sessions and the power required varied based on the vehicle model, charger type, and state of battery discharge. While the average power demand to charge most vehicles was 3-6 kW, the load from some electric vehicles that are using fast charging can be as much as 20 kW.

As the number of installed inductive chargers in the grid increases, the load profile of the network will be significantly modified, due to the high charging power served from this type of chargers. The additional charging demand may provoke grid issues such as voltage excursions, network overloading, etc. For that reason, an Energy Management System (EMS) is necessary in order to minimize potential disturbances in the normal operation of the grid. Additionally, the developed and used EMS must propose several services to EV drivers [8, 17, 24].

The energy management system fulfils the following objectives:

- monitoring the operation of the ChS - consumption in real-time in order to identify the demand flexibility that can be offered to support network operation;

- to enable the remote control of the maximum charging rate of the stations under emergency network operational conditions;

- user awareness of the location, the availability and the electricity cost of the contactless charging stations - the EMS makes EV drivers aware of the locations of the existing dynamic charging infrastructures in order to be able to decide the most convenient place for charging their EV in respect to their trip destination.

- to offer booking services to EV owners - enabling them to book the most suitable ChS at the most convenient time, considering their trip destination as well as the electricity energy prices.

The EMS architecture is presented in Fig. 9 and it comprises three components: the user awareness module, the monitoring module and the decision module. At any time, EV owner needs to find info from the user awareness module about the exact location of the nearest available ChS. Based on this info EV drivers can reschedule their driving route to the desired destination in order to reach the most convenient and available charging scenario.

The monitoring module is responsible for the interaction between the ChS and the EMS and for remotely controlling the maximum allowable charging rate of all the charging stations in a given area. The actual charging rate is defined by the battery

management system of the EV which cannot be higher than the one defined by the EMS, respectively the current loading of the electrical network.

The decision module is responsible for purchasing energy from the wholesale market and supplies the charging demand of EV drivers. Finally, the decision module offers demand response services to the market operator. In case of network operational issues (voltage excursions or network equipment overloading), the decision module can support the problematic grid area by reducing the charging rate of the charging stations located at that area.

The presented EMS architecture, from charging application point of view, is usable for all EV charging technologies – plug-in, contactless, static and dynamic. Its intelligence is based on the possibility to process the charging/booking requests from EV users as well as the demand response requests for network support in real-time. The EMS is responsible for the information management coming from the ChS, EV users and market operator and enables the interaction and information exchange between them.

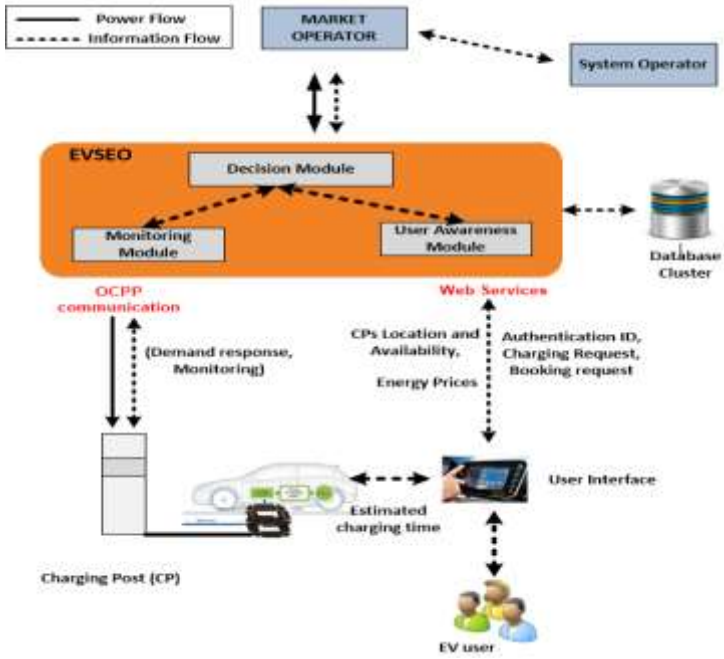


Figure 9

The energy management system–main modules and communications

Conclusions

This paper outlines some technical problems and results concerning design and implementation of dynamic EVs charging. The necessary power electronics and control units, that guarantee correct charging process, are described in the

presented architecture of dynamic ChS. The concept of two hierarchical communication levels between ChS, EV charging components and EM has been defined and the main features of used units have been discussed. The new research results considered are:

- defining the magnet coupling dependences as a function of misalignment and transmitting and receiving coils dimensions;
- dependence of power transfer efficiency Vs. magnet coupling, IPT quality factor and misalignment;
- reliable solution for control of transferred energy, and admissible misalignment, using magnet sensor and proper construction of its passive part;
- original construction of protective box of transmitting coils, built in a road, using polymer concrete reinforced with fibre glass, enabling to minimise the thickness of the cover.

All achieved results have been taken into account during design and practical implementation of dynamic road infrastructure. They are validated through dynamic inductive EV charging system with power of 35 kW (P_{MAX} = 50 kW) and efficiency of transferred energy (grid to EV) close to 90-92% at a distance between transmitting and receiving coil of 100 mm.

The proposed theoretical analysis, design considerations and practical work were done in the frame of EU FP7 project FastInCharge [8].

References

[1] Jin H. and Rim C. T., "KAIST wireless electric vehicles - OLEV," SAE International, Vol. 1, pp. 1-10, 2011

[2] Wu H. H., Gilchrist A., Sealy K., Israelsen P. and Muhs J., "A review on inductive charging for electric vehicles," in Electric Machines & Drives Conference (IEMDC), 2011 IEEE International, 2011, pp. 143-147

[3] Miller J. M., Scudiere M. B., McKeever J. W., and White C., "Wireless power transfer," in Oak Ridge National Laboratory's Power Electronics Symposium, 2011

[4] Covic G., Boys J. T., "Modern Trends in Inductive Power Transfer for Transportation Applications", IEEE Selected Topics in Power Electronics, Vol. 1, No. 1, March 2013

[5] Garnier L., Chatroux D., " Understanding the unbalancing of a battery pack to choose the best balancing solution", PCIM 13, 14-16 May 2013, Nurnberg, Germany, ISBN 978-3-8007-3505-1

[6] Chopra S., Bauer P.,"Analysis and design considerations for a contactless power transfer system", Telecommunications Energy Conference (INTELEC), 2011 IEEE, Oct. 2011

[7] Kamath H., "Program on Technology Innovation: Impact of Wireless Power Transfer Technology", Electric Power Research Institute (EPRI), California, 2009

[8] "Innovative fast inductive charging solution for electric vehicle", part of 7th Framework Program of EU, www.fastincharge.eu

[9] "On-road charging of electric vehicles"- The fabric project, part of 7th Framework Program of EU, http://www.fabric-project.eu/

[10] Covic G., Boys John T., Budhia M., Huang C., "Electric Vehicles – Personal transportation for the future", World Electric Vehicle Journal Vol. 4, Shenzhen, China, Nov 5-9, 2010

[11] OLEV Technologies /On Line Electric Vehicles/, http://olevtech.com

[12] Conductix Wampfler, www.ipt-technology.com

[13] ICNIRP – International Commission on Non – Ionizing Radiation Protection, EMF Guidelines 1998, 1999, 2001, 2008, 2009, 2010, 2014, www.icnirp.org/en/publication

[14] Bombardier, http://primove.bombardier.com

[15] UNPLUGGED project, http://unplugged-project.eu/

[16] Madzharov N., Tonchev A., "IPT Station for static and dynamic charging of Electric Vehicles", International Scientific Conference PCIM 2014, Nuremberg, Germany

[17] Madzharov N., Tonchev A., "Inductive high power transfer technologies for Electric Vehicles", Journal of Electrical Engineering, Vol. 65, No. 2, 2014, 125-128, ISSN 1335-3632

[18] Madzharov N. D., Ilarionov R., Tonchev A., "Systems for dynamic Inductive Power Transfer", Indian Journal of Applies Research, Vol.: 4, Issue: 7, July 2014, ISSN-2249-555X, IF: 2.1652

[19] Donati A., N. Madzharov, A. Melandri, F. Sighinolfi Induction sealing device for producing portable food packages –, US patent - US 8,286,406,B 2 - Otc 16, 2012

[20] Li S., Mi C., Wireless Power Transfer for Electric Vehicle Applications IEEE journal of emerging and selected topics in power electronicS, Vol. 3, No. 1, March 2015, pp. 4-17

[21] Dichev D., Koev H., Bakalova T., Louda P. A Measuring Method for Gyro-Free Determination of the Parameters of Moving Objects. Metrology and Measurement Systems, 23 (1), 2016, 107-118, ISSN 0860-8229

[22] Kraev G., N. Hinov, D. Arnaudov, N. Ranguelov and N. Gradinarov, „Multiphase DC-DC Converter with Improved Characteristics for Charging Super capacitors and Capacitors with Large Capacitance", Annual Journal of Electronics, V6,B1,TU of Sofia, Faculty of EET, ISSN 1314-0078, pp. 128-131, 2012

[23] Bankov N., Vuchev Al., Terziyski G., Operating modes of a series-parallel resonant DC/DC converter. – Annual Journal of Electronics, Sofia, 2009, Volume 3, Number 2, ISSN 1313-1842, pp. 129-132

[24] Karfopoulos E., Hatzoplaki E., Safalidis G., Karakitsios I., Kamarinopoulos A. and Hatziargyriou N., Energy Management System for fast inductive charging network: The FastInCharge project, MedPower Conference 2014, IET, 2014, Athens, Greece

[25] http://www.ifm.com

Modeling and Solving an Extended Parallel Resource Scheduling Problem in the Automotive Industry

Mónika Kulcsár-Forrai, Gyula Kulcsár

Department of Information Engineering, Faculty of Mechanical Engineering and Informatics, University of Miskolc, 3515, Miskolc-Egyetemváros, Hungary, aitkfm@uni-miskolc.hu, iitkgy@uni-miskolc.hu

Abstract: This paper presents an extended model for solving time-varying resource-constrained scheduling problems. The motivation for our research comes from the automotive industry. The problem is to create fine schedules for a complex manufacturing system to satisfy diverse customer demands. The detailed characteristics of the analyzed scheduling problem and the developed solving approach are described in this paper. To consider the impact of the assistant processes that are connected to the manufacturing primary processes, we elaborated a problem-transformation procedure and a new extended scheduling model that can manage time-varying availability constraints of parallel resources, unit processing times, job-dependent release times and due dates. This paper also presents slack-oriented and JIT-oriented algorithms that can solve the resource-constrained scheduling problems. The research results have been successfully applied and tested in practice.

Keywords: scheduling; resource availability constraint; multi-objective optimization; production planning and control; manufacturing operations management

1 Introduction

Production planning and scheduling systems deal with the allocation of limited resources to production activities to satisfy customer demands over the actual time horizon. Planning and scheduling tasks can be expressed as optimization problems, in which the main goal is to create plans that meet constraints and maximize production performance. These optimization models are very different in practice, corresponding to the characteristics of the real production systems and their business environments.

A hierarchical approach is one of the possible ways for solving such planning and scheduling problems. The hierarchical optimization approach means that the

decision process will run in a layered way by ordering the decisions according to their relative importance. This decomposition technique uses a suitable optimization model at each level of the hierarchical decision-structure. At a given level the applied model extends the decision variables, the constraints and the objective functions of the problem arriving from the upper level.

Production planning and scheduling process typically works according to the rolling horizon principle. This means that an initial plan is created for the actual time horizon, and its first part is executed. The system creates the plan for the next period by considering the previous partially overlapped period in advance. Then, the initial plan may be modified or re-planned by considering the changes and disturbances. The actual plan can also be periodically revised due to the uncertainties that may occur in the business and in the production processes.

Results of production planning and scheduling are usually not applicable to managing operational manufacturing since the created plans are rough and large-scale solutions, and they refer to aggregate resources. The role of fine scheduling (detailed scheduling) is to make a precise executive fine program for a short time horizon (for weeks/days/shifts) that concerns every detail.

To realize the created production fine schedule in practice, the complex decision making has to cover the primary processes and also the most important supplementary (e.g. logistical) and assistant processes (e.g. instrument supply) of the production.

In this paper, we present fine scheduling models and algorithms for solving real-life problems in the automotive industry. We focus on modeling and solving the scheduling problems of vehicle seat element manufacturing. To solve the fine scheduling problem, not only the main manufacturing processes, but the configuration-preparation processes have to be considered.

This paper is organised as follows: Section 2 briefly describes the examined production system. Section 3 reviews fundamental models for production scheduling, while Section 4 introduces a new solution for fine scheduling. Section 5 proposes an extended parallel resource scheduling model and new solving algorithms, and Section 6 shows an application of the theoretical results in practice. Finally, conclusions are given in the last section.

2 The Examined Production System

In the examined vehicle manufacturer workshop, seat elements are made for different types of cars. The customers (vehicle-assembly enterprises) generate product demands (orders) for specified product types and numbers of items. These production orders have to be fulfilled within strictly prescribed time limits.

The plant manufactures the seat elements (end products) on circle-shaped manufacturing systems. The manufacturing systems of the plant complete the orders together. It is possible to produce the specified product type on more than one path. Every single path can perform a given amount of rounds (cycles) in one shift, furthermore it possesses a specified number of attachment points (positions). There are shape carriers in the system. A given type of shape carrier can be connected to a given position of a specified path. The shape carrier can be one or two sided depending on its formation. It is possible to attach tools (molds) to the left and/or right side of the shape carrier. The attachments are determined by technological rules. Strict rules prescribe:

- what kind of products,
- on which path,
- in what kind of position,
- on what type of shape carrier,
- on which side and
- with what kind of other products they can be manufactured.

A path can be considered as a production line that works according to an independently defined calendar. The basic unit of the calendar is the shift. These basic working time intervals are equal (e.g. eight hours) long. In every shift the production lines can be adjusted. This means that it is possible to exchange a given number of shape carriers which are connected to the positions. The unit of one exchange is a given configuration that consists of one carrier and its connected mold(s). A complete exchange happens when we take the current configuration from the position and attach another prepared configuration to the same position.

The constructions of the production lines (paths) are different, so the total number of executable rounds in one shift can differ. In the system, various numbers of molds and carriers are usually available for manufacturing products.

The preparation (assembly and disassembly) of the tool configuration, which is needed for the production, is carried out by skilled workers. The preparation task is time consuming. Therefore, the number of the performable configuration preparations during one shift in the plant is restricted by tight capacity constraints.

3 Fundamental Models for Production Scheduling

Discrete manufacturing processes include a large variety of diverse and very distinct technologies that require specific models when creating or improving their efficient control systems. In this regard, there is a special demand for adequate modeling, formulation and solving of scheduling problems [20].

In the professional literature, many books and papers deal with scheduling models and methods. There are well-structured books that focus on manufacturing scheduling (e.g. [16], [17], and [21]). Several review and survey papers can also be found on this topic, for example [1], [5], and [11]. Scheduling plays an important role not only in manufacturing but in many different service industries. Pinedo et al. [18] presented an overview of some of the more important scheduling problems that appear in various service industries.

The scheduling problems lead to optimization tasks. Their complete review and classification exceed the scope of this paper; there is a vast amount of literature that deals with such problems. Therefore, only some of the most important solving approaches are shortly mentioned. The main categories are as follows:

- Mathematical programming approach (e.g. linear, nonlinear, integer programming, disjunctive programming, set assigning, set partitioning, set packing and set covering, etc.).
- Exact optimization approach (e.g. branch and bound methods, dynamic programming, etc.).
- Constraint programming approach (e.g. constraint satisfaction and constraint programming).
- Heuristic approach (e.g. basic scheduling rules and composite dispatching rules, etc.).
- Iterative improvement approach (e.g. beam search, local search and genetic algorithms, etc.).

To solve a scheduling problem in practice, we have to deal with at least three important issues. The first issue is the resource environment. This takes into consideration all features of the resources concerned (machines, workplaces, workers, tools, etc.) and characteristics of the relations among them. We also have to pay special attention to the features of the operations to be executed. The second issue focuses on the job characteristics and constraints. This group includes all the technological rules, manufacturing restrictions, job execution features and alternative process plans. The third issue is the production control policy that specifies the priorities, requirements, objective functions and key performance indices. The possible variants of these groups of issues result in many scheduling problems.

The simplest resource environment of scheduling problems is represented by the single machine type model, which refers to jobs containing one operation to be performed. If the single resource is replaced by a given set or group of resources, the scheduling problems become models of parallel resources. In this case, the jobs are executed simultaneously on different resources (machines or workplaces). Depending on the working resources capability of the parallel models, three basic resource environments are categorized [16], [21]. The first is the model of identical parallel machines, where each job can be processed on any machine, and

the processing time only depends on the dedicated job. The second variant is the model of uniform parallel machines, where the processing time of a dedicated job varies according to the machine speed. The third variant is the model of unrelated parallel machines, where the processing time varies according to the job and the machine. In this case, each machine can work at different speeds on the jobs. Several papers give detailed reviews on parallel machine scheduling, for example [4], [7], and [22]. However, Weng et al. [19] and later Lamothe et al. [13] have shown that the scheduling problems on machines with limited flexibility, setup and secondary resource constraints were poorly studied.

The shop scheduling models involve more machines and more jobs containing more than one operation. However, each operation can be performed on a given machine. According to different prescriptions, specific models can be formed from this general shop, as follows [21]:

- Job shop: the set of operations is job dependent, and each job may have a special precedence chain relation (operation sequence).
- Flow shop: a special case of the job shop in which the number and sequence of operations is fixed for any job.
- Open shop: a special case of the general shop in which there are no precedence relations between the operations.
- Mixed shop: a combination of the above models.

In the above-enumerated models, each operation can be assigned to a given machine (dedicated resource). A further extension of the general shop model is the flexible shop model, where the flexibility feature refers to the machine assignment possibilities. In the flexible models (e.g. flexible flow shop and flexible job shop, etc.) a given operation can be performed on any of the machines of a specified machine group. In this way, the scheduling problem is supplemented with machine selection tasks. The suitable machines from the group can simultaneously work identically or uniformly or even in an unrelated way [6], [23]. These models can be considered as the combined models of the shop models and parallel machine models.

The extended flexible shop models represent a new generation of the scheduling problem class (e.g. extended flexible flow shop and extended flexible job shop, etc.). It often occurs in the manufacturing systems that there are some resource objects (e.g. integrated production lines and cells) that can perform more than one operation as a unit. In this case, some operations can be grouped into larger units such as technological steps or even execution steps. These collecting steps can be considered as basic units for scheduling [8], [9], and [15]. The extended flexible shop models support the usage of alternative technological routings and resources. Parallel machine models are the functional building blocks of these extended models to support the parallel realization of the execution steps.

For applying scheduling models in practice, it is very important to consider the machine eligibility constraints. For example, Lin and Li [24] studied the parallel machine scheduling problem with unit-length jobs in which each job is only allowed to be processed on a specified subset of machines. Many other researchers also have considered the machine eligibility constraints in parallel machine scheduling problems. Lee et al. [14] studied the most general case of machine eligibility constraints as well as special cases of nested and inclusive eligible sets.

From the practice-oriented point of view, flexible scheduling approaches have to pay special attention to the job execution constraints related to the release times and due dates. As is well known, the general case of this problem with only a single machine is NP-hard if the optimization objective is to minimize the maximum lateness. The NP-hard (non-deterministic polynomial hard) property indicates that there is probably no polynomial-time algorithm to reach the optimal solution of the problem. (The precise definition of the NP-hard problem is given in [21].) Consequently, the parallel and extended variants of the problems are also NP-hard. Special cases of the problem have been analyzed in recent years. For example, Lazarev et al. [2] considered the problem with only a single machine and identical processing times for all jobs.

In most of the scheduling problems, it is assumed that the resources (machines) are continually available in time, and the number of resources is fixed in each problem instance. These simple types of availability constraints can only limit the number of parallel executable tasks. One other type of availability constraint is connected to calendar elements or shifts. The resource availability constraints specify the time intervals or windows in which the resources can perform the jobs. Ma et al. [25] provided a detailed review of this topic. Kaabi and Harrath [12] reviewed the models and results related to the parallel machine scheduling problem under availability constraints. Nevertheless, it can be seen that parallel resource scheduling problems with due-date related objective functions, resource-availability constraints, distinct release times and due dates are poorly studied in the literature.

Our problem has similarities to the problem addressed by Gharbi and Haouari [3] in the sense that both models have parallel resources, distinct release times and due dates. However, in our model, each resource has its own list of availability time intervals, not only a single availability time window, and the objective functions are also different. Brucker [21] presented the model $P \mid p_i=1;\ r_i\ integer \mid L_{max}$. This is similar to our scheduling model for configuration preparation in that Brucker's model also uses unit processing time, integer release times and due dates, but we even take into account the distinct availability time constraints of each machine.

In this paper, we study a special parallel resource scheduling problem in which time-varying resource constraints, unit processing times, job-dependent release times and due dates are considered to minimize the maximal tardiness and

earliness. In the literature we have not found any research papers considering this set of constraints and objectives.

4 A New Solution for Production Fine Scheduling

The product-type dependent production intensity of the examined manufacturing system can only be modified slowly because of the defined constraints. Therefore, the service of diverse production orders generates a serious production fine scheduling problem. Based on the necessary information the fine schedule is usually created for a one-week time interval in advance. It is recommended to keep product dependent stock levels adjusted individually due to the long reaction time of the manufacturing system.

During the development, we focused on heuristic and knowledge-intensive searching techniques because the addressed full scheduling problem is NP-hard. At the beginning of the development we started out from our previous models applied successfully in other situations [10].

In our approach, all the issues (batching, assigning, sequencing and timing) are handled simultaneously (Figure 1). The values of the decision variables of the problem are set by a multi-operator and multi-objective searching algorithm. The developed scheduling software modifies the actual schedule iteratively and prepares new solutions with consistent changes (modifications) based on multiple neighboring operators, execution-driven fast simulation, and overloaded relational operators.

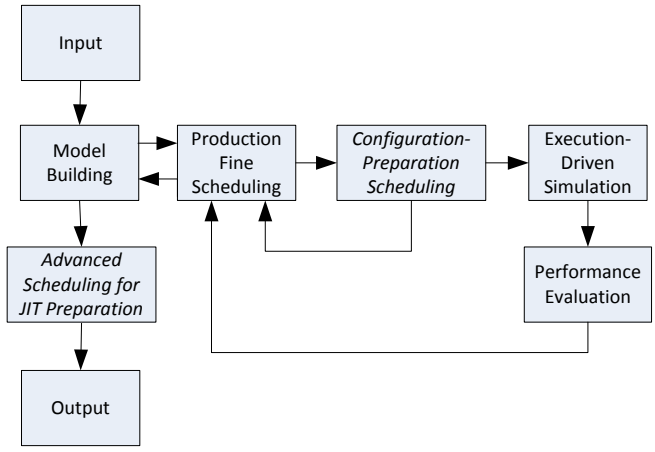


Figure 1
Simplified flow chart of the integrated scheduling approach

Based on the given input data, the model builder component defines the required model objects and initializes them with start values (attributes). Its tasks include the creation of the starting status of resources (paths, positions, carriers, molds), the specification of the internal production orders, and furthermore the definition of the restrictions and objective functions.

After carrying out the availability, applicability, and feasibility studies on the created object model, the builder process defines the currently indexed relationships of the entire system. Using these indices in each decision-making situation (e.g. assignment, selection) of the solving process, the alternatives of choices can easily be retrieved, so the chain of decision making produces feasible solutions.

The core of the implemented solver explores iteratively the space of the feasible solutions and creates neighbor candidate solutions by modifying the decision variables of the fine schedule according to the problem space characteristics. For each shift, the fine schedule specifies the configuration of shape carriers and molds to be run and the kinds of products to be manufactured in each position of each production line.

The candidate schedules are simulated by using an execution-driven fast simulation algorithm that represents the real-world environment with capacity and technological constraints. In this execution-driven simulation, the product units are passive, and they are processed, moved and stored by active system resources such as production lines, material handling devices, and buffers. The numerical tracking of the entities (product units, shape carriers and tools) provides detailed data of the manufacturing.

By using the results of the simulation, the actual values of the key performance indicators (KPIs) can be calculated. To express the shop floor management's goals, we use the following objective functions to be minimalized in a multi-objective optimization problem:

- The maximum product shortage at the due dates of the production orders;
- The sum of product shortages at the due dates of the production orders;
- The number of tardy production orders;
- The number of set-up activities;
- The maximum number of set-up activities in one shift;
- The number of product types with surplus;
- The sum of product surpluses;
- The maximum product shortage at the end of the time horizon;
- The sum of product shortages at the end of the time horizon;
- The sum of the priorities of tardy production orders;

- The maximum priority of tardy production orders;
- The number of product types with tardiness;
- The maximum product shortage (compared to zero);
- The sum of product shortages (compared to zero);
- The maximum tardiness of production orders;
- The sum of the tardiness of production orders;
- The number of configuration preparations;
- The sum of unused capacity of the production lines.

Every objective function has dynamically changing importance and sets of values to be taken into consideration. The relative quantification of the currently examined solution can be performed by comparing it to the best solution found so far. The mathematical model of this qualification was described in [8].

In every iteration of the searching algorithm, the actual solution has to be examined to decide whether the candidate fine schedule is feasible from the point of view of configuration preparation. Each candidate fine schedule requires well-defined preparatory activities for the tool configurations to be used. These tasks have to be scheduled on time-varying capacity-constrained parallel resources (skilled workers) to achieve zero tardiness. If this is possible, then the production fine schedule is feasible, otherwise it cannot be executed. Successful adaptation of this approach to practice is highly influenced by the efficiency of the solving algorithm of this sub-problem.

When the final solution of the production fine scheduling is being prepared, it is very important to pay special attention to the robustness of the configuration preparations. In the searching iterations, the algorithm focuses on the minimization of the maximal lateness. In the end, the final solution can be made more sophisticated by using an advanced JIT-oriented scheduling algorithm that adjusts the release time of the preparatory tasks to be as close as possible to the due dates, considering the necessary preparation times and the prescribed safety time intervals.

The complexity of the concrete industrial scheduling problem is described in Section 2. The full (complete) production scheduling problem is handled by using an advanced multi-objective and multi-operator searching algorithm. In each iteration of the searching algorithm, the built-in scheduling sub-problem of the configuration preparations has to be solved. These two scheduling problems constitute a two-level decision hierarchy, in which each level uses its own specific optimization model. The solution of the problem created at the higher level gives input data, constraints and criteria to the lower level. The given set of constraints is extended with the new constraints of the built-in problem at the lower level of the hierarchy.

The built-in sub-problem is a special scheduling problem. The "job" means one preparatory task at the lower level. We have to schedule the preparatory tasks (assembly the configurations of part-adequate tools and shape carriers) required by manufacturing primary (main) processes. The current production schedule (the solution of the overall problem) generates dynamically the jobs to be scheduled and their release times and due dates for the built-in sub-problem (a given set of configurations is used on the production lines and then a given subset has to be re-assembled by due dates). In the following parts of the paper, we focus on modeling and solving this sub-problem.

5 Modeling and Solving the Extended Parallel Resource Scheduling Problem of Configuration Preparations

5.1 The Description of the Problem

The scheduling problem of the configuration-preparations (jobs) can be summarized as follows:

- There are n jobs J_i ($i=1, 2, \ldots, n$). They are independent of each other and bound by the earliest starting and the latest completion times.
- The availability time intervals of the resources are defined by a calendar. This calendar consists of lists. Each resource has its own list, which is built up from time intervals. Each interval means a shift. The shifts do not overlap, and they are sorted by the starting times.
- The sets of available resources are classified by the shifts. Each set consists of uniform parallel resources. The resources mean skilled workers, who are able to prepare a prescribed number of configurations separately in shifts.
- The shift calendar is an input data structure and the maximal number of the executable preparatory tasks (jobs) can differ according to the calendar.
- The goal is to create a schedule to minimize the maximal tardiness of the configuration preparations by keeping the restrictions.

5.2 Problem Transformation

The problem described in the previous section is difficult to solve in its original form. Many papers can be found in the literature on parallel machine scheduling, but there is no suitable model for the examined time-varying resource availability features of the current problem. Therefore, we elaborated a problem-transformation procedure. Using this procedure we transform the problem to an advanced parallel machine scheduling problem, in which the actual number of the available machines depends on the time.

The essence of the transformation is the following:

- We give serial numbers in the form of decimal integers to the shifts in the global system joint to the plant. These serial numbers are called slots (s). The slots create a connected series that replaces the time axis. The last slot is denoted by s_{max}.
- The processing time of the configuration preparation cannot be longer than one shift. This fact comes from the applied technology. Each processing time takes one (unit value) slot ($p_i = 1$).
- The time data of the jobs are also transformed to slots: the earliest starting time is converted into the serial number of the next shift (r_i), the due date is converted into the serial number of the target shift (d_i), and we search for the completion time in a form that expresses the serial number of the assigned shift (C_i).
- The lateness of the job is also measured in slots: $L_i = C_i - d_i$. The tardiness is also calculated in slots: $T_i = max\ (0.\ L_i)$.
- The group of assembly workers changes with the set of parallel virtual machines (resources). Each virtual machine can work on one job at one time, and each job can only be executed on one virtual machine at a particular time. The number of the available virtual machines can differ according to the slot. This time-variable number of machines is denoted by $P(s)$. The original limitation in the number of executable jobs in the shift gives the concrete number of the available virtual machines in the specific slot.

The $P(s)$ values (limits) define the time-varying resource availability constraints. For example, the limit is 3 in the second shift, and 5 in the seventh shift. In the transformed model, these constraints define the number $P(s)$ of virtual machines (special skilled workers) in slot s, so each job has a unit processing time (equal to the shift length) without loss of generality. For example, one worker team performs three jobs, or there are three workers and each worker performs separately one job.

Based on the formal description $\alpha \mid \beta \mid \gamma$ commonly used in the literature (e.g. [21]) and the symbols introduced above, the new scheduling problem can be formalized as follows:

$$P(s) \mid p_i = 1;\ r_i = integer;\ d_i = integer \mid L_{max} \tag{1}$$

5.3 Slack-oriented Solving Algorithm

For solving the transformed scheduling problem of the configuration preparations (1), we developed a relatively simple algorithm that gives an optimal solution (Figure 2).

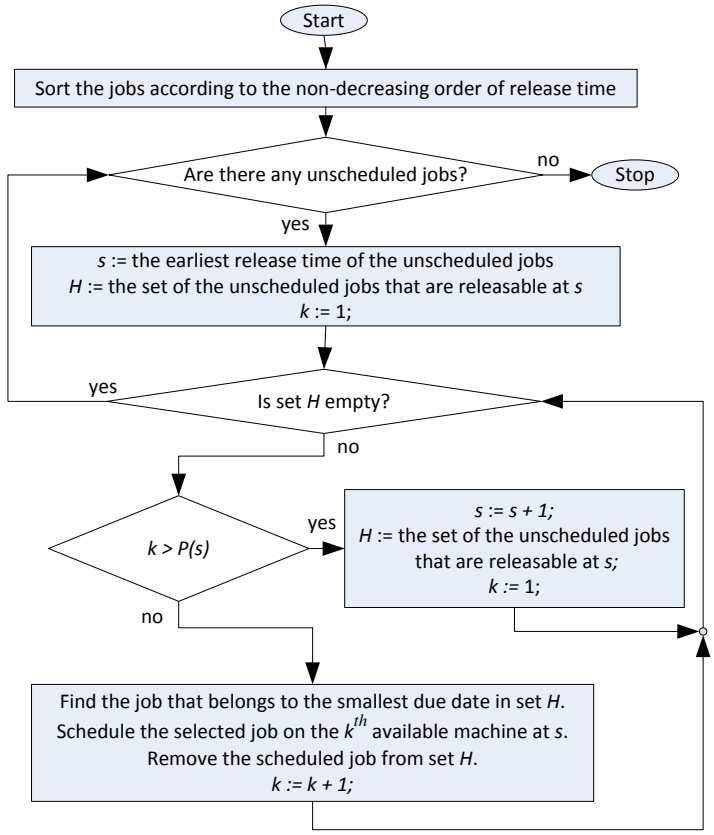


Figure 2
Simplified flow chart of the slack-oriented solving algorithm

The key element of the solving algorithm is that we store the actual number *P(s)* of the available parallel machines in each slot and schedule the releasable jobs according to the well-known SDS (Smallest Dynamic Slack first) rule. This rule

selects the releasable job that has the smallest dynamic slack time. The slack of a given job is equal to the due date minus the sum of the actual time and the remaining processing time: $slack_i = max(d_i - p_i - t, 0)$, where t denotes the actual time of the decision. As the processing time p_i of all jobs is one unit time, in each situation the job with the smallest dynamic slack can be achieved by using the EDD (Earliest Due Date first) rule. This means that if we can find a free machine at an intermediate slot s and there are at least one unscheduled and releasable job, then we schedule the job J_i with the nearest due date d_i.

The proposed algorithm creates a solution that minimize the maximum lateness. It produces an optimal solution in polynomial running time. If an ordered input data structure is used for jobs ($r_1 \leq r_2 \leq ... \leq r_n$), the algorithm runs in $O(n \log n)$ time.

If the maximum lateness (L_{max}) is not greater than zero, then the configuration-preparation schedule is able to fully serve the execution requirements of the examined production fine schedule. The algorithm calculates the value C_i and thereby creates the required solution (schedule) at the same time. The job J_i (configuration preparation) must be performed in the shift of slot C_i.

5.4 Optimality of the Slack-Oriented Algorithm

To prove the optimality of the presented algorithm, we show that all optimal solutions can be transformed into the result of the slack-oriented algorithm by retaining the optimal value of the objective function.

Each job has a serial number (index i) in order of release times ($r_1 \leq r_2 \leq ... \leq r_n$). Let S_a be the schedule created by the slack-oriented algorithm. Let S_b be an optimal schedule. Two vectors store the completion time of the jobs according to the two solutions (S_a and S_b). Let index x be the smallest job index where the C_x value (completion time) is different in S_a and S_b. This means that the first $x-1$ jobs are processed in the same slot according to S_a and S_b. We suppose that the value of index x is maximal because our assumption covers the relationship between these two solutions.

Therefore, the job J_x is carried out in slot C_x according to S_a while the same job J_x is processed in a later slot according to S_b. In this situation, two possible cases may be distinguished.

In the first case, a given machine is free in slot C_x according to S_b. Therefore, the job J_x can be moved to slot C_x on the free machine. This modification does not spoil the maximal lateness because the target job starts earlier, so the schedule S_b remains optimal.

In the second case, there is no free machine in slot C_x according to S_b. Therefore, there exists a job J_y that is carried out in slot C_x according to S_b, but the same job J_y is carried out in a later slot according to S_a. The due date d_y of job J_y is greater

than or equal to the due date d_x of job J_x. This follows from the fact that the slack-oriented algorithm scheduled the job J_x in an earlier slot than the job J_y according to S_a. As the schedule S_b is optimal, we can interchange J_x and J_y in schedule S_b, and the value of the objective function does not increase.

In both possible cases, the modified schedule S_b remains optimal, and the value x has increased after modification. This result contradicts the assumption that value x is maximum, so the modification can continue until the index x is greater than the number of jobs. Finally, each job J_i is completed in the same slot C_i according to schedule S_a and S_b. This proves that the slack-oriented algorithm creates an optimal solution.

5.5 Advanced Scheduling for Just-In-Time Preparation

In practice, one very important expectation of production management is that the manufacturing control system can fulfill the execution conditions of the released production fine schedule according to the paradigm "Just-In-Time". For example, the needed tools should not be prepared too early, because if some kind of uncertainty or unexpected events occur between the preparation completion time and the actual starting time of the inducing operation, then the released production fine schedule has to be modified, so some preparatory activities will have been done unnecessarily. Consequently, it is appropriate to schedule the implementation of each preparatory task as close as possible to the starting time of the related operation by taking into account the necessary preparation time.

In scheduling the configuration-preparation tasks, the presented slack-oriented algorithm is focused on the minimization of the maximum lateness. If there is a solution in which the maximum lateness is not greater than zero, then we can further refine the schedule so that the JIT principle can be validated. For this purpose, we developed an advanced algorithm that is able to reduce the maximal earliness without violating the due dates. The earliness of the job J_i is measured in slots: $E_i = max\ (0.\ -L_i)$.

The essence of the JIT-oriented algorithm is the following:

1. Start from the schedule S_a generated by the slack-oriented algorithm. Create an LDD (Latest Due Date) list by sorting the jobs in the non-increasing order of due dates.
2. Select the first job from the LDD list.
3. Examine the loading of the slots. In order to find a free machine in a suitable slot, start from the due date of the selected job and go backward in time (slot by slot) until a free machine is found or the original slot of the selected job is reached.

4. If a free machine can be found in a later slot than the original slot, then break the searching loop and move the selected job into the first-found free machine of the latest suitable slot.
5. If there is no free machine in the suitable slots, then the selected job remains in its original place.
6. Delete the selected job from the LDD list. If the LDD list is not empty, then go to 2, otherwise stop.

This JIT-oriented algorithm converts the schedule S_a into a new feasible schedule S_{JIT}, where the maximal tardiness is not increased, and the maximal earliness is reduced. It is easy to see that if a selected job has been moved from its original slot to a later slot, then its earliness is reduced, but the modification does not violate the due date or release time restrictions because of the well-defined boundaries of the searching loop. The modifications are carried out on the jobs in the non-increasing order of due dates. Therefore, jobs can be moved into the places which are freed by one of the previous job movements so the algorithm can achieve the maximum improvement. In the worst case, each job remains in its original place, and thus the tardiness and earliness are not changed.

The presented JIT-oriented algorithm creates a very sharp schedule in each situation. To increase the flexibility of the solution, we extended the algorithm with a set of job-dependent control parameters that specify the safety slack of each job. In this case, the free machine searching loop of the algorithm starts at an earlier (time) slot than the due date. For each job, the initial value of the first examined slot is equal to the difference of the given due date and the given safety slack control value of the job. This technique is well suited to the sophisticated safety requirements of the manufacturing control and to creating not only effective but also robust schedules.

6 Transfering the Theoretical Results into Practice

The scheduling problems outlined in this paper are inspired by a real case study concerning the plant of Fehrer Hungaria Járműipari Kft. specialized in vehicle seat products (Mór, Hungary). The firm produces different types of seat elements with variable series simultaneously. It is typical that the customers set very strict delivery due dates.

Many enterprise resources planning (ERP) and advanced production scheduling (APS) systems can be found in the market. Their functionalities cover a very wide range of different production systems and business environments. However, these general solutions cannot be applied directly to the operational production management in the plant under consideration because the created plans are based

on aggregated resources and they consider only the primary manufacturing processes.

We developed a new fine scheduling software based on the presented approach, models and algorithms. Our software can automatically create short-term execution plans that cover every important detail and process of the analysed production system. The generated schedules specify the tasks that the manufacturing system should perform in the planned time horizon. These detailed solutions can be executable directly at the shop floor level.

The application has useful graphical user interfaces for supporting user interactions in order to increase the flexibility of the production fine scheduling and control process. For example, when the process engineers want to declare mandatory configurations to be used for test manufacturing in a given time interval, an advanced editor module is available for them to express their exact requirements. The automatic scheduler takes into account these constraints. For this purpose, the solver engine is equipped with blocking techniques that make it possible to manage the modifiable and the protected configuration exchanges. The software offers many formats to show the solutions. The most important results are the fine schedule of the production processes, the schedule of the configuration preparations, and the values of the performance indicators. The fine schedule can be displayed as a list of the configuration exchanges with the corresponding data that specify which carriers and tools have to be attached to which position in which shift and what kind of product types have to be produced. The schedule of the configuration preparations declares exactly what pre-assembly activity has to be carried out in which shift.

The developed production fine scheduling system integrates many new functional components. The most important functions are as follows:

- Multi-objective production fine scheduling;
- Scheduling the configuration preparations;
- Managing the time-varying availability-constrained resources;
- Managing the shared accessible resources;
- Managing the process engineers' requirements;
- Managing the product-type dependent stock levels.

The multi-objective approach and these advanced functionalities of the fine scheduling software effectively help satisfy the requirements of shop floor management in practice.

For testing and evaluating the proposed models and algorithms, we used real industrial problems. Practical experience confirms that the production orders can be fulfilled with minimal tardiness and the manufacturing processes can be realized with a minimal number of configuration preparations and exchanges. The

accumulation of excessive stock can be avoided and safe levels of product-type dependent stock can be maintained. The utilization of the production lines can be increased. The importance of these goals can vary over time, so our software supports the user in expressing the actual importance of the objective functions by adjusting the priorities.

Conclusions

In this paper, we summarized our research results on the practice-oriented modeling and solving of fine scheduling problems related to vehicle seat element manufacturing. Extended models and advanced scheduling algorithms were presented to adapt to the concrete requirements of real-life situations by taking into consideration the specific characteristics of modern manufacturing.

We introduced the full problem and the proposed solving approach (fine scheduling); in addition the main part of the paper presented the concrete model and the solving algorithms of a built-in sub-problem, which in itself is a meaningful and important scheduling problem. This sub-problem focuses on scheduling only the preparatory activities (jobs) required by the manufacturing primary (main) processes.

The full production scheduling problem is NP-hard, so we handle this problem by using an advanced multi-objective and multi-operator searching algorithm to create near-optimal solutions. In each iteration of the searching algorithm, the built-in scheduling sub-problem has to be solved in order to decide whether the current full production schedule is feasible from the point of view of the configuration preparation. The new model $P(s)| p_i=1; ri=integer; d_i=integer | L_{max}$ and the proposed algorithms are intended to solve only the built-in sub-problem. This model includes a special resource environment that consists of time-dependent sets of parallel machines, while a set of independent jobs with release time constraints, due dates, and unit processing times is considered. To minimize the maximum lateness, we proposed a new slack-oriented solving algorithm that produces the optimal solution in polynomial running time. For supporting the Just-In-Time paradigm in manufacturing control, we introduced a JIT-oriented version of the scheduling algorithm that is able to reduce the maximal earliness without violating the due dates. To increase the robustness of the released schedule, safety slack control parameters can be used in the scheduling process.

The application of the proposed approach in practice showed that multi-operator and multi-objective fine scheduling based on simulation can determine what the actual manufacturing system should perform in the planned time horizon. To solve scheduling problems in real environments, the primary manufacturing processes have to be considered, and special attention has to be paid to the preparatory processes. The achievements and experiences of the software application have been very positive. The obtained results encourage the application of this

approach to other multi-objective optimization problems in production systems and processes.

The presented new scheduling models and algorithms can be applied effectively to solving different scheduling problems. One of the possible cases is the scheduling of manufacturing processes to which time-dependent resource constraints of assistant processes are connected. Another potential case is the scheduling of direct manufacturing jobs with due dates on time-dependent sets of parallel resources.

Our research and development project highlights the importance of modeling for the treatment of production planning, scheduling and control problems; in addition it emphasizes the interconnection of theoretical results and practical demands. The main purpose of this paper was to share our experiences and results with researchers and industrial practitioners working in the fields of production information engineering.

Acknowledgements

This research was partially carried out in the framework of the Center of Excellence of Mechatronics and Logistics at the University of Miskolc.

This research was partially connected to the TAMOP-4.2.1.B-10/2/KONV-2010-0001 project with support by the European Union, co-financed by the European Social Fund.

Software development and application of the research results in practice is supported by Fehrer Hungaria Járműipari Kft. (Mór, Hungary).

References

[1] A. Allahverdi, C. T. Ng, T. C. E. Cheng, M. Y. Kovalyov, A Survey of Scheduling Problems with Setup Times or Costs, European Journal of Operational Research, Vol. 187, pp. 985-1032, 2008

[2] A. A. Lazarev, D. I. Arkhipov, F. Werner, Scheduling Jobs with Equal Processing Times on a Single Machine: Minimizing Maximum Lateness and Makespan, Optimization Letters, pp. 1-13, 2016

[3] A. Gharbi, M. Haouari, Optimal Parallel Machines Scheduling with Availability Constraints, Discrete Applied Mathematics, Vol. 148, pp. 63-87, 2005

[4] C. Koulamas, The Total Tardiness Problem: Review and Extensions, Operations Research, Vol. 42, pp. 1025-1041, 1994

[5] D. Lei, Multi-Objective Production Scheduling: a Survey, The International Journal of Advanced Manufacturing Technology, Vol. 43, Issue 9-10, pp. 926-938, 2009

[6] D. Quadt, H. Kuhn, A Taxonomy of Flexible Flow Line Scheduling Procedures, European Journal of Operational Research, Vol. 178, pp. 686-698, 2007

[7] E. Mokotoff, Parallel Machine Scheduling Problems: a Survey, Asia-Pacific Journal of Operational Research, Vol. 18, pp. 193-242, 2001

[8] Gy. Kulcsár, F. Erdélyi, A New Approach to Solve Multi-Objective Scheduling and Rescheduling Tasks, International Journal of Computational Intelligence Research, Vol. 3, Issue 4, pp. 343-351, 2007

[9] Gy. Kulcsár, F. Erdélyi, Modelling and Solving of the Extended Flexible Flow Shop Scheduling Problem, Production Systems and Information Engineering, Vol. 3, pp. 121-139, 2006

[10] Gy. Kulcsár, M. Kulcsárné Forrai, Detailed Production Scheduling Based on Multi-Objective Search and Simulation, Production Systems and Information Engineering, Vol. 6, pp. 41-56, 2013

[11] H. Aytug, M. A. Lawley, K. Mckay, S. Mohan, R. Uzsoy, Executing Production Schedules in the Face of Uncertainties: a Review and Some Future Directions, European Journal of Operational Research, Vol. 161, pp. 86-110, 2005

[12] J. Kaabi, Y. Harrath, A Survey of Parallel Machine Scheduling under Availability Constraints, International Journal of Computer and Information Technology, Vol. 3, Issue 2, pp. 238-245, 2014

[13] J. Lamothe, F. Marmier, M. Dupuy, P. Gaborit, L. Dupont, Scheduling Rules to Minimize Total Tardiness in a Parallel Machine Problem with Setup and Calendar Constraints, Computers & Operations Research, Vol. 39, Issue 6, pp. 1236-1244, 2012

[14] K. Lee, J. Y.-T. Leung, M. L. Pinedo, Scheduling Jobs with Equal Processing Times Subject to Machine Eligibility Constraints, Journal of Scheduling, Vol. 14, Issue 1, pp. 27-38, 2011

[15] M. Kulcsárné Forrai, F. Erdélyi, Gy. Kulcsár, A New Extended Model for Solving Flexible Job Shop Scheduling Problems, Proceedings of the International Conference on Innovative Technologies, IN-TECH, pp. 325-328, 2013

[16] M. L. Pinedo, Planning and Scheduling in Manufacturing and Service, 2nd ed., Springer-Verlag New York, 2009

[17] M. L. Pinedo, Scheduling Theory, Algorithms, and Systems, 3rd ed., Springer-Verlag New York, 2008

[18] M. Pinedo, C. Zacharias, N. Zhu, Scheduling in the Service Industries: an Overview, Journal of Systems Science and Systems Engineering, Vol. 24, Issue 1, pp. 1-48, 2015

[19] M. X. Weng, J. Lu, H. Ren, Unrelated Parallel Machine Scheduling with Setup Consideration and a Total Weighted Completion Time Objective, International Journal of Production Economics, Vol. 70, pp. 215-226, 2001

[20] P. Bikfalvi, F. Erdélyi, Gy. Kulcsár, T. Tóth, M. Kulcsárné Forrai, On Some Functions of the MES Applications Supporting Production Operations Management, In G. Bognár, T. Tóth (eds.): Applied Information Science, Engineering and Technology: Selected Topics from the Field of Production Information Engineering and IT for Manufacturing: Theory and Practice, (Topics in Intelligent Engineering and Informatics; 7), Berlin: Springer-Verlag, pp. 103-129, 2014

[21] P. Brucker, Scheduling Algorithms, 5th ed., Springer-Verlag Berlin Heidelberg, 2007

[22] T. C. E. Cheng, C. C. S. Sin, A State-of-The-Art Review of Parallel-Machine Scheduling Research, European Journal of Operational Research, Vol. 47, pp. 271-292, 1990

[23] W. Wang, Flexible Flow Shop Scheduling: Optimum, Heuristics, and Artificial Intelligence Solutions, Expert Systems, Vol. 22, Issue 2, pp. 78-85, 2005

[24] Y. Lin, W. Li, Parallel Machine Scheduling of Machine-Dependent Jobs with Unit-Length, European Journal of Operational Research, Vol. 156, Issue 1, pp. 261-266, 2004

[25] Y. Ma, C. B. Chu, C. R. Zuo, A Survey of Scheduling with Deterministic Machine Availability Constraints, Computers & Industrial Engineering, Vol. 58, pp. 199-211, 2010

Information-Theoretic Analysis of Iris Biometrics for Biometric Cryptography

Sasa Adamovic, Milan Milosavljevic, Mladen Veinovic, Marko Sarac, Aleksandar Jevremovic

Department of Informatics and Computing, Singidunum University
32 Danijelova Street, 11000 Belgrade, Serbia
(sadamovic, mmilosavljevic, mveinovic, msarac, ajevremovic)@singidunum.ac.rs

Abstract: This paper presents a rigorous information-theoretic analysis of iris biometrics with the aim to develop optimized biometric cryptosystems. By estimating local entropy and mutual information, we identify the iris regions that are most suitable for these purposes. Parameter optimization of the appropriate wavelet transform produces higher entropy and low mutual information in the transformation domain. This establishes an effective framework for the development of systems for the extraction of truly random sequences from iris biometrics, while not compromising its proven authentication features.

Keywords: iris biometrics; image analysis; information theory; image texture; biometric cryptosystems

1 Introduction

According to some estimates, the entire field of information protection should make a radical qualitative leap and a shift of the fundamental paradigm of computer security towards a paradigm of information-theoretic security [1]. Such a shift would allow the creation of an entire class of cryptographic mechanisms whose compromisation would be independent from the attacker's computing power. All this points to a new position of this discipline within the general theory and practice of information protection systems. Information analysis of source biometric data is crucial for construing concrete solutions of a biometric cryptosystem with a theoretically guaranteed performance rate.

Biometrics has established itself as a significant source of cryptologic parameters in the domain of reliable and practically acceptable authentication. Biometric systems are based on physical and behavioral characteristics of human beings such as fingerprint, voice, face, iris and others. The strength and resistance of these systems are directly related to the natural amount of information present in a biometric source. In order to estimate the maximum quantity of information, one

must have a good understanding of biometric data specific to a particular source, as well as the technology used to precisely read and extract information.

The original concept of "Biometric encryption" was applied to fingerprints in 1994. The pioneer in this area is Dr. George Tomko, the founder of Mytec Technologies, Toronto, Canada. Ever since, numerous researchers have contributed to this and other related technologies. Besides the Biometric Encryption, the term biometric cryptosystems is also used. We shall use the abbreviation "BC" in the remainder of this text to denote biometric cryptosystems. Generating keys for various cryptographic purposes based on biometric data is an important idea. A prime example of one such system is given in [2], where authors achieve a promising result (FRR (false rejection rate) = 0,47%, FAR (false acceptance rate) = 0%, key length = 140 bits) in an iris biometry application. This system enables a multipurpose use of a biometric template, without the possibility of compromisation. This result opens a door to a wide application in cryptographic protection mechanisms. Furthermore, the authors carried out the analysis of noise and errors that occurred while forming the iris biometric template. Pertaining results enabled them to select an adequate code, which is optimized with regard to the maximum allowed capacity determined by the iris biometric source. A 2D Gabor wavelet was used to extract 2048 bits of phase information which produces approximately 249 degrees of freedom [3].

Most BC systems applied to a variety of biometric characteristics produce fairly long keys (140 bits [2], 186 bits [4], 240 [5] bits). This imposes the following question: what is the true quantity of consistent information available within biometric data, based on which it is possible to generate a cryptographic key? This very information serves as the material for key generation. In that case, the key itself cannot be longer than the quantity of biometric information. If this were to happen, it would only speak of overly high performance settings for the debugging code, which inevitably leads to FAR being greater than zero. In addition to acceptable FRR values, algorithms suggested by the majority of authors [4, 5, 6, 7] have FAR values greater than zero percent. From our point of view, practical application of such BC systems is unacceptable. Also, other algorithmic solutions [2, 8, 9, 10] were proposed that result in both FAR being equal to zero percent and FRR having acceptable values. However, it remains to be determined whether those solutions fully utilize the true capacity of the system and to identify the system's maximum level of effectiveness [11].

Considering the above-stated, we assume that a strong information-theoretic foundation and the application of the Theory of Perfect Cyphers are indispensable for successfully developing BC systems. The theory was proposed by Shannon [12]. The information-theoretic analysis of biometrics, as a special information source, would provide concrete solutions for development of BC systems with theoretically guaranteed performance.

This paper is concerned with a rigorous information-theoretic analysis of iris. We apply measures of information (entropy, local entropy, and mutual information) to identify iris regions that are most suitable for generation of cryptologic keys. Also, optimization of the parameters of the transform function produces higher entropy and reduced mutual information in the transformation domain. This establishes the foundation for development of a BC system for estimation of truly random or consistent bits from the iris biometric source. In addition, the authors are concerned with the complex procedure of processing iris biometric data [13]. Biometric data contains various types of noise that may significantly increase the degree of variability, which further alters the quality of information obtained.

The remainder of this work is organised as follows. The next section discusses the biometric database used, which is followed by a detailed information-theoretic analysis consisting of calculating local entropy over a texture of iris image after the normalization phase, determining optimal parameters in the transformation domain (iris coding phase), modeling of the iris information source by means of Shannon approximation models, measuring mutual information between identical and different irises, and the setup of an information-theoretic foundation for iris biometrics. The final section comprises the conclusion and an overview of the contributions of our work.

2 Information Analysis and Experimental Results

We used the CASIA Iris Image Database version 4.0, for the experimental portion of our work. This database was created by researchers from the Institute of Automation, Chinese Academy of Sciences, and it contains several thousand iris images [14]. Several versions of the database were offered free of charge to the international biometric research community. Over 4000 users from 70 countries have downloaded the CASIA database so far and a vast number of researchers have used it in their work.

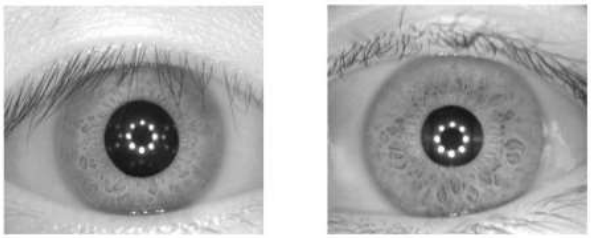


Figure 1
Samples of iris images from the CASIA-Iris V4 database

Fig. 1 provides test samples of iris images in the gray-scale format, acquired by means of special cameras.

2.1 Analysis of Iris Image Texture

In the first part of the information analysis we measure the entropy of an iris texture image after the normalization phase. The number of iris rings is 20 with 240 points on each. By means of the Daugman's rubber sheet [15] model, we obtain a gray-scale image with a resolution of 20 x 240 pixels. Pixel depth is 8 bits, whereby each pixel is represented by a gray shade from the 0 to 255 range of decimal values. Fig. 2 depicts the normalization process and the resulting rectangular iris texture.

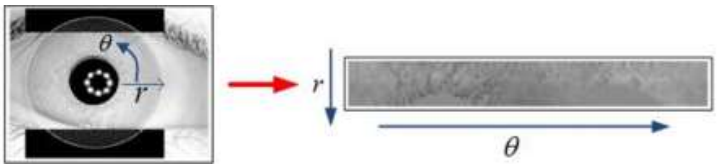


Figure 2
Iris texture in the first phase

In the next step, we calculate the local entropy over the iris texture we obtained in the manner described above. Due to the nature of data, we employed the method commonly used to determine the entropy of two-dimensional signals [16, 17]. This method essentially utilizes the Shannon entropy [18]. Entropy is most conveniently defined as an average quantity of information or the measure of uncertainty of an information source (iris, in our case). For a known probability p, entropy of an event is calculated by:

$$H = -\sum_{i=1}^{l} p_i \log_b p_i \quad (1)$$

where p_i pertains to symbol probabilities obtained through image histograms.

The value of local entropy varies based on the chosen window size. The window is square-shaped and it represents the number of included neighboring pixels. The values obtained are represented by means of a binary logarithm, where 1 bit is the unit of quantity of information. The chosen method allows for the use of a varying number of neighboring pixels, which is quite similar to Shannon's approximation models – the concept used in modeling natural language as an information source.

Fig. 3 depicts a 3D model of local entropy for the chosen windows size of 9x9. The model represents average local entropy values for each individual pixel position. The illustration reveals that the first circular iris region (next to the pupil) has a larger local entropy. This is clearly seen on the y-axis that displays the quantity of information (i.e., achieving 5 bits per pixel out of the maximum possible 8 bits per pixel). Also, a closed circular contour in the X, Z, clearly points out to the higher entropy region. The average local entropy in this experiment differs significantly in the first (4.4412 bits per pixel) and the second region (3.6020 bit per pixel). Based on the results obtained, we adopt the division of iris by regions whereby the first region becomes the primary interest of our research.

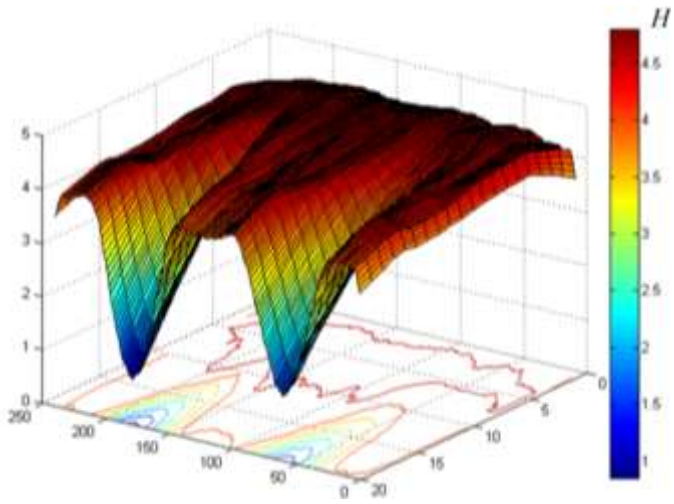


Figure 3
A 3D information model of local entropy (CASIA Iris Image Database version 4.0)

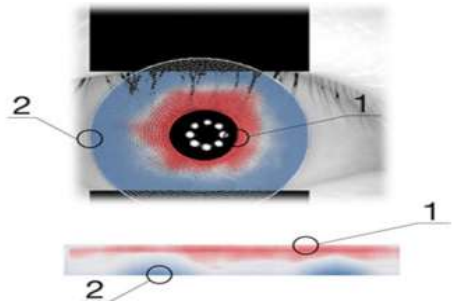


Figure 4
Division of iris by regions (red color (1) - higher information value, blue color (2) - lower information value)

Fig. 4 illustrates the first region of iris, determined to be of higher information value, whereas the second region is characterized by a decline of quality as measured by local entropy.

Low quality of the second region is attributable to eyelids and lashes, but also the automatic segmentation phase where algorithms do not attain 100% accuracy. Regardless of the shortcomings, we observe that both the left and the right side of the iris in the second region have lower entropy (area not covered by eyelids and lashes). Therefore, the second region cannot be used to extract material for generating cryptologic parameters. Moreover, due to the impreciseness of segmentation algorithms, practical applications of BC systems lead to higher FRR parameters. Researchers attempt to remedy this problem by designing various concatenated security codes that often lead to FAR values greater than zero percent.

2.2 Analysis of Optimal Parameters in the Transform Domain – Iris Coding

We conduct the next information analysis on a biometric iris template code or iris information source following the coding phase. There are several important parameters for the algorithm utilized in the coding phase. This phase results in an iris biometric template. The success of this phase depends on the optimal choice of parameter values used to provide high entropy of the iris code and the maximum possible quantity of consistent bits. The following comparative analysis was carried out in the part of the process where iris code is formed. In fact, by analyzing the biometric template – iris code, we conduct an analysis over the iris information source.

The parameters of interest include radial and angular resolution (r and θ); in other words, the parameters that produce the number of points in the iris image that will be coded in each iris, as well as filter parameters used to extract only unique iris characteristics. Filter parameters include: filter number N, wavelength λ_n (in pixels), bandwidth given as σ/f, and the multiplicative factor between center wavelengths of successive filters α.

It is known that altering the wavelength parameter $\lambda_n$of the filter provides the opportunity to increase the entropy of the source and the number of consistent bits. For an in-depth discussion and technical details please see [19] and [20].

Figure 5
Biometric template – binary iris code

Fig. 5 shows an example of the iris code (in binary format) with masked portions of iris code that contain errors caused by eyelashes and lids. The following results were obtained using a methodology similar to that applied in the preceding information analysis. We analyzed iris code at the binary matrix level over which we calculated local entropy. The dimension of the iris code after the coding phase is 20 x 480 pixels, with the pixel depth of 1 bit. Afterwards, we carried out a comparative information-theoretic analysis encompassing the entire iris code.

We conducted a comparative analysis of the same iris population with the aim to confirm our assumption from the previous analysis. Three biometric templates (iris code) were generated for each iris for parameter values $\lambda_n = \{12,18,24\}$.

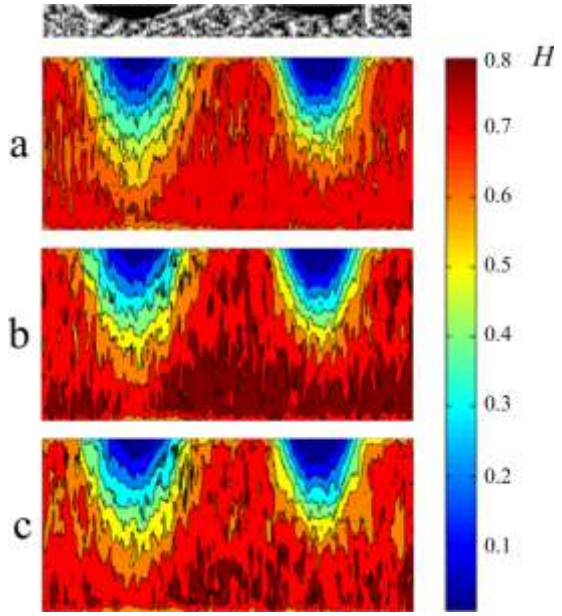


Figure 6
Iris coding transformation domain - varying the λ parameter a: λ = 12, b: λ = 18, c: λ = 24

As could be seen in Fig. 6 (a), the average information quantity or local entropy is equally distributed over the entire surface of iris code for parameter value λ = 12. This results in an average information quantity of 0.8412 per bit. Judging from the analysis of iris texture previously described, we cannot expect equal quality of information in both iris regions.

Fig. 6 (b) is the result of parameter value λ = 18. In the first region, we observe a characteristic information that is not equally present in the second region of the iris code. We obtain an average information quantity of 0.8189 per bit. This characteristic information corresponds to the results illustrated in Fig. 3.

The last measurement uses the value of λ = 24 and is depicted in Fig. 6. (c). The characteristic information is slowly vanishing from the first region, while it is almost nonexistent in the second region. This time, we obtain an average information quantity of 0.7982 per bit.

Varying the λ parameter is important for identifying the optimum filter values, which in turn produce stable and consistent bits with maximum entropy. Please note that by maximum entropy we actually refer to the best achieved compromise between maximum entropy and the largest number of consistent bits. Consistent bits comprise the characteristic information of the iris biometric source. This is of paramount importance for achieving a sound theoretical framework for development of BC systems. In our case, this compromise is arrived at for a filter bandwidth of λ = 18.

Table 1
Local entropy values by iris code regions

Window 9 x 9 / **Filter parameters**	**Region 1 - iris (bit per pixel)**	**Region 2 - iris (bit per pixel)**	**Regions 1 and 2 - iris (bit per pixel)**
λ = 12	0.9351	0.7251	0.8412
λ = 18	0.9125	0.6997	0.8189
λ = 24	0.8901	0.6776	0.7982

Table 1 presents summary results that clearly establish a significant difference between local entropy values of the first and the second region.

2.3 Analysis of Mutual Information between the Same and Different Irises

In this portion of information-theoretic analysis we use $\lambda = 18$ as the optimal filter bandwidth value in the iris coding transformation domain. By applying Shannon's approximation models [18], we determine the maximum entropy in the first region of the iris code. Adopting a method of approximation is crucial for properly estimating mutual information of identical and different irises.

The method as a whole is comprised of simple algorithms that were particularly developed for approximation models of orders II to V. We analyzed only the first iris code region. In the coding phase, we formed the matrix, row by row, based on the radial vectors in the normalization phase. The first row represents bits obtained through the first iris ring (radial vector). The rings are indicated in ascending order with the first being located closest to the pupil and the tenth farthest away from it since we only use the first iris region.

For instance, for an order II approximation, we assume a set of 4 possible messages, where messages are represented by numbers 1 to 4. For an order III approximation, we use numbers 1 to 8. Similar reasoning applies to higher order approximations. We assume that all messages have equal probabilities. For an order II approximation, iris code is decoded using a bigram. Fig. 7 provides an example of iris code decoding for an order II approximation by means of a dictionary. The process is similar for higher order approximations, with the number of words in the dictionary and the word length being increased.

Fig. 8 shows entropy levels for approximations of order II to V. Approximation V results in an entropy of 0.8208 per bit, which is the maximum value achieved by optimizing parameters in the transformation domain. Upon a closer look, entropy values for order V approximation are almost identical to local entropy for $\lambda = 18$.

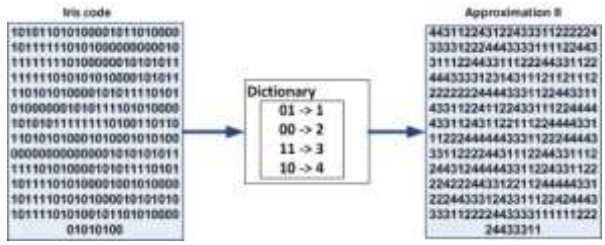


Figure 7
An example of order II approximation

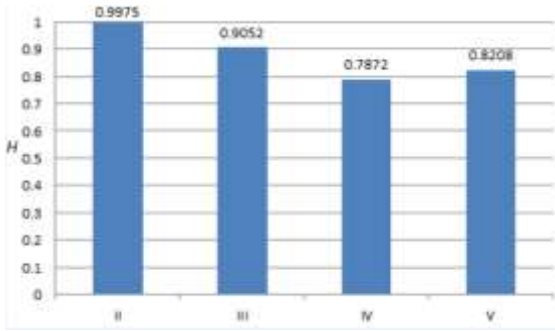


Figure 8
First iris region entropy for various approximation models (II - V)

Approximating for orders higher than V is possible, provided that all words in the dictionary exist. For this particular experiment (iris code) and order six approximation, many of the words in the dictionary had zero probability. This would not lead to a reliable entropy level. Hence, we restrict iris code to order V approximation. To that end, order V approximation is adopted for estimating mutual information. This approximation results in the entropy of 0.8208 per bit, which sums up to 3490 bits in the first iris code region.

2.4 Modeling of Iris Information Source using Shannon's Approximation Models

The following important analysis provides the calculation of mutual information between identical and different irises for the first region over the data obtained by order V approximation. The significance of this analysis is rather high concerning the security of BC systems. The method we use to measure mutual information I(iris x;iris y) between the two iris code regions is given by the following expression (2):

$$I(A,B) = \sum_{b\in B}\sum_{a\in A} p(a,b) * \log\left(\frac{p(a,b)}{p(a)p(b)}\right) \quad (2)$$

where:

- $p(a,b)$- joint probability distribution function of A and B
- $p(a)$ - marginal probability distribution function of A
- $p(b)$ - marginal probability distribution function of B

In the sense of probability theory, relative entropy of a system measures the distance between two probability distributions. In this way, mutual information is defined as (3):

$$\mathrm{I(A;B)} = \mathrm{H(A)} + \mathrm{H(B)} - \mathrm{H(A,B)}; \quad (3)$$

$$\mathrm{I(A;B)} = \mathrm{H(A)} - \mathrm{H(A|B)} = \mathrm{H(B)} - \mathrm{H(B|A)};$$

where:

- $H(A)$ – marginal entropy of A
- $H(B)$ – marginal entropy of B
- $H(A|B)$ – conditional entropy of A
- $H(B|A)$ – conditional entropy of B
- $H(A,B)$ – joint entropy of A and B

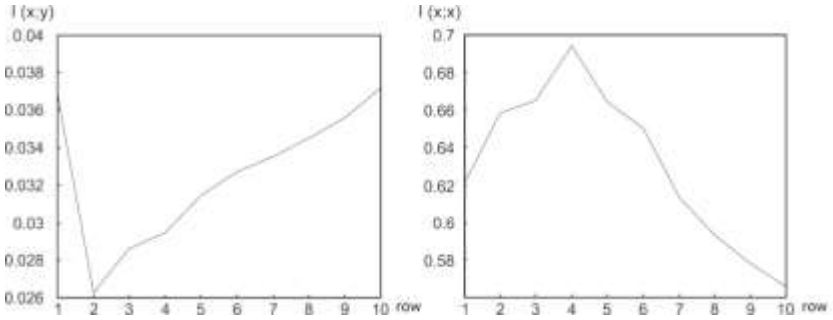


Figure 9

Mutual information between irises: left - different irises, right - the same irises

Fig. 9 illustrates the relationship of mutual information between the same and different irises by rows. Rows are shown on the x-axis. There are 10 rows in the first region and they are numbered in an ascending order, starting from the row closest to the pupil and ending farthest apart from it. Mutual information for two iris codes of a single person is $\mathrm{I(X;Y)} = 0.2078$ per bit or 997 bits totally. Please note that one always desires maximum mutual information between different images of the same iris.

In case of irises belonging to different people (Fig. 9 (a)), the average quantity of mutual information is $\mathrm{I(X;Y)} = 0.1065$ per bit or 511 bits totally. Such results guarantee the presence of consistent bits, in particular for the analysis of mutual information between the same rows. The interval between rows 3 and 6 contains

the maximum mutual information (same person irises), while the interval from the row 2 to row 6 (different people irises) measured minimum mutual information. We attribute this result to the algorithm parameters in the transformation domain.

2.5 Degrees of Freedom vs. Entropy

The complexity of iris code is approximately determined by measuring the Degrees of freedom (DOF, hereafter) over the corpus of different iris codes. DOF is calculated by means of all mutual Hamming distances as a binomial probability distribution.

DOF is also defined as a minimum number of independent coordinates that fully describe the state of a system. For the CASIA database, DOF is 1068, with 1D Gabor wavelet being used for coding of the source [21]. This is an exceptional result and guarantees the uniqueness and independence between different iris templates.

Table 2
Comparison of 1D and 2D wavelet demodulation

Iris template	Wavelet filters	DOF
2048 bits	2D Gabor	249
9600 bits	1D Gabor	1068

Table 2 compares the sizes of generated iris code and the DOF obtained between iris codes generated by 1D and 2D Gabor wavelets. When using a 1D Gabor wavelet, the generated iris code amounts to 9600 bits, whereas a 2D Gabor wavelet results in the iris code of 2048 bits [3].

Table 3
DOF after optimizing the wavelet filters

	Region 1 and 2 (9600 bits)	Region 1 (4800 bits)	Region 2 (4800 bits)
λ_n	DOF	DOF	DOF
12	2946.8	2217.5	935.5
18	1367.3	**1346.1**	396.7
24	654.5	785.0	199.3

Table 3 presents the DOF values by regions and for the iris code as a whole, obtained after the optimization of parameters in the transformation domain. We use three values of the λ_n parameter. For $\lambda_n = 18$ in the first iris region (size of 4800 bits) we measured DOF = 1346. This is a significant improvement compared to the data displayed in Table 2. Furthermore, the iris authentication features have not been compromised in any way.

2.6 Information-Theoretic Framework of Iris Biometrics

We begin by introducing the notation [12] needed to establish the information framework of iris biometry.

- $I(Y; Y')$ - mutual information between images of the same iris;
- $I(X; Y)$ - mutual information between images of different irises;
- $H(X)$ - entropy of iris code;
- $H(X, Y)$ - joint entropy of iris codes for two different eyes;
- $H(Y, Y')$ - entropy of two iris codes of the same person;
- $H(K)$ - joint entropy of two iris codes for the same eye;

Let us assume that the irises Y and Y' belong to Alice and are used in a certain BC system. On the other hand, iris X belongs to Eve, a potential attacker.

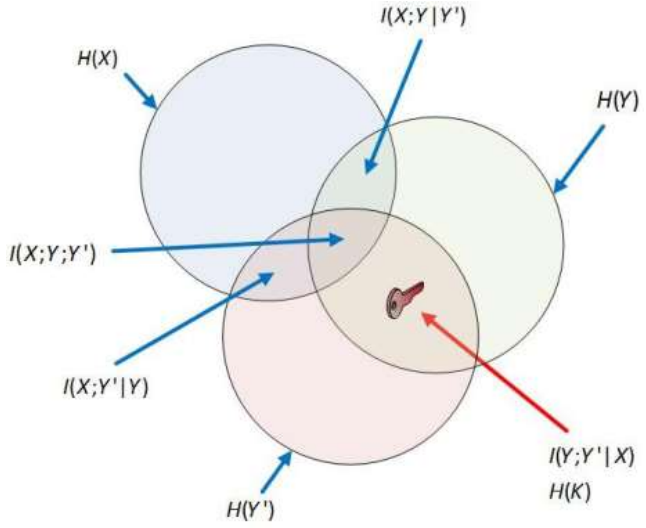


Figure 10

Graphic presentation of information-theoretic framework for development of BC systems

Fig. 10 illustrates an information-theoretic maximum for estimation of biometric keys used for development of BC systems. The diagram contains three random variables X, Y and Y'. It is important to note that the mutual information $I(X; Y; Y')$ is symmetric provided that the three variables are equally independent. A greater dependence between Y and Y', so is the mutual information $I(Y; Y')$ greater. This also entails that the information $I(Y; Y'; X)$ is also greater. In this case, pairwise mutual information $I(X; Y)$ and $I(X; Y')$ have decreased, while the joint information $I(X; Y; Y')$ and $I(Y; Y'|X)$ have potentially increased.

Based on the analyses and measurements of entropy, we can numerically represent the quantity of information that is usable for creating an efficient BC system scheme. It is estimated that on average, the overall quantity of information present in the first region of an iris is $H(\text{iris } X) = H(\text{iris } Y) = H(\text{iris } Y') = 3940$ bits.

Mutual information between images of the same iris is $I(Y;Y') = 997$ bits, while mutual information between images of different irises is $I(X;Y) = 511$ bits. Joint entropy of two iris codes of two people is $H(X,Y) = 7281$ bits, whereas the joint entropy for images of the same iris is $H(Y,Y') = 6709$ bits. Also, we estimate that the overall quantity of useful information (i.e., entropy of the key) is $H(K) = I(Y;Y'|X) = 788$ bits, while the expected mutual information between all three variables (two same-person irises and another person's iris) is $I(X;Y;Y') = I(Y;Y') - I(Y,Y'|X) = 997 - 788 = 209$ bits.

2.7 Discussion

This work presents a method based on complex information-theoretic analysis of iris biometric that aims to extract homogeneous regions of high entropy. Successful extraction of these regions facilitates the development of effective systems for generation of cryptographic keys. Our method includes modeling of information sources – iris biometric. Shannon's model approximations, created real conditions for the application of information measures (entropy, mutual entropy, conditional entropy, joint entropy) to better understand the quality of the iris as biometric data. We also emphasized the importance of optimization wavelet parameters to achieve better results in the transformation domain. The results achieved in the work [11] prove this claim. At the same time, this approach allows for the application of simpler error correction codes with equal False Accept Rate levels, which reduces the overall complexity of this class of systems.

Conclusions

The main aim of this research was to enable the development of a professional class of systems for generation of long keys. We set out to meet the demands of modern cryptosystems relying on the existing components for coding biometric sources that encompass the entire process, starting from the choice of biometry, through imaging and ending with biometric templates.

The information-theoretic analysis used herein for the iris biometric data has confirmed our doubts. Moreover, it has led us to formulate clear goals in terms of raising the bar for system efficacy close to the theoretic maximum. In order to identify iris regions with the richest content of consistent information, we estimated entropy, local entropy, mutual information and employed Shannon approximations to model the information source. We performed parameter optimization of the appropriate wavelet transform with the aim to obtain the highest possible entropy and lowest possible information in the transformation domain.

Numerous authors have designed the schemes for systems that generate cryptographic keys based on the whole iris region. We demonstrated that the whole region cannot be used to develop such systems. The authors bypass the issues of low quality region and insufficient key length by increasing the capacity of error correction codes [11]. For this very reason, it is common among such systems to have FAR values above zero, which is unacceptable from our point of view. In addition, the keys generated in such manner rarely pass the common tests of cryptologic randomness (that include randomness and unpredictability).

Since the topic of this paper lies between biometrics and cryptography, we highlight the necessity of introducing the information-theoretic analysis in the increasingly popular field of biometric cryptography. This should be done with the aim of producing a firm bond between the two disciplines in a manner that is fully compliant with the cryptographic principles and characteristic features of biometric data.

We believe that the information-theoretic analysis employed in the course of development of this system guarantees high security performance needed for applications in law enforcement, military, government and diplomacy.

Acknowledgement

This work was supported by the Ministry of Science and Technological Development of the Republic of Serbia through the project TR32054.

References

[1] Matthieu Bloch, Joa o Barros: Physical-Layer Security: From Information Theory to Security Engineering (Cambridge University Press, 1st edn. 2011)

[2] F. Hao, R. Anderson, J. Daugman: Combining Crypto with Biometrics Effectively, IEEE Transactions on Computers, 2006, 55 (9) pp. 1081-1088

[3] J. Daugman: The Importance of Being Random, Statistical Principles of Iris Recognition, Pattern Recognition, 2003, 36 (2) pp. 279-291

[4] H. A. Garcia-Baleon, V. Alarcon-Aquino, O. Starostenko, et al.: Bimodal Biometric System for Cryptographic Key Generation Using Wavelet, Mexican International Conference on Computer Science- IEEE, 2009, pp. 186-196

[5] F. Hao and C. W. Chan.: Private Key Generation from On-Line Handwritten Signatures, Information Management & Computer Security, 2002, 10 (2) pp. 159-164

[6] P. Tuyls, A. H. M. Akkermans, T. A. M. Kevenaar, et al.: Practical Biometric Authentication with Template Protection, AVBPA'05 Proceedings of the 5th international conference on Audio- and Video-Based Biometric Person Authentication, 2005, pp. 436-446

[7] M. van der Veen, T. Kevenaar, G.-J. Schrijen, et al.: Face Biometrics with Renewable Templates, Proc. SPIE 6072, Security, Steganography, and Watermarking of Multimedia Contents VIII, 60720J, February 2006

[8] F. Monrose, M. K. Reiter, Q. Li, S. Wetzel.: Cryptographic Key Generation from Voice, In Proceedings of the 2001 IEEE Symposium on Security and Privacy, May 2001

[9] T. C. Clancy, N. Kiyavash, D. J. Lin.: Secure Smart Card-Based Fingerprint Authentication, Proc. ACM SIGMM Workshop Biometrics Methods and Application (WBMA), 2003

[10] J. Daugman: How Iris Recognition Works, Circuits and Systems for Video Technology. IEEE Transactions on, 2004, 14, pp. 21-30

[11] S. Adamovic, M. Milosavljevic, M. Veinovic, M. Sarac, A. Jevremovic: 'Fuzzy Commitment Scheme for Generation of Cryptographic Keys Based on Iris Biometrics', IET Biometrics, 2016, DOI: 10.1049/iet-bmt.2016.0061 IET Digital Library, http://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2016.0061

[12] C. E. Shannon.: Communication Theory of Secrecy Systems, Bell System Technical Journal, 1949, 28, pp. 656-715

[13] S. Adamovic, M. Milosavljevic.: Information Analysis of Iris Biometrics for the Needs of Cryptology Key Extraction, Serbian Journal of Electrical Engineering, 2013, 10, 1, pp. 1-12

[14] 'Biometrics Ideal Test', http://biometrics.idealtest.org, accessed 15 October 2012

[15] T. Johar, P. Kaushik.: Iris Segmentation and Normalization using Daugman's Rubber Sheet Model, International Journal of Scientific and Technical Advancements, 2015, 1 (1) pp. 11-14

[16] S. Adamovic, A. G. Savic, M. Milosavljevic, et al.: Texture Analysis of Iris Biometrics based on Adaptive Size Neighborhood Entropy and Linear Discriminant Analysis, International Scientific Conference – Sinteza, Serbia, pp. 658-660, April 2014

[17] R. C. Gonzalez, R. E. Woods, S. L. Eddins: Digital Image Processing Using MATLAB, New Jersey, Prentice Hall, 2003

[18] C. E. Shannon.: A Mathematical Theory of Communication, Bell System Technical Journal, 1948, 27, pp. 379-423, 623-656

[19] T. Lee.: Image Representation using 2D Gabor Wavelets, IEEE Transactions of Pattern Analysis and Machine Intelligence, 1996, 18 (10) pp. 959-971

[20] Raymond W, Yueng.: A new Outlook on Shannon Information Measures, IEEE Transactions on IT., 1995, 37 (3) pp. 466-474

[21] L. Masek.: Recognition of Human Iris Patterns for Biometric Identification Iris Recognition, http://www.csse.uwa.edu.au/~pk/studentprojects/libor/., accessed 15 October 2012

OLOUD - An Ontology for Linked Open University Data

Rita Fleiner[1], Barnabás Szász[2], András Micsik[3]

[1]Department of Applied Informatics, John von Neumann Faculty of Informatics, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, fleiner.rita@nik.uni-obuda.hu

[2]Faculty of Informatics, University of Debrecen, Kassai út 26, 4028 Debrecen, Hungary

[3]Institute for Computer Science and Control, Hungarian Academy of Sciences, Lágymányosi u. 11, 1111 Budapest, Hungary, andras.micsik@sztaki.mta.hu

Abstract: The Ontology for Linked Open University Data (OLOUD) is a practical approach to model course information at a typical Hungarian university. OLOUD aims to integrate data from several sources and provide personal timetables, navigation and other types of help for students and lecturers. The modeled domains include curricula, subjects, courses, semesters and personnel, but also buildings and events. Although there are several ontologies for the mentioned domains, selecting a set of ontologies fitting our use case was not an easy task. We summarize problems we met such as missing links, inconsistencies as well as many overlaps between ontologies. Finally, OLOUD acts as a glue for a selection of existing ontologies, and thus enables us to formulate SPARQL queries for a wide range of practical questions of university students.

Keywords: Linked Open Data; Linked Open University Data; Ontology; OWL

1 Introduction

One aspect of the Smart University or Smart Campus concept is to improve the teaching and learning environment using modern data fusion and data consumption techniques. A campus has a large number of people with a substantial set of common information needs [1]. Therefore, it is of great importance in this area to establish a common data model which enables the interconnection of fragmented data from heterogeneous data sources.

In this paper, we focus on university course information as a special segment of the open data in the higher education domain: The aim is to facilitate the implementation of Smart Universities by defining a common data model for

course information. We chose the ontological representation as the most modern description method for the problem domain. The ontology can be used as a data model to influence or integrate traditional SQL databases, as an RDF schema in triple stores, and in reasoners or rule systems to enhance collected data. The original objective was to develop a generic data model for university 'course related' data. During the work, we noticed that though the Bologna Process ensures a certain level of compatibility for education systems in the EU, this does not reach deeper constructs of the educational model. We found that especially the meanings of course, subject and study programme are quite different in currently available educational models in Europe. Therefore, we decided to base our work on the Hungarian concepts in this field. During the ontology development, we relied on the existing concept definitions and data structures used by university information systems in Hungary such as the Neptun student information system[1] or the Moodle e-learning platform[2].

The aim of this paper is to describe the process of developing the Ontology for Linked Open University Data[3] and to show how it was used to generate Linked Data for courses at the Óbuda University. In order to achieve our goal, we defined the following partial objectives:

- to review the existing ontologies in this field and evaluate their possible usage,
- to reveal the use cases and objectives of the new ontology and define the basic concepts in the domain,
- to present the ontology development process,
- to describe the generation of university 'course related' data as LOD according to the ontology,
- to demonstrate the possible use of the above ontology and data by presenting SPARQL queries according to the questions in the use cases.

In Section 2 potential use cases are explored for the possible uses of our ontology. In Section 3 we explore existing work related to our goals, ontologies that can be used to model university 'course related' open data and identify gaps in existing ontologies. Section 4 describes our new ontology and how it re-uses other existing ontologies. Section 5 is about the evaluation and dataset generation. Finally, we conclude in Section 6.

[1] https://neptun.uni-obuda.hu/
[2] https://elearning.uni-obuda.hu/
[3] http://lod.nik.uni-obuda.hu/oloud/

2 Use Cases, Objectives and Concepts

Presenting 'course related' information requires a lot of data originating from multiple information systems at a typical university. As these systems are usually not fully integrated and the access to the data is limited, significant effort is necessary to successfully navigate through the potential obstacles. Foreign students unfamiliar with local specifics might find it even more challenging. With our data model, we want to support the generation and the management of integrated university 'course related' data and the appearance of future mobile and web applications that are built on the use of this data. In the following, the major concepts and tasks are defined that we aim to support with our Linked Data approach.

2.1 Concepts in Hungarian Higher Education

Figure 1 summarizes the major concepts for university students and teachers in Hungary concerning university courses, subjects, curricula and study programme. Though the Bologna Process ensures a certain level of compatibility for education systems in the EU, this does not reach deeper constructs of the educational model. We found that the meaning of these concepts varies in currently available models.

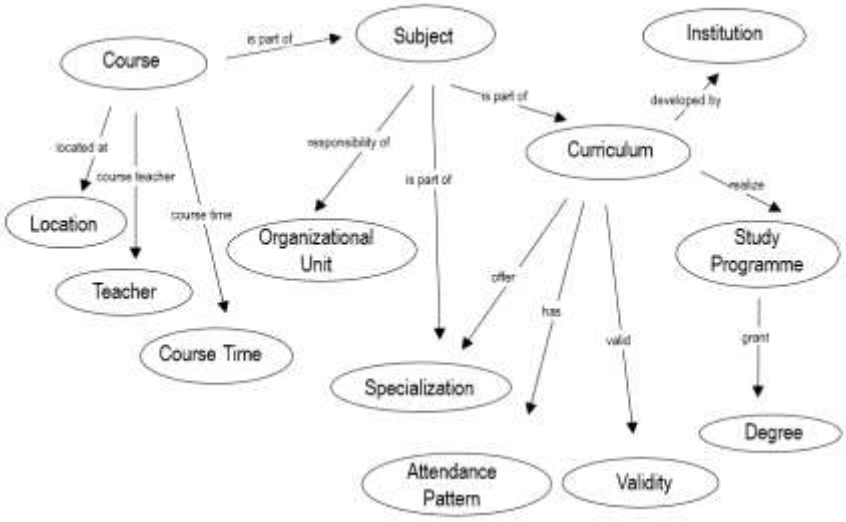


Figure 1
Use case concepts

Table 1 demonstrates how different ontologies use different labels for more or less similar concepts. Therefore, we give a short summary of how these terms are interpreted in Hungarian education. After enrolling to the university each student is assigned to a **Curriculum** (in Hungarian "tanterv"), which is a set of **Subjects** (in Hungarian "tantárgy") and their relations (i.e. dependencies among the subjects). Curriculum might specify **Specializations** (in Hungarian "specializáció"), which are sets of compulsory and optional subjects. Each Curriculum has a specific **Attendance Pattern** (in Hungarian "munkarend", e.g. full-time, part-time, correspondence), and a result as a specific **Degree** (in Hungarian "fokozat", e.g. BSc, MSc, BA, MA, PhD). A Curriculum is in many-to-one relationship with a **Study Programme** (in Hungarian "szak", e.g. Computer Science Engineer), offered by the university. The curriculum is the specification how the Study Programme can be completed. A Study Programme determines the qualification that a student will get after the successful completion of his/her studies. A Study Programme must be accredited by an external body. A Curriculum is valid for a given time interval, meaning that a student can be assigned to it only if his/her enrollment time falls into this period. For each Subject there is an **Organizational Unit** responsible for it. **Courses** (in Hungarian "kurzus") are advertised based on a Subject, have temporal (**Course Time**) and spatial (**Location**) attributes and one or more assigned **Teacher**(s).

Table 1

Possible mapping of core terms

AIISO	**Teach**	**XCRI-CAP**	**Bowlogna**	**MLO-Adv**
Programme	StudyProgram	Course	Study Track	Learning Opportunity Specification
Subject			Subject	Learning Opportunity Specification
Course	Course		Teaching Unit	Learning Opportunity Instance

In the case of Hungarian universities, it is crucial to understand the difference between Subjects and Courses. Course is the elementary unit of the educational process, where date, location, teacher and students are assigned. Course is the framework, which has a specific training type (like lecture, practice or seminar) and has some requirements that students must complete. Courses are organized in individual sessions in a weekly or a custom cadence within a semester, and such a course event is called the session of the course. Subject is a higher-level component of the training process, it is the unit of the curriculum with a specified training content, and fulfillment is rewarded with a number of credits. It may contain more courses, which all must be completed for the completion of a subject.

2.2 Use Cases

In the following we list some of the tasks we aim to support with our Linked Data approach. Courses are organized into a series of lectures, lab exercises or seminars, either in a weekly or in a custom cadence within a semester. There might be multiple labs advertised for a course, so students can choose the most suitable to their circumstances. This creates the challenge of assembling a personal timetable for students and lecturers avoiding conflicts and considering personal preferences and requirements. Students would benefit from an integrated view containing course description (title, identifier, abstract, and dependencies), course time and location. A personal information service may provide students with on-demand information about their daily schedule, navigation to the next lecture, overlaps of classes, etc.

There is also a need for 'long term' planning. Quite often there are no predefined course timetables at Hungarian universities, just a list of courses to be completed, and a dependency graph among the courses, which defines the prerequisites for each. Some courses are advertised only in every second semester. Some universities recommend a specific order of courses, but following such an order breaks easily if for example a single course is not completed in the suggested semester. Thus, students face a kind of constraint satisfaction problem to solve at the beginning of each semester. For this purpose, students need a personal advisor recommending the best way for them to fulfill the curriculum requirements. This advisor needs to consider where the student is on his/her roadmap, what courses they should focus on, what are the personal preferences (e.g. preferred number of courses or credits per semester) and what courses are being advertised.

University resources (rooms, equipment) are used by multiple faculties. They can be booked for regular courses, exams in the exam period and other events. Different types of events may have separate registries, thus blocking an overall view of anticipated resource usage. One needs at least an overall list of reservations by the reserving person, location and date.

According to the above use cases a set of competency questions was defined as the requirement of the OLOUD ontology.

- What are the attributes of a given curriculum, like start of validity, end of validity, study programme, degree, language, attendance pattern?
- What type of specializations exist in each curriculum?
- What are the compulsory subjects of a given curriculum?
- What are the compulsory subjects of a given specialization?
- What are the attributes of a given subject, like responsible teacher, credit number, degree, language, attendance pattern?
- What type of courses consists of a given subject?
- What are the courses for a given subject in a specific semester?
- What are the prerequisite subjects of a given subject?

- What are the attributes of a given course, like the teacher, the course type, the requirement type and hours?
- On which day a given course is, what time does it start and how long it is?
- What is the location of a given course?
- What courses are at the same time partially or completely overlapping?
- What is the course schedule (with course identifier, time and lecturer) for a specific lecture hall or lab?
- What is the navigation route between two course locations?
- What are the dates of the individual sessions of a given course?
- What event is going to be at a specific location at a specific time?

In the following we use the above formulated questions as competency questions for the evaluation of the ontology.

3 Related Work and Ontologies

In this section, existing work is discussed related to linked data in the educational field. Linked Universities[4] and Linked Education[5] are two European initiatives created to enable education with the power of Linked Data. Linked Universities is an alliance of European universities engaged in exposing their public data as linked data. It promotes a set of vocabularies describing 'academic related' entities. LinkedEducation.org is an open platform aimed at further promoting the use of Linked Data for educational purposes.

The Open University in the UK was the first university that created a linked data platform to expose information from its departments. The evolution process of the Open University Linked Open Data platform is described in [2], [3]. This process started as a research experiment and evolved to a data hub for the open content of the university. The platform is now the key information service at the Open University, with several applications and websites exploiting linked data through data.open.ac.uk and establishing connections with other educational institutions and information providers. In the publications, the authors describe the main milestones and tasks accomplished to achieve this state. The Open University datasets can be classified into the following six groups: open educational resources, scientific production, social media, organizational data, research project output and publication metadata.

The main difference with Óbuda University is the lack of navigation and timetable data at the Open University. There are 125 classes and 785 properties from 57 public vocabularies to describe the data at the Open University. Their main effort

[4]http://linkeduniversities.org /
[5]http://linkededucation.org

in the modelling was to reuse the most matching terms from existing vocabularies directly instead of being restricted to the semantics of only a few widely-used ontologies. The large number of the used vocabularies, the redundancy in the data and in the used properties are the consequences of their approach.

The general process for building linked open university data and a use case at Tsinghua University are described in [4]. Procedures like choosing datasets and vocabularies, collecting and processing data, converting data into RDF and interlinking datasets are studied. The datasets unfortunately are not available through public SPARQL endpoint.

The Lucero project analyzed open educational datasets in 2012 [5]. Linked Open Datasets in four universities and four broader educational projects were studied and the most commonly used vocabularies, classes and properties were described. In this case, no representations for course, semester or lecture room concepts were found.

The state of linked data for education is studied in [6]. They collect existing datasets explicitly related to the education field, extract key information, and analyze them. The goal is to better understand what is already available to application developers in this area, what common practices are being used and how the considered datasets connect with each other through common content and vocabulary reuse. They found 144 different vocabularies used in 'education related' datasets. The most popular vocabularies are not specific to education, but are used to represent general concepts and relations, such as resource metadata (Dublin Core [7]), people (FOAF [8]), topics (SKOS [9]), time (W3C Time Ontology [10]) and bibliography (BIBO [11]). More education specific vocabularies are also widely used, such as the Academic Institution Internal Structure Ontology (AIISO [12]), or the Model of Learning Opportunities (MLO [13]).

We found a very useful review of vocabularies and ontologies for modelling course information in higher education [14]. Parts of the following section were influenced by this work.

3.1 Ontologies for Education

The scope of this section is to review existing ontologies that are available to describe various aspects of educational courses and to evaluate whether they can be used to model 'course related' information in the Hungarian higher educational landscape. For simplicity, in the following we refer to ontologies, vocabularies and lighter schema constructs as ontology uniformly. Table 2 provides an overview about ontologies related to our use case and their description targets. There are big differences in the interpretation and in the elaboration of the used terms in the various ontologies.

Table 2
Coverage of relevant ontologies

Ontology	Course	Subject	Curricula, Study Prog.	Specialization	Degree	Teacher	Organization	Time	Location
FOAF						✓	✓		
Vcard						✓	✓		
Event	✓							✓	
W3 Time								✓	
iLoc									✓
Aiiso +	✓	✓	✓	✓		✓	✓		
Teach	✓		✓	✓	✓	✓		✓	✓
XCRI			✓		✓		✓		
Course-Ware	✓								
VIVO	✓			✓		✓		✓	
MLO, ECIM	✓	✓			✓		✓	✓	
Bowlogna	✓		✓	✓	✓	✓	✓		

AIISO (Academic Institution Internal Structure Ontology) provides classes and properties to describe the structure of an academic institution. It is designed to be used in conjunction with the Participation ontology [15], which stands for describing the roles that people play within groups. Participation has only one class, but any domain can extend it by creating subclasses for their own roles within their areas of expertise. AIISO Roles [16] is an example for such extension; it describes roles that people play in an academic institution. The AIISO ontology proved to be useful in our work because it distinguishes at class level the Course and the Subject concepts. These classes are subclasses of KnowledgeGrouping. There is a note in AIISO that this class became deprecated. Probably it was a plan of the authors, but there is not any information on how and when this would be done. AIISO offers only a few properties to describe courses (i.e. code, description, teaches and responsibility) and is mainly used to connect subjects and courses with the organizational structure of the university.

In our work we experienced problems in reusing some properties from the AIISO ontology because they were defined as rdf:Property and in the automatic conversion to OWL the development tool Protégé[6] decided wrongly to use owl:ObjectProperty instead of owl:DataTypeProperty. Another difficulty we faced was that although there exist classes like Programme and Module, their precise meaning is not defined, so their usage is ambiguous. Besides all this AIISO was chosen to form the basis of our data model because the structure of the concepts in this ontology fit into the Hungarian system the best.

[6] http://protege.stanford.edu/

TEACH [17] is a lightweight vocabulary providing detailed properties to describe a course, but it does not model the provider of the course. The concepts in TEACH lack some important features that are essential for our purposes. For example, the concept 'Subject' is necessary to describe university courses, which does not exist in TEACH. Another problem was that one would expect an owl:DataTypeProperty based on the example data for TEACH, but the ontology itself declared the specific property (e.g. teach:courseDescription, teach:ects) as owl:ObjectProperty. These shortcomings were fixed, and the corrected version of the TEACH ontology (http://lod.nik.uni-obuda.hu/teach-fixed.owl) was used in the first phase of our work. In TEACH one can find further problems though. For example, the properties hasAssignment, hasAssignmentMaterial, and hasCourseMaterial have appeared in the index of terms of TEACH. However they are not defined in the vocabulary specification. The classes Student and Lecture are defined but there is no property available in the ontology to relate them to a Course. A potential disadvantage of the TEACH ontology is that it is not linked into other ontologies and some of the definitions are missing or do not have domains or ranges specified. Because of the above problems of this vocabulary, we decided not to use it.

XCRI-CAP [18] is the abbreviation for eXchanging Course Related Information, Course Advertising Profile. The term *course* in the UK is equivalent with the term *study programme* in Hungary, thus XRI-CAP does not contain the description about the course and subject in our terminology. XCRI-CAP is the UK standard for describing study programme marketing information. XCRI represents a lot of data about the provider and the programme and it also differentiates between a programme and the particular presentation of it. XCRI-CAP is in XML format and does not exist in RDF.

The ReSIST Courseware Ontology [19] is a simple ontology with only four classes and many properties like title, teacher, credits, prerequisites, assessment method, etc. It was developed within the ReSIST project between 2006 and 2009. It is an early ontology without any usage at present. The trouble with this ontology is that it is closely related to the Aktors ontology, which is no longer defined anywhere online.

The Metadata for Learning Opportunities (MLO) Advertising ontology is similar in a way to XCRI-CAP, because its purpose is to standardize the specifications for describing and exchanging information about learning opportunities. It can be considered the European equivalent of the British Standard XCRI-CAP for advertising learning opportunities. MLO-Adv contains the following four classes:

- Learning Opportunity Object: an abstract resource used within the context of education or training. It has the following three subclasses:
- Learning Opportunity Provider: a person or organization that offers the learning opportunities

- Learning Opportunity Specification: description of a learning opportunity, consisting of information that will be consistent across multiple instances of the learning opportunity.
- Learning Opportunity Instance: single occurrence of a learning opportunity, it might have a particular date or location.

MLO includes some properties from Dublin Core Elements such as contributor, date, description, identifier, subject, title, and type. ECIM [20] is an extension of MLO, which provides a common format for representing credits awarded for completion of a learning opportunity. XCRI, MLO and ECIM ontologies are similar in that they differentiate between a course specification and a course instance or course offering. The specification contains information about a course or a study programme that remains consistent from one presentation to the next, whereas the instance defines those aspects that vary between presentations for example location or start date. This has the advantage that there will be a smaller amount of data that needs to be updated between years and offerings.

The goal of the VIVO ontology [21] is to represent academic research communities, and thus it enables the discovery of researcher interests, activities, and accomplishments. In a later phase of our work VIVO may be useful to represent research groups within the university, including researchers' grants and external roles. Currently, its focus is quite different from the focus of OLOUD, for example VIVO has its own Course class, but its main properties are credits and prerequisites.

The Bowlogna ontology [22] describes terms used by the Bologna process. It can represent the departments, the teaching units together with information about their ECTS credits and teaching language. It can also be used to store students' examinations, their results and degrees. Although it aimed at providing a standard schema for European universities, in our modeling work we did not find any usage of it.

The main goal of this study was to reveal the usability of the above ontologies in our use case. The following general consequences were drawn:

- It is crucial to understand the meaning of the main concepts and the relationships among them in case of each ontology. Unfortunately, in most cases these ontologies use essential concepts without defining their meaning (i.e. what do concepts like course, subject, module, study programme exactly mean and how do they relate with each other?).
- A specific ontology is usable only if the definitions of the main concepts fit into the use case. Furthermore, in the decision process it is important to see what kind of implementation is used in case of a certain property or relationship (e.g. defining temporal information of a course can be achieved in various ways, but does the actual one satisfy our requirements?).
- The correctness of the formal description of the ontology is important. If it contains shortcomings and mistakes, its reuse is cumbersome.

- Currently there is no other ontology suitable for the use case scenarios described in the paper. Existing ontologies miss properties and thus cannot provide a full description of teaching activities. Furthermore, existing ontologies contradict each other in the naming and semantics of subject, course, curriculum, etc.

Basically, we had to find a set of ontologies filling all capability columns with a minimal number of overlaps. The selection criteria were also determined by several rational considerations, like availability, maintenance, usage and modularity. Before the final selection was made, we had to harmonize the term usage and adapt terms to the Hungarian system if it was possible at all. The final decision was to base our work on the AIISO ontology, because the structure of the concepts Course-Subject-Programme in this ontology fit into the Hungarian system the best.

4 Ontology Development

The main motivation of developing a new ontology born during the first trial implementation: In the domain of university courses there are existing ontologies and our initial approach was to build a linked open university dataset using these existing ones. It turned out that the existing ontologies in this field do not cover fully our requirements and some of them contain mistakes. Facts and connections related to the course-subject-curriculum concepts cannot be fully described by the existing ontologies, only very partially. This recognition confirmed the purpose of the new ontology development.

4.1 Methodology

The five-star model of good Linked Data vocabulary use [23] acted as a guideline during our work. Our aim was to design a 4-star vocabulary. The following rules were applied to restrict the potential interpretations of the defined classes and properties towards their intended meaning:

- Dereferenceable human readable information should exist about the ontology (e.g. a web page documenting it).
- The ontology should be described by a formal language, like OWL.
- The ontology should be linked to other ontologies.
- The ontology should contain metadata about itself (e.g. authors, modification date, used ontology language, status of the ontology terms, license information, etc.).

OLOUD was developed based on the Uschold and King methodology, which consists of the following steps [24], [25]:

1. Identify the objectives of the ontology development and the intended usage (see Section 2.2); determine the necessary formalization level (see Section 4.3).
2. Specify the ontology by outlining the domain. This includes the identification and the clear textual definition of key concepts and relations (see Section 2.1). Furthermore, setting up identifiers for concepts and relations is necessary.
3. Formalize the terms defined in the specification using a formal language (see Section 4.3).
4. Integrate with existing ontologies. During specification and formalization, it is an important step to research third party ontologies for potential reuse and inclusion (see Section 4.2).
5. Evaluate the fruition of the objectives and the completeness of the ontology based on a predefined (generic and ontology specific) criteria (see Section 5.2).
6. Specify the documentation principles, which should be aligned to type and objective of the ontology (see Section 4.3).

4.2 Integration with other Ontologies

In the process of creating a schema for a Linked Open Dataset it is advisable to reuse as much as possible of the available ontologies or vocabularies. There are quite different vocabulary reuse strategies [26]. The two basic forms are (1) reusing classes and properties from existing vocabularies directly, and (2) establishing links at schema-level. The second case means defining new classes as either sub-classes or equivalent classes and properties as sub-properties or equivalent properties of the classes and properties of the reused ontology. The reuse strategies can be influenced by various factors, like reuse only one (or a few) domain specific vocabulary to provide a clear data structure, or reuse only popular vocabularies to make the data easier to be consumed.

In case of the development of the OLOUD ontology our strategy was the following. First, the necessary concepts (classes and properties) were identified. Then, re-usable vocabularies which could serve to express the defined concepts were chosen according to criteria such as wide usage, OWL 2 compatibility, and regular maintenance. The study for vocabulary selection was detailed in Section 3, and it resulted in the choice to use AIISO as the basis and then use other ontologies to fill in the gaps. Ontologies that are not specific to education are used to represent general concepts and relations, such as resource metadata (Dublin Core), people (FOAF), time (W3C Time and Temporal Aggregates Ontology [27]), events (Event [28]), address (vCARD [29]) and indoor location (iLOC [30]). In case of necessary classes and properties missing from the previous ontology list, new OLOUD terms were introduced. If it was possible the new terms were linked on schema level to the above ontologies with the rdfs:subPropertyOf or rdfs:subClassOf properties.

Integration work posed the problem of fragmentation. In several cases an ontology was needed only for a single property (e.g. address). FOAF and vCard are similar ontologies, but each lacks some important properties, and thus both had to be used to fill in the holes. In the integration process, it was revealed that too many ontologies were needed to express desired goals, and some ontologies were hard to reuse because of the inaccuracy and mistakes in them.

4.3 Ontology Description

Figure 2 represents the overview of the new ontology: the main classes, the highlighted object properties connecting them and the essence of the class hierarchy as well.

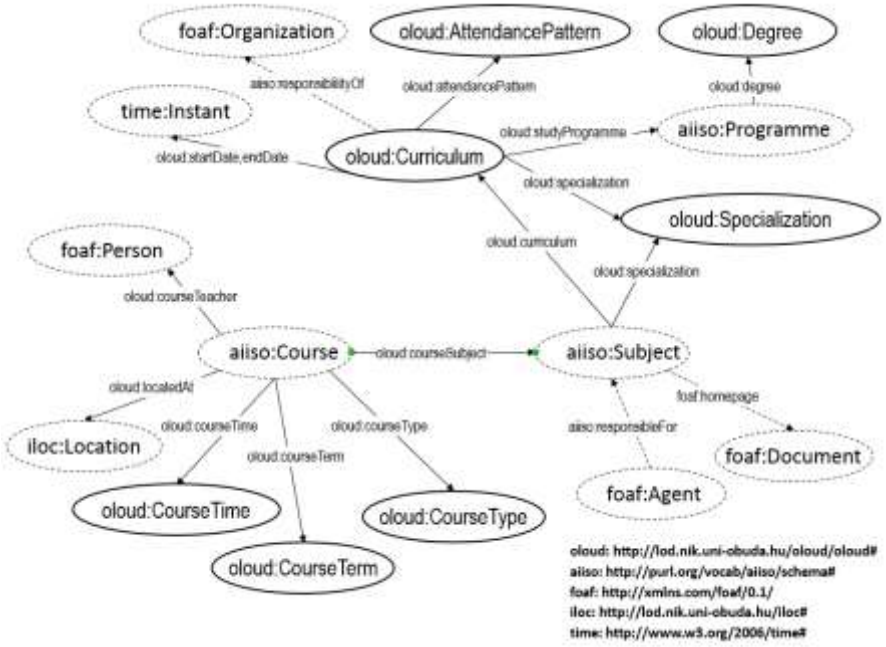


Figure 2
Overview of the main classes and properties in OLOUD

Classes defined in OLOUD are written with *oloud* prefix, classes and properties needed from other ontologies are written with their prefix and marked with a dashed line. The most important classes in OLOUD are *Curriculum, Subject, Course* and *Programme*. Curriculum class is defined as a subclass of the aiiso:KnowledgeGrouping class, Subject, Course and Programme classes are used directly from AIISO. In the following, these classes are described with their direct connections.

Curricula at Hungarian universities contain the list of subjects with their dependencies (i.e. each subject can have various prerequisite subjects). A specific Curriculum entity is connected to the Subject entities with the *curriculum* property. The faculty or department of the university responsible for the Curriculum is determined with the *aiiso:responsibilityOf* property. Each Curriculum has a period of validity determined by the *startDate* and *endDate* properties. The possible Specializations in each Curriculum are determined with the *specialization* property. The Attendance pattern, the Study programme and the Degree of the study are set by the *attendancePattern, studyProgramme* and *degree* properties. The language of the studies according to the Curriculum is given with the *dcterms:language* property.

Subjects are featured by their name, code, number of credits, person and organization responsible for it: *foaf:name, aiiso:code, subjectCredit, aiiso:responsibilityOf*. A Subject entity belongs to a specific Curriculum entity. The connection between a Subject and its Courses is set by the *courseSubject* property defined in OLOUD, since AIISO does not provide any property to connect these concepts. The prerequisite conditions between Subject entities are set by the *subjectRequires* property.

The entities of the Course class are the actual instances of subjects having spatial, temporal and type descriptions, identification number, name and instructor: *locatedAt, courseTime, courseTerm, courseType, aiiso:code, foaf:name* and *courseTeacher*. To describe entities of the Curriculum, Subject and Course classes properly some auxiliary classes were introduced: *StudyProgramme, Degree, AttendancePattern, Specialization, CourseTerm, CourseType*. The aiiso:Programme class is used to represent the Study Programme concept and Specialization is defined as a subclass of the aiiso:Module.

Location class from the iLOC ontology is used to represent all the necessary entities describing indoor locations for Course and Event entities. Courses and events can be assigned to Rooms, and Rooms are connected via a network of POIs (Points of Interest), which can be doors, hallway connections, etc. The offices of lecturers can also be included in the description of campus buildings.

Entities providing temporal description of courses in OLOUD are based on OWL Time and Temporal Aggregates Ontologies. Our objective was to enable SPARQL queries according to date, time and duration and to define course time as recurring events. These objectives can be satisfied with the above ontologies. We suggest using a separate ontology module for the 'time related' concepts. In this module subclasses are defined for classes in OWL Time and Temporal Aggregates Ontologies, facilitating the generation of entities describing recurring events.

The OLOUD ontology consists of two modules: OLOUD-BASE [31] and OLOUD-TIME [32]. The former describes all the 'university related' concepts, uses the prefix *oloud* and namespace *http://lod.nik.uni-obuda.hu/oloud/oloud#*. The latter provides the necessary classes and properties to describe course time

data as recurring events, uses the prefix *otime* and namespace *http://lod.nik.uni-obuda.hu/oloud/otime#*. In OLOUD-BASE there are 7 classes, 16 object properties, 5 data properties and 14 individuals, while in OLOUD-TIME 6 classes are defined at this moment.

OWL 2 RL was chosen as the formal language for OLOUD. The advantage of this ontology approach is that new classifications can be inferred by rules and class restrictions, such as subjects announced for the current semester, subjects meeting the prerequisite criteria in case of a specific student or course announcements having various properties. The OLOUD Ontology was implemented in a self-documenting way. Based on the request MIME type it can be downloaded in different formats including the human consumable HTML output, which is automatically generated from the following properties: rdfs:label, rdfs:comment, rdfs:domain, rdfs:range. The ontology description was implemented as metadata best practice described in [33], by adding the recommended metadata instances and addressing the outlined policies. The Ontology is licensed under the terms of Creative Commons 3.0[7]

5 Ontology Evaluation and Dataset Generation

5.1 Evaluation

The role of the evaluation is to verify the fulfillment of the initial goals and the completeness of the ontology based on the predefined criteria. The evaluation was regularly carried on during the process of ontology development. The development tool – Protégé – was leveraged for validation purposes. The first phase of the evaluation was the definition of a class and the corresponding properties within Protégé. Inconsistency was discovered in several cases as Protégé was not able to create a property for the intended purpose. The root cause of these issues was mostly flaws in the imported third party ontologies. The continuous evaluation also included immediate trials of the new concepts, creating individuals and properties for these new individuals with the help of Protégé. During the ontology evaluation, it was inspected whether the original objectives and expectations were met. We tested the complex use cases with implementing different SPARQL queries answering the questions in section 2 [34]. The OOPS! Ontology Pitfall Scanner was also used to check our OWL [35]. The minor issues the scanner found were fixed.

[7] http://creativecommons.org/licenses/by/3.0/

5.2 Triplification – the Óbuda University Use Case

To evaluate and validate the ontology LOD triples were created based on public data at the Óbuda University. The triples were also tested with Protégé for consistency. On our LOD server[8] we currently serve the dataset using Marmotta[9]. The resulted linked data is organized into six graphs[10] according to the type of the entities (i.e. subjects, courses, events, persons, location and others). At this moment, the database contains about 1000 entities with more than 6000 triples. Example data can be found in [34]. Triples of the different classes were derived from different sources, and the technique – the actual method the triples were created by – depended on the actual source. The location data was created manually based on building layout diagrams of Óbuda University, while the subject and course data was automatically generated with PHP scripts from relational database dumps extracted from the electronic administration system of Óbuda University. The university event descriptions were generated by scraping data from the university webpage.

Expressing 'time related' data of recurring events of course instances was not an easy task. There are multiple ways to model temporal information, but probably the most used ontology for this purpose is the OWL Time Ontology. It provides basic constructs to define and describe points and intervals bounded with a start and endpoint in the temporal space. OWL Time provides two approaches to describe a point of time: either using the xsd:datetime datatype or using the DateTimeDescription class. While the first one offers an easy way to define a point of time by a well-structured string, it lacks some of the features the DateTimeDescription class provides. On the other hand, manually modeling and maintaining DateTimeDescription entities are error prone and tiring because these require at least 7-8 triples in a format that is reusable in a semantic sense.

Temporal Aggregates ontology was used to express temporal information of courses as recurring events (e.g. lectures on every Monday from 8 am until 9.30 am in the 2015 Fall semester). The precise implementation of such information as Linked Data needs the introduction of several additional entities, hence the management of such information is time consuming and error prone. In case the given LOD dataset contains lot of temporal information, manually publishing all the necessary triples would be cumbersome. We used self-unfolding URI scheme for the time entity and an attached template to auto generate the required triples based on the information in the URI. The generation of course time data was implemented as an automatized process described in [36] using SPARQL Construct queries that can be executed in a scheduled manner.

[8] http://lod.nik.uni-obuda.hu/marmotta/
[9] http://marmotta.apache.org
[10] http://lod.nik.uni-obuda.hu/marmotta/core/admin/contexts.html

Different namespaces were used for ontology concepts and for the generated data instances. The T-Box (terminology) of the Ontology is identified with the following URI schema: http://lod.nik.uni-obuda.hu/oloud/oloud#{class or property name} and the A-Box (instance data) is identified with http://lod.nik.uni-obuda.hu/data/{instance_ID} URIs. The structure of the instance_ID for the different classes were chosen according to the nature of the specific class. For example, the subject code (used by the university to refer to the subject) was a proper choice for the instance_ID of the Subject class, because it is unique among the different subjects. For the instances of the Course class the course code (used by the university) had to be complemented with the semester code, because the course code is unique only within a single semester.

Conclusions

The starting point of our work was to implement useful, "smart" services for university students based on linked data. We realized that there are too many ontologies or vocabularies for the domain, and none of them is suitable for our purpose. We created the OLOUD ontology, which amalgamates selected ontologies and fills the missing links between existing concepts.

During the ontology development, we relied on the existing concept definitions and data structures used in the Hungarian landscape such as the Neptun student information system or the Moodle e-learning platform. This was necessary because of the specialties of the Hungarian educational system (probably all national systems have smaller or bigger differences from others), and also because all previous ontologies for education were specific to some goals, and neither of them aimed at a holistic description of the domain.

In Hungary, the structure of the training programs in higher education is unified, the meanings of basic terms like courses, subjects, specializations, curricula are treated uniformly. This is not derived from acts of legislation, but from everyday practice, which is characterized by the overwhelming use of a specific electronic student administration system at universities in Hungary (called Neptun). Thus, most Hungarian universities follow the same data model (i.e. the ones standing behind Neptun) in their workflow.

Although there is a uniformly used electronic student administration system in Hungarian universities, the need for an ontology for university open data still exists. The reason is complex: (1) open university data exist in more sources (not only in Neptun), (2) the availability of public data from Neptun is cumbersome and it is far from the requirements of a 4-star dataset, (3) the reuse and the integration of open data from several data sources is difficult.

There is a need for a common understanding of the basic terms of the educational process. This can help foreign students to find the way in their university studies, and the interoperability between universities in different countries. The OLOUD ontology provides the basis of several ongoing student projects, which either

integrate new datasets for the university or implement new services on top of the OLOUD dataset. In the future, our plan is to collect and transform to linked data all knowledge that is practical for the daily life at Óbuda University. Furthermore, we wish to proceed with various application developments using the generated LOD dataset. For example, a 'curriculum assistant' mobile application helping students to select their courses at the start of the semester might be useful. In the future, the OLOUD ontology can be extended with new features. For example, the need to list information about the ongoing and past research of the university might induce the extension of the OLOUD model.

References

[1] Rohs, M., & Bohn, J. (2003, May) Entry Points into a Smart Campus Environment-Overview of the ETHOC System. In Distributed Computing Systems Workshops, 2003. Proceedings. 23rd International Conference on (pp. 260-266) IEEE

[2] Daga, E., d'Aquin, M., Adamou, A., & Brown, S. (2015) The Open University Linked Data-data. open. ac. uk. Semantic Web Journal. http://www.semantic-web-journal.net/content/open-university-linked-data-dataopenacuk

[3] d'Aquin, M., Brown, S. (2012) Linked Data at the Open University: From Technical Challenges to Organizational Innovation http://www.slideshare.net/mdaquin/linked-data-at-the-open-university-from-technical-challenges-to-organizational-innovation

[4] Ma, Y., Xu, B., Bai, Y., & Li, Z. (2011, December) Building Linked Open University Data: Tsinghua University Open Data as a Showcase. In Joint International Semantic Technology Conference (pp. 385-393) Springer Berlin Heidelberg

[5] The Lucero project. So, what's in linked datasets for education (2012) http://lucero-project.info/lb/2012/04/so-whats-in-linked-datasets-for-education/

[6] d'Aquin, M., Adamou, A., & Dietze, S. (2013, May) Assessing the Educational Linked Data Landscape. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 43-46) ACM

[7] DCMI Metadata Terms (2012) http://purl.org/dc/terms/

[8] Brickley, D., & Miller, L. (2014) FOAF Vocabulary Specification 0.99. Namespace Document 14 January 2014-Paddington Edition

[9] Miles, A., & Bechhofer, S. (2009) SKOS Simple Knowledge Organization System Reference. W3C recommendation, 18, W3C

[10] Hobbs, J. R., & Pan, F. (2006) Time Ontology in OWL. W3C working draft, 27, 133

[11] D'Arcus, B., & Giasson, F. (2009) Bibliographic Ontology Specification. URL: http://bibliontology.com/specification

[12] Styles, R., & Shabir, N. (2008) Academic Institution Internal Structure Ontology (aiiso) http://vocab.org/aiiso/schema

[13] WS-LT, C. E. N. (2008) Metadata for Learning Opportunities (MLO)–Advertising. In Workshop Agreement. CEN. http://www.estandard.no/files/CWA15903-00-2008-Dec.pdf

[14] Barker, P. (2015) A Short Project on Linking Course Data http://blogs.pjjk.net/phil/a-short-project-on-linking-course-data/

[15] Styles, R., Wallace, C., & Moeller, K. (2008) Participation Ontology http://vocab.org/participation/schema

[16] Styles, R., Wallace, C. (2008) Academic Institution Internal Structure Ontology Roles (AIISO Roles) http://vocab.org/aiiso-roles/schema

[17] Kauppinen, T., Trame, J., & Westermann, A. (2012) Teaching Core Vocabulary Specification (TEACH ontology) http://linkedscience.org/teach/ns/

[18] Stubbs, M. (2006). XCRI| eXchanging Course-Related Information. Techn. rep., Manchester Metropolitan University, Aytoun Street, Manchester, M1 3GH

[19] Rodriguez, B., Millard, I. (2006) ReSIST Courseware Ontology http://courseware.rkbexplorer.com/ontologies/courseware

[20] Educational Credit Information Model (ECIM) (2010) ftp://ftp.cen.eu/CEN/Sectors/TCandWorkshops/Workshops/CWA16077.pdf

[21] Börner, K., Conlon, M., Corson-Rikert, J., & Ding, Y. (2012) VIVO: A Semantic Approach to Scholarly Networking and Discovery. Synthesis Lectures on the Semantic Web: Theory and Technology, 7(1) 1-178

[22] Demartini, G., Enchev, I., Gapany, J., & Cudré-Mauroux, P. (2013) The Bowlogna Ontology: Fostering Open Curricula and Agile Knowledge Bases for Europe's Higher Education Landscape. Semantic Web, 4(1) 53-63

[23] Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman II, C. (2014) Five Stars of Linked Data Vocabulary Use. Semantic Web, 5(3) 173-176. http://www.semantic-web-journal.net/content/five-stars-linked-data-vocabulary-use

[24] Uschold, M., and M. King. Towards a Methodology for Building Ontologies. IJCAI'95 Workshop on Basic Ontological Issues in Knowledge Sharing. Diss. Ed. D. Skuce, 1995

[25] Uschold, M. Building Ontologies: Towards a Unified Methodology. Technical Report University of Edinburgh Artifical Intelligence Applications Intiture AIAI TR (1996)

[26] Schaible, J., Gottron, T., & Scherp, A. (2014) Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling. In The Semantic Web: Trends and Challenges (pp. 457-472) Springer International Publishing

[27] Pan, F. (2005) Temporal Aggregates for Web Services on the Semantic Web. IEEE International Conference on Web Services (ICWS'05) DOI=http://dx.doi.org/10.1109/ICWS.2005.118

[28] Raimond, Y., & Abdallah, S. (2007) The Event Ontology. Technical report, 2007, http://motools. sourceforge. net/event

[29] Iannella, R., & McKinney, J. (2013) vCard Ontology for Describing People and Organisations, http://www.w3.org/TR/vcard-rdf/

[30] Szasz, B., Fleiner, R., Micsik, A. & Simon-Nagy, G. (2016) iLOC – An Indoor Ontology. http://lod.nik.uni-obuda.hu/iloc/iloc-20160409.owl

[31] Szasz, B., Fleiner, R., Micsik, A. (2016) OLOUD: Ontology for Linked Open University Data. http://lod.nik.uni-obuda.hu/oloud/oloud-20160609.owl

[32] Szasz, B., Fleiner, R., Micsik, A. (2016) Ontology Extension for Temporal Data. http://lod.nik.uni-obuda.hu/oloud/oloud-time-20160609.owl

[33] Vandenbussche, Pierre-Yves, and Bernard Vatant. "Metadata recommendations for linked open data vocabularies." Version 1 (2011): 2011-12

[34] Szasz, B., Fleiner, R., Micsik, A. (2016) OLOUD – Ontology for Linked Open University Data. http://lod.nik.uni-obuda.hu/oloud/index.html

[35] Poveda-Villalón, M., Suárez-Figueroa, M. C., & Gómez- Pérez, A. (2012) Validating Ontologies with OOPS!. In Knowledge Engineering and Knowledge Management (pp. 267-281). Springer Berlin Heidelberg

[36] Szasz, B., Fleiner, R. and Micsik, A. (2016) Linked Data Enrichment with Self-Unfolding URIs. In 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI 2016) pp. 305-309

Dynamic Resource Allocation in Cloud Computing

Seyedmajid Mousavi[1], Amir Mosavi[2,3,4], Annamria R. Várkonyi-Kóczy[2,5], Gabor Fazekas[1]

[1]Faculty of Informatics, University of Debrecen, Kassai Str. 26, 4028 Debrecen, Hungary, {majid.mousavi & fazekas.gabor}@inf.unideb.hu

[2]Institute of Automation, Kandó Kálmán Faculty of Electrical Engineering, Óbuda University, Bécsi út 94-96, 1431 Budapest, Hungary, amir.mosavi@kvk.uni-obuda.hu, varkonyi-koczy@uni-obuda.hu

[3]Norwegian University of Science and Technology, Department of Computer Science, 7491 Trondheim, Norway

[4]Institute of Structural Mechanics, Bauhaus Universität-Weimar, Marienstraße 15, 99423 Weimar, Germany

[5]Department of Mathematics and Informatics, J. Selye University, Elektrarenska cesta 2, 945 01 Komarno, Slovakia

Abstract: Utilizing dynamic resource allocation for load balancing is considered as an important optimization process in cloud computing. In order to achieve maximum resource efficiency and scalability in a speedy manner this process is concerned with multiple objectives for an effective distribution of loads among virtual machines. In this realm, exploring new algorithms, as well as development of novel algorithms, is highly desired for technological advancement and continued progress in resource allocation application in cloud computing. Accordingly, this paper explores the application of two relatively new optimization algorithms and further proposes a hybrid algorithm for load balancing which can contribute well in maximizing the throughput of the cloud provider's network. The proposed algorithm is a hybrid of teaching-learning-based optimization algorithm (TLBO) and grey wolves optimization algorithm (GW). The hybrid algorithm performs more efficiently than utilizing every single one of these algorithms. Furthermore, it well balances the priorities and effectively considers load balancing based on time, cost, and avoidance of local optimum traps, which consequently leads to minimal amount of waiting time. To evaluate the effectiveness of the proposed algorithm, a comparison with the TLBO and GW algorithms is conducted and the experimental results are presented.

1 Introduction

A cloud is created from numerous physical machines. Each physical machine runs multiple virtual machines which are presented to the end-users, or so-called clients, as the computing resources. The architecture of virtual machines is based on physical computers with similar functionality. In cloud computing, a virtual machine is a guest program with software resources which works like a real physical computer [1]. Yet, high workload on virtual machines is one of the challenges of cloud computing in the allocation of virtual machines. The task, requested by a client, has to wait to be allocated to the work and the resources needed. This strategy is independent of the executive priority of the tasks. However, the client who owns the task may offer larger value for it to try to raise his/her priority and eventually may succeed in taking control over the resources needed. Users can consume services based on the service level agreement that defines their needs of quality of service (QoS) parameters [2]. Yet, the multipurpose nature of the scheduling in the cloud computing environment has made it extremely difficult to manage. Therefore, scheduling has to create a compromise between service quality costs to come up with a suitable service which belongs to a multi-objective optimization problems family [3]. Several methods are available for a multi-objective scheduling problem [4]. The current methods of allocation of resources, such as FIFO [1] and Round-Robin [5] which are used in the cloud, do unfair allocation regardless of priority between tasks.

Resource allocation in the cloud environment is utilized to achieve customer satisfaction with minimal processing time. Reducing the fees of leasing resources in addition to ensuring quality of service and improving throughput for trust and satisfaction of the service provider is considered as another objective. In dynamic scheduling, the basic idea is the request allocation at the time of implementation of programs. In addition to the cost estimation, in the static method, dynamic scheduling consists of two other main sections of system state estimation and decision-making [6]. To recap, clients are interested in having their tasks completed in the shortest possible time and at the minimum cost which cloud servers should receive. On the other hand, the cloud providers are interested to maximize the use of their resources and also to increase their profits. Obviously these two objectives are in conflict with each other and often they are not satisfied with the traditional methods of resource allocation and scheduling mechanisms available [7]. Yet the goal is to direct the resource allocation to be performed in a way that is acceptable to both the users and the suppliers.

Resource allocation is a technique that ensures the allocation to virtual machines when multiple applications need different resources of CPU and input/output memory [2]. In cloud computing there are two technical restrictions. Firstly, the capacity of the machines is physically limited; secondly, priorities for the implementation of the tasks should be in harmony with maximizing the efficiency of resources. Ultimately, the waiting time and the completion time are to be

reduced, in order to decrease the cost of system implementation. Classical methods for achieving a fully optimized solution are very time-consuming and in some cases are impossible. Traditional approximate methods [5] are reported inconclusive and inaccurate for solving optimization problems and are often trapped in local optimum.

Virtual machines in distributed systems have different usage conditions including; the cost of utilizing them and also different processing power. The tasks initiated by users may also have a different amount of information. In addition, to assign any task on any machine, a preparation time between tasks is also considered. This time-delay, which varies in different resources, is considered negligible in this study. The most important problem in this process is the order process and how the placement of tasks on resources is conducted. In fact by increasing the productivity of resources, the response time can be reduced and, simultaneously, can improve the total cost for resource utilization and load balancing. The load-balancing index is calculated based on target variables. The target variables include:

- The time of completing the latest task among virtual machines
- The average cost paid by the user for use of the resources
- Efficiency caused by the impact of load balancing based on completion time and cost of doing them.

To conduct this study in the realm of cloud computing system development, a number of assumptions are made with the following characteristics. In these assumptions, the resource and virtual machine are considered as one entity.

- Tasks are independent.
- Distributed environment is heterogeneous and dynamic.
- All tasks must be done.
- Each task is performed only by one virtual machine.
- Everything is done exactly once.
- Each virtual machine has different and specified processing speed.
- Each virtual machine has a special price that must be paid for the use of it in time.

Traditional approximation and resource allocation methods (see, e.g. [7]) due to the multi-objective and dynamic nature of the problem and also difficulties in dealing with local optimum need advancement and major improvement. Consequently, the purpose of this paper is set to address the research gap in cloud computing. To do so a hybrid approximation algorithm for resource allocation is proposed. In this study the performance of two algorithms i.e. teaching-learning-based optimization (TLBO) [8] and grey wolves optimization (GW) [9], in comparison with the proposed algorithm, are discussed. These two algorithms are currently used as approximation algorithms for establishing load balancing, based on time and cost between resources and efficiency. The balance is established between three target assessment variables for evaluating the proposed approximation algorithm.

2 Related Works

For resource allocation in distributed scheduling, Xu et al. [5] present a non-dominated sorting genetic algorithm-based multi-objective method (NSGA-II) [10]. They aimed at minimizing the time and cost in load balancing using resources to achieve Pareto optimal front. They used self-adaptive crowding distance (SCD) to overcome the crowding distance. In addition, in their proposed method, a mutation operator is included in the traditional algorithm of NSGA-II to avoid premature convergence. In this method, the strategy that the algorithm uses for improving the efficiency and performance of the intersections, has not been effective, as the solutions are trapped in local optimum.

Salimi et al. [14] introduced a multi-objective tasks' scheduling using fuzzy systems and standard NSGA-II algorithms for distributed computing systems in [6]. The authors aimed at minimizing implementation time and costs while increasing the productivity of resources. Their study is associated with the load balancing in the distributed system. They use the indirect method and fuzzy systems and do implementation of the third objective function to solve this problem. However dealing with three objectives has not been efficiently done in their work. In [7], Cheng provides an optimized hierarchical resource allocation algorithm for workflows using a general heuristic algorithm. In this model, the main objective is the coordination between the tasks and duties assigned to the service. The purpose is to service in accordance with the operational needs to perform properly the tasks and observe the priority between them. This model accomplishes workflow tasks scheduling aimed at load balancing by(?) dividing the tasks to different levels. Further, mapping and allocation of each level of tasks to resources is directed according to the processing power.

Gomathi and Karthikey [8] introduce a method for assigning tasks in a distributed environment using Hybrid Particle Swarm Optimization algorithm (HybPSO) [11]. HybPSO is used to meet the user needs and increase the amount of load balancing with productivity. The goal is to minimize the task completion time among processors and create load balancing. This method assures that each task is assigned to exactly one processor. In this method, each solution is shown as a particle in the population; each particle is a vector with n dimension which is defined for scheduling n independent tasks.

In [9], authors introduce a heuristic method based on particle swarm algorithm [12] for tasks' scheduling on distributed environment resources. Their model considers the computational cost and the cost of data transfer. Their proposed algorithm optimizes dynamic mapping tasks to resources using classical particle swarm optimization algorithm and ultimately balances the system loads. This optimization method is composed of two components. One of them is the scheduling operations task and the other one is particle swarm algorithm (PSA) to obtain an optimal mix of the tasks to resources' mapping.

Table1 presents a summary of the related works done in the field of tasks' scheduling. This table includes the objectives of tasks' scheduling, the algorithms used in these methods, the simulation environment of the algorithms, and the year in which they were developed. Table1 does not include the GW and education-based learning algorithms and/or any variations of these algorithms.

Table1
Summary of the works done in the field of resources' allocation

Author	Evolutionary algorithm	Environment	Targets	Year	Simulation tool
Xue et al [13]	multi-target genetic	Cloud	• Reduce the longest termination time among resources • Reduce the resources cost • Load balancing	2014	Matlab
Salimi et al [14]	multi-target genetic	Grid	• Reduce the longest termination time among resources • Reduce the resources cost • Load balancing	2014	GridSim
Cheng [15]	genetic	Cloud	• Reduce the longest termination time among resources • Load balancing	2012	Java environment
Gomathi & Karthikey [8]	swarm optimization	Cloud	• Reduce the longest termination time among resources • Load balancing	2013	Java environment
Pandey et al [17]	swarm optimization	Cloud	• Reduce costs associated with load balancing	2010	Amazon EC2
Wu et al [18]	swarm optimization	Grid	• Reduce the longest termination time among resources • Reduce the workflow time • Load balancing	2012	Ad-hoc VC++ toolkit
Izakian et al [19]	swarm optimization	Cloud	• Reduce the longest termination time among resources • Reduce the workflow time • Load balancing	2010	Java environment
Banerjee et al [20]	Ant colony	Cloud	• Reduce the longest termination time among resources • Load balancing	2009	Simulated cloud
Mousavi & Fazekas [21]	Ant colony	Cloud	• Reduce the longest termination time among resources • Reduce the workflow time • Load balancing	2016	CloudSim
Ludwig & Moallem [22]	Ant colony	Grid	• Reduce the longest termination time among resources • Load balancing	2011	GridSim
Babu & Krishna [23]	Bee colony	Cloud	• Reduce the longest termination time among resources • Load balancing	2013	CloudSim
Zhao [24]	swarm optimization	Cloud	• Reduce the longest termination time among resources • Reduce the resources cost	2015	CloudSim
Abdullah & Othman [25]	Simulated Annealing	Cloud	• Reduce the longest termination time among resources • Load balancing	2014	CloudSim

The literature review shows that traditional methods which are used for optimization, may be definitive and accurate, yet they are often trapped in local optimum. In fact, due to the dynamic nature of distributed environment and

heterogeneous resources, in such a system, the scheduling process must be done automatically and very quickly. That is why the scheduling process is recognized as an NP-complete problem [26]. Traditional approaches are not dynamic and suitable to solve such a scheduling problem. These approaches contain a large search space; facing a large number of possible solutions and a tedious process to find the optimal solution. There is currently no efficient method available to solve these problems. In such circumstances, the traditional approach has been set to find a fully optimized solution instead of finding the semi-optimal solution, but in a shorter time. In this context, IT professionals are focused on exploratory methods. Therefore, metaheuristic algorithms which have a global overview, as they ensure convergence to solution and do not fall into the trap in local optimum, are of importance. Consequently, the GW algorithm is chosen for this purpose. In addition, the TLBO algorithm is used in a hybrid form with GW to improve local optimization and increase accuracy.

3 Proposed Method

Methodology is based on bonding the algorithms of TLBO and GW. With such hybridization, it is aimed at speeding up the process while maintaining the improvement of local optimization and increasing the accuracy. In the following, the problem is described. Further TLBO and GW algorithms are introduced as the primary solutions to the described problem. The proposed methodology then emerges from bonding of two algorithms.

3.1 Description of the Problem

There is a distributed network in a cloud environment with resource systems $S_1,S_2,S_3,\ldots,S_n$. The resources are ready to serve in the distributed network for various nodes. Different jobs are sent for the source systems by nodes. The overall goal of this system is that, an agreed scheduling on the resources' allocation is to be obtained to perform the jobs. Consequently, with the resources allocation, the load balancing will be increased [27]. Here the scheduler is responsible to allocate one or more jobs to artificial machines in a distributed system [28]. In other words, the agreement on job scheduling is done by the scheduler. The scheduler provides a scheduling for resource allocation [29].

Several jobs are allocated and processed in parallel with each other at time - t - in the distributed system. The number of variables T_k is permutation between jobs and resources, this variable is called P, and its value is calculated as follows:

$$P = n^m \quad (n \text{ is number of tasks and } m \text{ is the number of sources}) \tag{1}$$

As it is described in Figure 1 each node includes several jobs. Each job requires a series of specific resources. The problem can be introduced as *Job* → $j_1,j_2,\ldots,j_n$ *and Resource* → $R_1,R_2,\ \ldots\ ,R_m$.

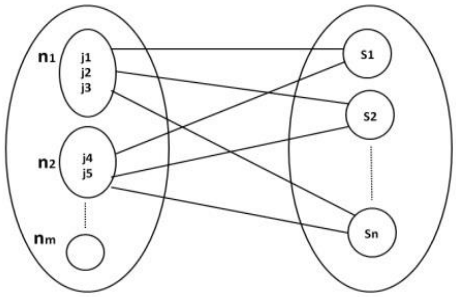


Figure 1
Resource allocation in a distributed environment

If in the particular example, the resources $R_1,R_2,\ \ldots\ ,R_m$ have the same capacity and processing power and $j_1,j_2,\ldots,j_n$ all need 1% of the processing. The advanced model can be defined in a form describing what jobs in which resources should be used in order to achieve the maximum load balancing, average response time, and minimum cost. For the exact solution of the problem, all possible allocation modes must be calculated and the best mode chosen. Due to the large number of modes (exponential), the problem is an example of set packing problems, which is of *NP*-complete type.
Optimization function is defined for resource i and job j. y_i is the number of resources (package). The objective function and mathematical programming model that should be optimized are as follows:

$$Min\ B = a*\left(1- L_{(y\,j)}\right)+ b*C_{(y\,j)} + c*T_{(y\,j)} \tag{2}$$

S.t.

$$\sum_{i=1}^{n} w_i x_{i\,j} \le K y_j, \forall j \quad , \quad \sum_{j=1}^{n} x_{i\,j} \le b_j, \forall j \qquad x_{i\,j}, y_i = 0,1 \quad \forall i,j$$

Where :

$$x_j = \begin{cases} 1 & job\ j\ is\ used \\ 0 & job\ j\ isnot\ used \end{cases} \quad , \quad y_j = \begin{cases} 1 & resource\ j\ is\ used \\ 0 & resource\ j\ isnot\ used \end{cases}$$

$x_{i\,j}$ represents that job j is assigned to resource i. C is the maximum capacity for each resource. w_i represents the amount of job i that is covered by the resource. The aim is to find the minimum number of virtual machines - Y_j - that minimize the objective functions. The values of L, C, and T (load balancing, cost, and response time) are considered based on the number of virtual resources, Y_j, where

a, b, c are variable based on cloud system. The variable of X_{ij} demonstrates that the i^{th} job is in j^{th} virtual machines, and if its value is equal to 0, it means that there is not any resource in j^{th} virtual machine and if its value is equal to 1, it means that there is enough resource to allocate the j^{th} virtual machine. Every job has the capacity of W_i. The first limitation indicates that total capacity of jobs can be placed at the maximum - K - available resources. The second limitation shows the maximum capacity of each virtual resource. b_j is the capacity of each virtual resource.

3.2 Grey Wolf Algorithm

Mirjalili et al. [30] introduce GW for solving engineering problems. GW is a new optimization algorithm inspired by behavior of grey wolves' hunting and their role hierarchies. The GW algorithm is benchmarked on 29 well-known test functions and the results on the unimodal functions show the superior exploitation of GW. Capability of exploration of new solutions in GW algorithm is confirmed by the results of multimodal function. The hierarchical structure and social behavior of wolves during the hunting process is modeled in the form of mathematical models and is used to design an optimization algorithm. The GW optimization algorithm emulates the hierarchical leadership and hunting mechanism of Grey Wolves in nature. Four types of GWs are considered as hierarchical structure in the social behavior of wolves for simulating, such as alpha, beta, delta, and omega. Furthermore, the three main steps of hunting - searching for prey, encircling prey, and attacking prey are implemented (Figure 2).

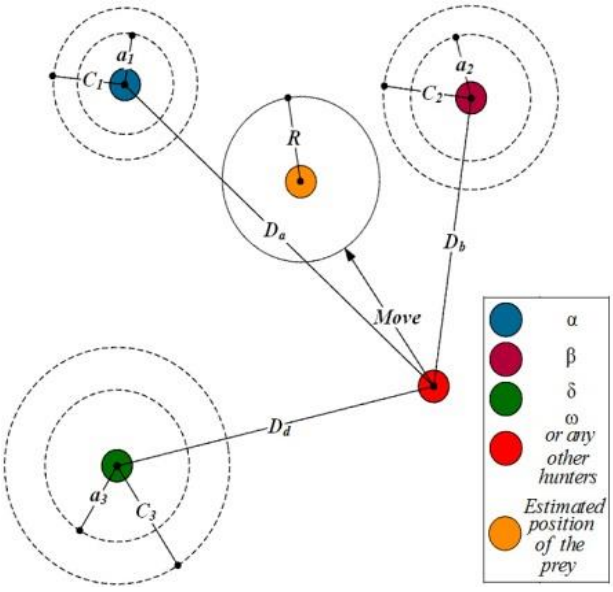


Figure 2
Grey wolves' motion in haunting [30]

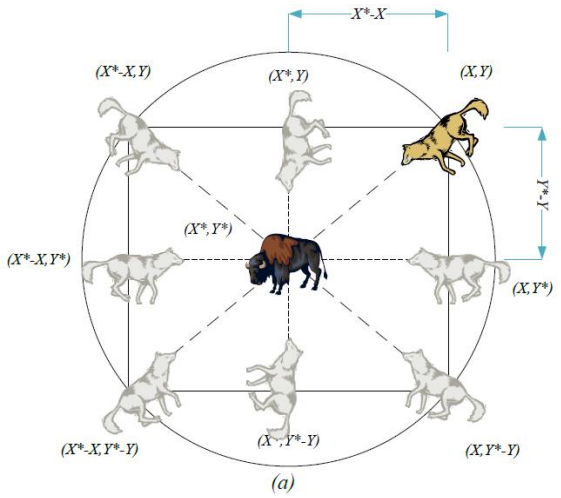


Figure 3
Updating wolves' position [30]

The Wolves' leader is called alpha and is primarily responsible for the prey. The second level of wolves, which helps the leader, is called beta. The third level of wolves is called delta and is designed to support alpha and beta. The lowest level is called Omega [31]. In general, the algorithm steps can be summarized as follows:

- The fitness of all solution levels are computed and three top solutions are selected as Alpha, Beta, and Delta (Figure 2) until the end of the algorithm. In fact, Alpha is the best fitness of solution. After Alpha, the Beta and Delta are the best solutions respectively.

- In each iteration, the three top solutions e.g. Alpha, Beta, and Delta have the ability to estimate the prey position, conducting it in each iteration using the following equations [30]:

$$\vec{D}_\alpha = \left|\vec{C}_1.\vec{X}_\alpha - \vec{X}\right|, \quad \vec{D}_\beta = \left|\vec{C}_2.\vec{X}_\beta - \vec{X}\right|, \quad \vec{D}_\delta = \left|\vec{C}_3.\vec{X}_\delta - \vec{X}\right| \tag{3}$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad where \quad \begin{cases} \vec{X}_1 = \vec{X}_\alpha - \vec{A}_1.(\vec{D}_\alpha) \\ \vec{X}_2 = \vec{X}_\alpha - \vec{A}_2.(\vec{D}_\alpha) \\ \vec{X}_3 = \vec{X}_\alpha - \vec{A}_3.(\vec{D}_\alpha) \end{cases} \tag{4}$$

First, the wolves put a ring around the prey, where *Xp*, is the hunting position vector. *A* and *C* are hunting vector coefficients. *X* is the wolves' positions and t stands for the stage of each iteration. *D* indicates the behavior of putting the ring around the hunt [30]. In each iteration, after determining the positions of Alpha, Beta, and Delta, the other solutions are updated in compliance with them. Hunting

information is defined by Alpha, Beta, and Delta. And the rest update their x positions accordingly. In each iteration, vector and consequently vectors b and c are updated. At the end of an iteration, Alpha wolf position is considered as the optimal point. This value is A. A value of A is an option value which is between (-2a, 2a). The absolute value of A is less than 1, so when the wolves are at the A distance from the prey, attack happens. At a distance of more than one, it is still(??) necessary that the wolves must converge toward each other [4]. In the GW algorithm, some main parameters like initial population size, vector coefficients, number of iterations, and the number of wolf levels are to be determined [32]. Then, the cost function of optimization which is minimized in this study is introduced. Afterward, the initial population is formed randomly and the fitness function is introduced. Then, in a loop on a regular basis, the position of the wolves' level is determined and the fitness function is calculated, and, using them, the new positions are calculated again. Iteration of this loop is specified according to the initial parameters.

3.3 Teaching-Learning-based Algorithm

Teaching-learning-based optimization algorithm (TLBO) [33] provides a novel approach to explore a problem space to find the optimal settings and parameters to satisfy the problem's objectives. The algorithm was introduced by Rao et al [33]. Similar to other evolutionary optimization techniques, TLBO algorithm is an algorithm derived from nature and works based on a teacher's teaching in a classroom. A teacher in the classroom, by expressing material, plays an important role in student learning and, if the teaching is effective, students learn the material better. In addition to the teacher factor, review of lessons by students would lead to better learning. This algorithm uses a total population of solutions to achieve the overall solution. A teacher tries to increase the level of students' knowledge by teaching and repeating the materials. Therefore the students can achieve a good score. In fact, a good teacher makes students closer to the level of his/her knowledge. The teacher is the most knowledgeable person of the class that shares his/her knowledge with the students. So the best solution (the best student of the class population) in the same iteration can act as a teacher. It should be considered that the students acquire knowledge based on the quality of teaching by the teacher and students' status (the average of class scores). This idea is the basis of Teaching-Learning-Based Optimization algorithm for solving optimization problems. The algorithm operates in two phases. The first phase is the teacher who shares his/her knowledge with students and the second phase is the review of courses by students in the same class. At the first stage, a teacher tries to improve scores of a class. In Figure 4, the Gaussian distribution function is used and the average scores acquired by students in the classroom is shown as M. M Parameter indicates the degree of the teacher's success in the classroom. In this figure, M_1 and M_2, respectively, show average scores of two separate classrooms with the same students.

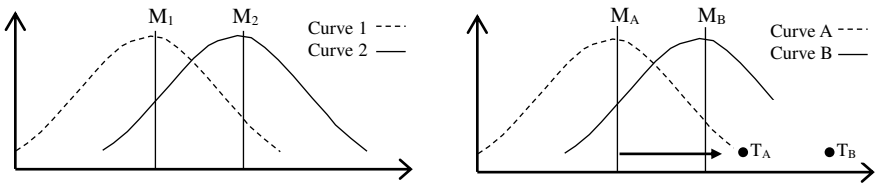


Figure 4
Average scores of students who were in classroom [33]

As it is shown in figure 4, the second teacher, with average scores of M2, has acted better than the first teacher with average score of M_1. T_A is the first grade teacher, who, with the best-case average scores of the first grade, M_A, moves to the T_A. It means that the academic level of students is approaching that of their teacher or equal with him/her. This creates a new population of the classroom which has shown an average of M_B and T_B. In fact, the students do not reach the knowledge level of the teachers, just close to it, which also depends on the level of classroom ability. Gaussian probability function is defined as follow [34, 36]:

$$f(X) = \frac{1}{s\sqrt{2p}} e^{\frac{-(x-m)^2}{2s^2}} \qquad (5)$$

In this formula, μ is the average score of students, which is shown as M_1 and M_2 in Figure 4.

3.4 Proposed Algorithm

Given more convergence power in the global optimality, the GW algorithm is used as a base algorithm in the proposed algorithm. This algorithm can also perform multi-objective optimization. The steps are as follows. In the initial state, a series of random numbers as the initial population are considered with uniform distribution and a basic solution is considered for the problem. Coefficients a, b, and c are initialized. Each solution is known as a wolf. In another word, each wolf is considered as a solution to the problem. These solutions or wolves have an answer. Wolves are divided into three categories; alpha, beta, and gamma. Yet on the basis of the fitness function, one of them gives a better answer to the fitness function. Then, the solution enters the main loop where after a few iterations the best solution for the fitness function is discovered. Based on the equations of the GW algorithm, the wolves' position is updated.

According to the first class of wolves, the new positions are fitted. Later on, more values for the probability of solution are considered. Correspondingly, the values of beta and gamma classes, the new positions of wolves, and their classifications can be obtained. If a suitable solution is found in the new classification, the algorithm is to be improved further. The best solution between the wolves is considered as the initial solution (initial population) for the teaching and learning algorithm. Further, the problem of the teaching and learning algorithm is solved

and the solution is considered as initial population to start again. In this stage the GW algorithm is implemented. If there is no improvement in GW algorithm, according to the teaching and learning, it tries to find a better solution. If the solution is trapped in local optimum, an algorithm based on teaching and learning can introduce the new area of space based on training, which may improve solution. Since the accuracy of GW algorithm in the local behaviour is high, after each stage the position of wolves is determined. This position can be improved by learning and training algorithms, and GW algorithm is implemented again. This process increases the accuracy of GW algorithm. It should be considered that in the GW algorithm, every wolf represents a solution in the solution space. The best and the most successful solution will be chosen among them at any stage according to the position of other wolves. The best solution of the GW is defined as the initial solution for teaching and learning. After learning and training, its output is implemented as the initial solution for the next iteration in the GW algorithm. The proposed algorithm is presented in table 2:

Table 2
Pseudo code of the proposed algorithm

```
Initialize the grey wolf population Xi=(i=1,2,…,n)
Initialize a,b and c
Calculate the fitness of each search agent
X1=the best Search gent
X2=the second best Search Agent
X3=the third best Search agent
While t<Max number of iterations)
   For each search agent
    Update the position of the current search agent by equation
    End for
Calculate the fitness of all search agents
Update X1,X2,X3
t=t+1
If not improve solution
 Begin
   sol_wolf=Solution_grey_wolf
       Initialize sol_wolf for initialize_solution for TLBO
       Sol_TLBO=Do TLBO with Initialize Population with sol_wolf
       Intialize the grey wolf population Xi= Sol_TLBO, Initialize a,b and c
       Calculate the fitness of each search agent
X1=the best Search agent, X2=the second best Search agent, X3=the third
best Search agent
 end
end while
return X1
```

The main advantage of this algorithm is that if there was no improvement in grey wolf algorithm, according to the teaching-learning process, we try to find a better solution. If the problem is stuck in local optimum, teaching-learning process can introduce the new area of space based on training phase, which may improve the solution. Because of the accuracy of grey wolf algorithm in the local behavior (defect of the grey wolf algorithm), after each iteration, the position of wolves is updated. These positions can be improved by the teaching-learning algorithm, and then grey wolf algorithm is repeated again. This process increases the accuracy of grey wolf algorithm. Figure 5 illustrates flow diagram of proposed algorithm.

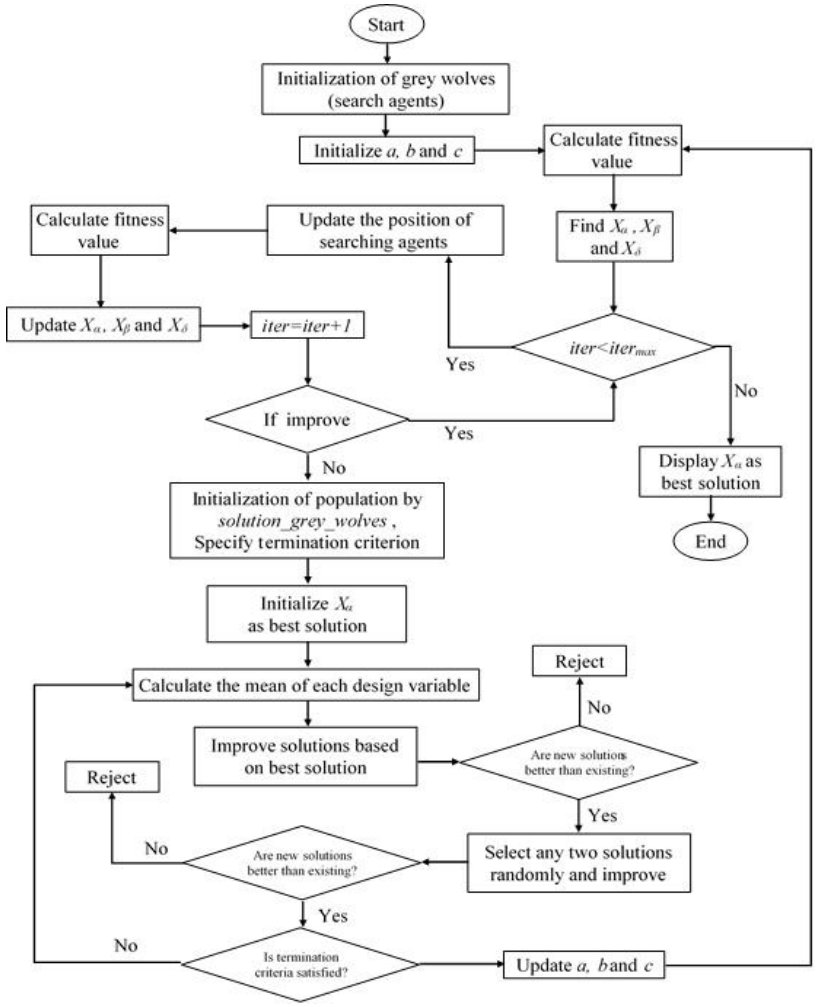


Figure 5
Flow diagram of proposed algorithm

4 Simulation

According to the assessment index, load imbalance and the number of resources allocated are compared with each other. In this problem, Matlab is used to simulate the proposed method and set a packing problem which is considered as a model of resource allocation in cloud computing. Packages in the set packing are considered as requests that are processed in the cloud virtual servers.

Amount of efficiency of each resource - $Ri_{efficiency}$ - is equivalent to what percentage of resource has been used compared to the total resource [28, 30, 35]. Formally, the coefficient of variation of resource efficiency is called lack of load balancing. This variable indicates to what extent, there is a deviation of productivity. According to statistical indicators, if this variable is zero, it means that absolutely all resources are used. This variable is equal to the quotient of productivity of standard deviation in resources when there are a number of resources. When the variable is close to zero, load balancing is done better. Standard deviation is:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Ri_{efficiency} - mean\,R_i)^2} \qquad (6)$$

Load balancing factor *(Flb)* and load imbalance factor *(NFlb)* are introduced using the following equations:

$$NFlb = \frac{s.R_i}{n} \quad i = 1...n \qquad (7)$$

$$Flb = \frac{C - s}{n} \qquad (8)$$

Where C is the capacity of each resource or virtual machine.

Teaching-Learning simulation parameters have been set as follows:

```
maxIt = 500;      % Maximum Number of Iterations
nPop = 250;       %for each test must be updated Population Size
```

maxIt is the number of iterations for TLBO algorithm. *nPop* is the proportion of the number of packages in an appliance packaging problem.

Grey Wolf simulation parameters have been set as follows:

```
max_iteration=500
Dim = number of your variables
a=0.0354,   b=38.3055,    c=1243.531
```

Coefficients a, b, c, have been selected according to the resource. *Dim* is the number of packets in the packaging of appliance problem and *max_iteration* is the number of iterations for the algorithm.

5 Experimental Results

There are 60 packages in the dataset binpack5 with the capacity of 100. Each package has different value. The best way to obtain optimal solution is dividing the sum of the values of packages by the capacity of 100. The optimal solution of the allocation is 20 boxes. Our proposed method has achieved an approximate solution of 23. The solution has acted better than the GW and TLBO algorithms alone, as is shown in Table 3.

Table 3

Comparison of allocation with the proposed algorithm in dataset binpack5

Dataset	GW	TLBO	Hybrid method	Optimal result
First part of binpck5	24	26	23	20

In Table 4, allocation in dataset boxes binpack5 is shown. Each box is shown as a bin. Each bin indicates the packages with weights. For example bin1 contains packages weighing 49, 5, 47 and 4. Other boxes are allocated different packages in the same manner.

Table 4

Allocated boxes for proposed method

bin1: 49.5,47.4,	bin7: 41,39.5,	bin13: 35,35,29.9,	bin19: 27.2,26.9,26.9,
bin2: 47.3,47.2,	bin8: 37.2,37,25.5,	bin14: 34.7,32,31.5,	bin20: 26.8,26.2,26.1,
bin3: 46.6,45,	bin9: 36.6,36.6,26.3,	bin15: 30.7,30.3,29.8,	bin21: 25.9,25.8,25.4,
bin4: 44.5,44.4,	bin10: 36.6,36.3,27.1,	bin16: 29.8,28.8,28.7,	bin22: 25.2,25.2,25.2,
bin5: 43.9,43,	bin11: 36.1,35.7,27.5,	bin17: 28.3,27.5,27.4,	bin23: 25.1,
bin6: 41.9,41.4,	bin12 35.5,35.1,29.2,	bin18: 27.3,27.3,27.2,	

In the second experiment, a set of binpacks was used. Hence, according to data dispersion and increased capacity of boxes, the problem became more difficult, but the results demonstrate that the proposed method is desirable. Figure 6 indicates this process. A comparison between the differences of optimum solution show a high performance.

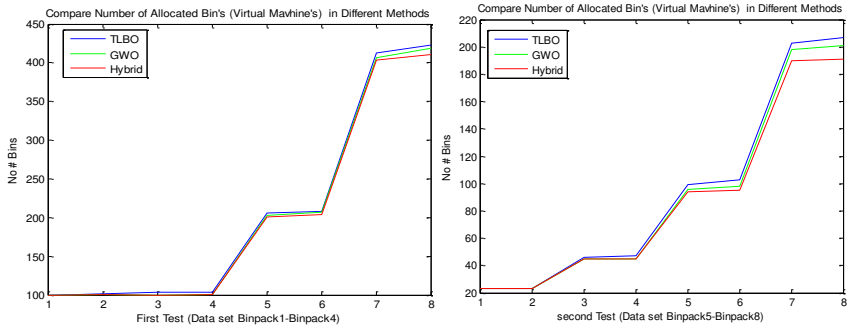


Figure 6

Comparison of the hybrid method with other methods to achieve optimal solution

According to a recent article [36], CGA-CGT and HI-HB methods are providing the best solution to set packing. The proposed hybrid method indicates a better performance than the other two methods (Table 5).

Table 5
Comparison of the proposed method with HI_BP and CGA-CGT

Experiment	Dataset	Number of Job	Resource capacity	HI_BP	CGA-CGT	Hybrid	Optimal solution
First	Binpack1-U250_00	250	150	100	100	100	99
First	Binpack1-U250_01	250	150	101	101	101	100
First	Binpack2-U250_00	250	150	100	100	100	99
First	Binpack2-U250_01	250	150	101	101	101	100
First	Binpack3-U500_00	500	150	204	201	201	198
First	Binpack3-U500_01	500	150	204	204	204	201
First	Binpack4-U1000_0	1000	150	404	404	403	399
First	Binpack4-u1000_1	1000	150	414	413	411	406
Second	Binpack5-T60_00	60	100	23	23	23	20
Second	Binpack5-T60_01	60	100	23	23	23	20
Second	Binpack6-T120_00	120	100	45	45	45	40
Second	Binpack6-T120_01	120	100	47	45	45	40
Second	Binpack7-T249_00	249	100	96	94	94	83
Second	Binpack7-T249_01	249	100	101	97	95	83
Second	Binpack8-T501_00	501	100	202	194	190	167
Second	Binpack8-T501_01	501	100	204	199	191	167

Figure 7 demonstrates the difference between the proposed approximate solution and the optimal solution. The difference is within acceptable limits. When data is increased, the proposed hybrid method has a better performance than the other two methods.

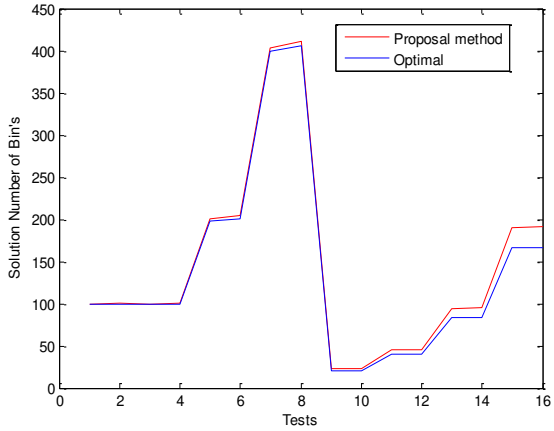


Figure 7
Solution's difference between the proposed method and the optimal solution

Nflb index indicates lack of load balancing. Decreasing this index demonstrates the increasing of load balancing in the cloud system. Increased load imbalance further indicates that maximum resource capacity is used. This means that the capacity of the resources is low. Therefore by increasing data, load balancing is performed better in the proposed algorithm. Figure 8 shows the load imbalance

between the methods in the first and second experimental set. The load imbalance index acts in a reverse manner compared with the load balancing index. Increasing of data will increase the performance of load balancing. The training learning method, due to the structure of incorrect understanding of training in the experiments, doesn't have a good load balancing with more data. The proposed method shows a good performance with increased data in load balancing. The load balancing can be calculated by the lack of load balancing. Load balancing can indicate appropriate allocation. The second experiment in Table 6 confirms the fifth to eighth data set, and then related charts are presented in Figure 8.

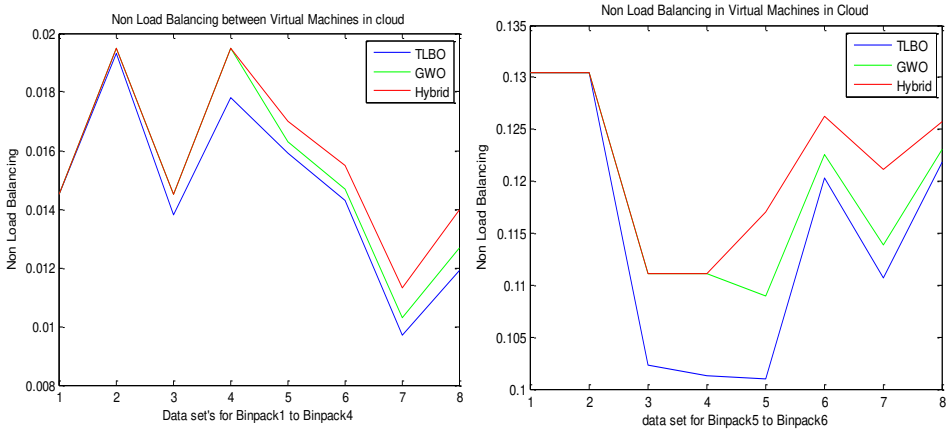


Figure 8

Load imbalance in the first test set (left) and load imbalance in the second test set (right)

Table 6

Comparison of the load imbalance between the proposed algorithm and other algorithms

Experiment	Dataset	Number of Job	Resource capacity	Nflb-TLBO	Nflb-GWO	Nflb-Proposed method
First	Binpack1-U250_00	250	150	0.0145	0.0145	0.0145
First	Binpack1-U250_01	250	150	0.0193	0.0195	0.0195
First	Binpack2-U250_00	250	150	0.0138	0.0145	0.0145
First	Binpack2-U250_01	250	150	0.0178	0.0195	0.0195
First	Binpack3-U500_00	500	150	0.0159	0.0163	0.0170
First	Binpack3-U500_01	500	150	0.0143	0.0147	0.0155
First	Binpack4-U1000_00	1000	150	0.0097	0.0103	0.0113
First	Binpack4-u1000_01	1000	150	0.0119	0.0127	0.140
Second	Binpack5-T60_00	60	100	0.1304	0.1304	0.1304
Second	Binpack5-T60_01	60	100	0.1304	0.1304	0.1304
Second	Binpack6-T120_00	120	100	0.1023	0.1111	0.1111
Second	Binpack6-T120_01	120	100	0.1013	0.1111	0.1111
Second	Binpack7-T249_00	249	100	0.1010	0.1090	0.1170
Second	Binpack7-T249_01	249	100	0.1203	0.1226	0.1263
Second	Binpack8-T501_00	501	100	0.1107	0.1139	0.1211
Second	Binpack8-T501_01	501	100	0.1219	0.1231	0.1257

Variable of relative changes percentage in comparison with the best answer can demonstrate the accuracy of the algorithm. Therefore, we calculate Robust Parameter Design *(RPD)* parameter, which is normalized change percentage against the best answer where $f_{heuristic}$ is metaheuristic value and $f_{optimal}$ is optimal value. The *RPD* equation is:

$$RPD = 100 \times \frac{f_{heuristic} - f_{optimal}}{f_{optimal}} \tag{9}$$

Figure 9 indicates improvements of the relative changes' percentage. With increasing jobs this performance is observed to be stable in the proposed method.

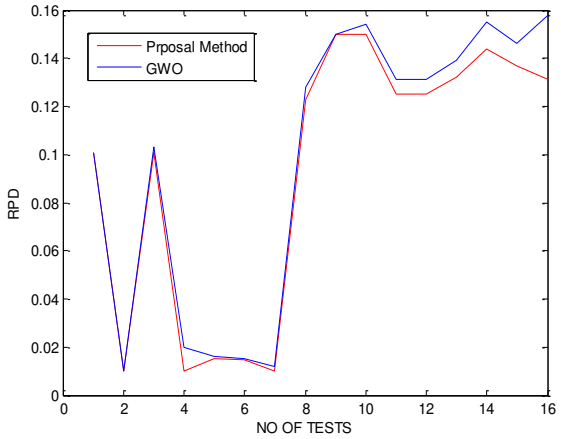


Figure 9

Relative changes' percentage of the optimal solution

This criterion shows the performance of the algorithm with increase of the jobs and indicates that the proposed method has appropriate performance with increasing data. Table 7 demonstrates relative changes' percentage in proposed algorithm. The average of relative changes' percentage is increased in an appropriate way in experimental data and has a better flow than GW and TLBO algorithms.

Table 7

RPD index in performance of proposed method

Experiment	Dataset	Number of Job	Resource capacity	Hybrid	Optimal solution	RPD
First	Binpack1-U250_00	250	150	100	99	1/99
First	Binpack1-U250_01	250	150	101	100	1/100
First	Binpack2-U250_00	250	150	100	99	1/99
First	Binpack2-U250_01	250	150	101	100	1/100
First	Binpack3-U500_00	500	150	201	198	3/198
First	Binpack3-U500_01	500	150	204	201	3/201
First	Binpack4-U1000_00	1000	150	403	399	4/399
First	Binpack4-u1000_01	1000	150	411	406	5/406

Second	Binpack5-T60_00	60	100	23	20	3/20
Second	Binpack5-T60_01	60	100	23	20	3/20
Second	Binpack6-T120_00	120	100	45	40	5/40
Second	Binpack6-T120_01	120	100	45	40	5/40
Second	Binpack7-T249_00	249	100	94	83	11/83
Second	Binpack8-T501_00	501	100	190	167	23/167
Second	Binpack8-T501_01	501	100	191	167	22/167

Conclusions

The performance of two relatively new optimization algorithms, i.e., TLBO and GW, along with a hybrid form of these two algorithms in dynamic resource allocation is described. To evaluate the performance of the proposed hybrid algorithm, a comparison with the TLBO and GW algorithms is conducted and the experimental results presented. It is reported that the proposed hybrid algorithm performs more efficiently than utilizing only one of these algorithms. It is further concluded that the main problem in the resource allocation of cloud scheduler is the lack of convergence in the optimal solution. Optimization of objective functions for the resource allocation at any time is one of the main problems in dynamic resource allocation. The evaluation of experimental results indicate that the proposed hybrid approach in high-volume data for resource allocation in cloud scheduler has better performance than the other two methods.

Acknowledgment

This work has been sponsored by Hungarian National Scientific Fund under contract OTKA 105846 and Research & Development Operational Program for the project “Modernization and Improvement of Technical Infrastructure for Research & Development of J. Selye University in the Fields of Nanotechnology and Intelligent Space”, ITMS 26210120042, co-funded by the European Regional Development Fund.

References

[1] J. Yao and H. Ju-Hou: Load Balancing Strategy of Cloud Computing Based on Artificial Bee Algorithm, Information Management, Vol. 51, 2012, pp. 185-189

[2] S. Zhao, X. Lu, and X. Li: Quality of Service-based Particle Swarm Optimization Scheduling in cloud Computing, Networks, Vol. 12, 2015, pp. 235-242

[3] A. Mosavi and A. Vaezipour: Reactive Search Optimization; Application to Multiobjective Optimization Problems, Applied Mathematics, Vol. 3, 2012, pp. 1572-1582

[4] S. Zhang and Y. Zhou: Grey Wolf Optimizer Based on Powell Local Optimization Method for Clustering Analysis, Discrete Dynamics in Nature and Society, Vol. 3, 2015

[5] U. Ayesta, M. Erausquin, E. Ferreira, and P. Jacko: Optimal Dynamic Resource Allocation to Prevent Defaults, Operations Research, Vol. 6, 2016, pp. 451-456

[6] A. Abur, and A. G. Exposito: Power System State Estimation: Theory and Implementation. CRC press, 2004

[7] A. Khetan, V. Bhushan, and S. Gupta: A Novel Survey on Load Balancing in Cloud Computing, Journal of Engineering & Technology, Vol. 6, 2013, pp.1-9

[8] B. Gomathi and K. Karthikeyan: Task Scheduling Algorithm Based on Hybrid Particle Swarm Optimization in Cloud Computing Environment, Journal of Theoretical and Applied Information Technology, Vol. 89, 2013, pp. 33-38

[9] E. Emary, Hossam M. Zawbaa, Crina Grosan, and Abul Ella Hassenian: Feature Subset Selection Approach by Grey-Wolf Optimization. Industrial Advancement, Springer International Publishing, Vol. 63. 2015 pp. 1-13

[10] A. Mosavi: Multiple Criteria Decision-Making Preprocessing Using Datamining Tools. International Journal of Computer Science Issues, Vol. 7, 2010, pp. 26-34

[11] P-Y. Yin, S-S. Yu, P-.P Wang, and Y-T. Wang: A Hybrid Particle Swarm Optimization Algorithm for Optimal Task Assignment in Distributed Systems. Computer Standards & Interfaces, Vol. 28, 2006, pp. 441-450

[12] R. C. Eberhart and J. Kennedy: A New Optimizer Using Particle Swarm Theory. Micro Machine and Human Science, Vol. 1, 1995, pp. 39-43

[13] S. Xue: An Improved Algorithm Based on NSGA-II for Cloud PDTs Scheduling, Journal of Software, Vol. 6, 2014, pp. 443-450

[14] R. Salimi, H. Motameni, and H. Omranpour: Task Scheduling Using NSGA II with Fuzzy Adaptive Operators for Computational Grids, Journal of Parallel and Distributed Computing, Vol. 74, 2014, pp. 2333-2350

[15] B. f Cheng: Hierarchical Cloud Service Workflow Scheduling Optimization Schema using Heuristic Generic Algorithm, Telecommunications, Vol. 15, 2012, pp. 92-95

[16] R. V. Rao, V. J. Savsani, and D. P. Vakharia: Teaching–Learning-based Optimization: an Optimization Method for Continuous Non-Linear Large Scale Problems, Journal of Information Sciences, Vol. 7, 2012, pp.1-15

[17] S. Pandey, L. Wu, S. Guru, and R. Buyya: A Particle Swarm Optimization-based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments, Information Networking and Applications, 2010, pp. 400-407

[18] L. Wu: A Revised Discrete Particle Swarm Optimization for Cloud Workflow Scheduling, Faculty of Information and Communication Technologies Swinburne University of Technology, Melbourne, Australia, 2012, pp. 1-5

[19] H. Izakian, B. Ladani, A. Abraham, and V. Snasel: A Discrete Particle Swarm Optimization Approach for Grid Job Scheduling, International Journal of Innovative Computing, Information and Control, Vol. 16 2014, pp. 4219-4252

[20] S. Banerjee, I. Mukherjee, and P. Mahanti: Cloud Computing Initiative using Modified ant Colony Framework, World Academy of Science, Engineering and Technology, Vol. 12, 2009, pp. 200-203

[21] SM. Mousavi and G. Fazekas: A Novel Algorithm for Load Balancing using HBA and ACO in Cloud Computing Environment, International Journal of Computer Science and Information Security, Vol. 16, 2016, pp. 48-52

[22] S. Ludwig and A. Moallem: Swarm Intelligence Approaches for Grid Load Balancing. J Grid Computing, Vol. 8, 2013, pp. 279-301

[23] LD. Dhinesh Babu and PV. Krishna: Honey Bee behavior Inspired Load Balancing of Tasks in Cloud Computing Environments, Science Direct, Applied Soft Computing, Vol. 13, 2013, pp. 2292-2303

[24] S. Zhao, X. Lu, and X. Li: Quality of Service-based Particle Swarm Optimization in Cloud Computing, Computer Engineering and Networks, 2015, pp. 235-242

[25] M. Abdullah and M. Othman: Simullated Annealing Approach to Cost-based Multi-Quality of Service Job Scheduling in Cloud Computing Environment, American Journal of Applied Sciences, Vol. 18, 2014, pp. 872-877

[26] MR. Garey, DS. Johnson, and L. Stockmeyer: Some Simplified NP-Complete Graph Problems. Theoretical Computer Science, Vol. 1, 1976, pp. 237-267

[27] Z. Bo, G. Ji, and A. Jieqing: Cloud Loading Balance Algorithm, Proceedings of IEEE 2nd International Conference on Information Science and Engineering, China, 2010, pp. 5001-5004

[28] M. Grabowski, C. Rizzo, and T. Graig: Data Challenges in Dynamic, Large-Scale Resource Allocation in Remote Regions, Safety Science, Vol. 34, 2013, pp.76-86

[29] D. Bertsimas, S. Gupta, and G. Lulli: Dynamic Resource Allocation: A Flexible and Tractable Modeling Framework, European Journal of Operational Research, Vol. 23, 2014, pp. 14-26

[30] S. A. Mirjalili, S. M. Mirjalili, and A. Lewis: Grey Wolf Optimizer, Advances in Engineering Software, Vol. 69, 2014, pp. 46-61

[31] M. H. Suleiman, Z. Mustafa, and M. R. Mohmed: Grey Wolf optimizer for Solving Economic Dispatch Problem with Valve-Loading Effects, APRN Journal of Engineering and Applied Sciences, Vol. 2, 2015, pp. 1619-1628

[32] C. Selvaraj: A Survey on Application of Bio-Inspired Algorithms, International Journal of Computer Science, Vol. 11, 2014, pp. 366-370

[33] R. V. Rao, V. J. Savsani, and D. P. Vakharia: Teaching–Learning-based Optimization: A Novel Method for Constrained Mechanical Design Optimization Problems, Computer-aided Design, Vol. 23, 2011, pp. 303-315

[34] N. Malarvizhi, V. R. Uthariaraj: Hierarchical Load Balancing Scheme for Computational Intensive Jobs in Grid Computing Environment, First International Conference in Advanced Computing, 2009, pp. 97-104

[35] A Várkonyi-Kóczy, A. R.: A Load Balancing Algorithm for Resource Allocation in Cloud Computing. In Recent Advances in Technology Research and Education, Vol. 660, 2017, pp. 289-299

[36] B. Yagoubi and Y. Slimani: Task Load Balancing Strategy for Grid Computing, Journal of Computer Science, Vol. 12, 2007, pp. 186-194

Robust Control of Single-Mast Stacker Cranes

Sándor Hajdu*, Péter Gáspár**

* Department of Mechanical Engineering, University of Debrecen
Ótemető u. 2-4, H-4028 Debrecen, Hungary, e-mail: hajdusandor@eng.unideb.hu

** Systems and Control Laboratory, Institute for Computer Science and Control, Hungarian Academy of Sciences
Kende u. 13-17, H-1111 Budapest, Hungary, e-mail: gaspar.peter@sztaki.mta.hu

Abstract: The stacker cranes in automated storage/retrieval systems (AS/RS) of warehouses often have very high dynamical loads. These dynamical loads may generate harmful mast vibrations in the frame structure of stacker cranes which can reduce the stability and positioning accuracy of these machines. The aim of this paper is to develop controller design methods which have proper reference signal tracking and mast-vibration attenuation properties. First, the dynamic modeling of single-mast stacker cranes by means of multibody modeling approach is summarized. Based on this modeling technique a $\mathcal{H}_\infty$ and a robust control design method are proposed for achieving the appointed purposes. The analyses of the controlled systems are carried out by time domain simulations.

Keywords: stacker crane; modeling uncertainties; robust control; multi-body model

1 Introduction

One of the most important materials handling machines in automated storage/retrieval systems (AS/RS) of warehouses is the stacker crane. These machineries realize the storage/retrieval operation into/from the rack structure of warehouse. The stacker crane frame structures are often subjected to very high dynamical loads due to the inertial forces in the acceleration and braking phases of moving cycles. These dynamical loads generate undesirable, low frequency and high amplitude mast-vibrations in the frame structure. These high amplitude mast-vibrations reduce the positioning accuracy and the stability of the stacker cranes. In extreme cases, the massive oscillations may damage the frame structure of these machines.

Because of the above-mentioned reasons, the harmful mast-oscillations must be reduced. This can be performed for example by means of controlling the traveling motion (towards the aisle of the warehouse) of the stacker crane. In this paper, some controller designing techniques (based on $\mathcal{H}_\infty$ approach) are developed

which can reduce the harmful mast-vibrations. In [2] and [4] authors introduce motion control techniques to attenuate the mast-vibration of stacker cranes. However, in these works the effect of lifted load position and magnitude on the dynamical properties of the structure is neglected during the controller design. The main purpose of this work is to develop a controller design method which takes varying lifted load position and magnitude into account and at the same time having proper reference signal tracking and mast-vibration attenuation properties.

In this paper, the so-called multibody modeling technique is applied to the dynamic modeling of single-mast stacker cranes. For more information about this modeling approach see the following books: [1, 3]. Some further examples of dynamic modeling of stacker cranes by multibody models can be found in [17] and [18]. Concerning the mathematical models of electric drive systems see, e.g. [13-16]

For control design purposes, $\mathcal{H}_\infty$ [5-8] and robust control [19] approaches are applied. The presented control design methods in this paper are based on the results of our previous work, see in [10]. The main contribution of this paper is the robust $\mathcal{H}_\infty$ position controller which can handle the model uncertainties due to varying lifted load conditions. First, the concept of $\mathcal{H}_\infty$ control is presented by means of a standard $\mathcal{H}_\infty$ control method (the so-called mixed-sensitivity loop shaping). After that, a more sophisticated method is developed for the robust $\mathcal{H}_\infty$ position control of stacker cranes. The method for the determination of weighting function parameters in robust control design is also proposed.

The structure of the paper is as follows. In Section 2 the background of dynamic modeling of single-mast stacker cranes is summarized. The state space representation of the model is also introduced. In Section 3 the mixed-sensitivity loop shaping control method for the positioning control of stacker cranes is presented. Section 4 proposes a robust control method which aim is the fast and vibration-free positioning control of stacker cranes in the presence of model uncertainties.

2 Modeling Aspects of Single-Mast Stacker Cranes

In this section the modeling considerations necessary to the control design are briefly summarized. Before the control design a suitable dynamic model must be generated, as mentioned in the introduction, for this purpose the multibody modeling approach is chosen. In this multibody model the continuous sections of the mast are approximated by rigid elements having lumped masses, its center points (i.e. nodes). These elements are interconnected by elastic hinges. More details of this multibody model, as well as the main parameters of investigated stacker crane, are presented in [9-11].

One of the most important steps of dynamic modeling is choosing the generalized coordinates for the governing equations of motion. Several equivalent choices of generalized coordinates exist, and with the proper selection the generation process of motion equations can be simplified. In this paper, the q_i vertical displacements of each node are applied for generalized coordinates. Let us denote the degrees of freedom (DOF) of the model by n_d. This way the generalized coordinate vector of the model can be expressed as: $q = \begin{bmatrix} q_1 & q_2 & \dots & q_{n_d} \end{bmatrix}^T$. Here q_1 is the vertical position of the bottom frame and q_{n_d} is the vertical position of mast-tip.

The detailed derivation of the dynamic equations for the before-mentioned multibody model and generalized coordinates can be found in [9, 11]. The matrix equation of motion can be generated in the following form (with the mass matrix M, the damping matrix K and the stiffness matrix S respectively):

$$M\ddot{q} + K\dot{q} + Sq = F\,. \tag{1}$$

In Equation (1) F is the vector of external excitation forces. In this work a single-input system is investigated, where the input signal of the model is the external force F_t acting on the bottom frame. Thus, in vector F only the first coordinate is nonzero.

The controller synthesis methods applied in this paper use the state space representation of the model, thus the matrix equation of motion (1) must be transformed into state space form. As mentioned before the input signal of the model is the external force acting in the direction of q_1 generalized coordinate. In the following steps of this work the model is applied in the synthesis of controller which realizes the positioning control of single-mast stacker cranes with reduced mast-vibrations. Therefore, two kinds of outputs are required in the state space presentation of the dynamic model. The first one is used to describe and investigate the mast-vibrations. This output is the inclination of mast, i.e. the position difference between the undermost point of mast and mast-tip. The output is denoted by z. The second output is the so-called measured output. This output is applied for the position control of the stacker crane and can be equal to the horizontal position or velocity of stacker crane. In this work the horizontal position of stacker crane, i.e. the first generalized coordinate is applied as measured output. The output is denoted by y.

The state space representation of the dynamic model is generated in the following form:

$$\dot{x} = Ax + B_1 d + B_2 u\,, \tag{2a}$$

$$z = C_1 x + D_{11} d + D_{12} u\,, \tag{2b}$$

$$y = C_2 x + D_{21} d\,, \tag{2c}$$

where x, u, y, d, z are the state vector, control input, measured output, disturbance input and performance output vectors, respectively. The matrices $A \in \Re^{n \times n}$, $B \in \Re^{n \times m}$, $C \in \Re^{p \times n}$, $D \in \Re^{p \times m}$ are the so-called system matrices. Here n is known as the order of the system and m, p are the number of all input and output variables of the system respectively. As can be seen in equation (2) the matrices B, C, D are usually partitioned according to the kinds of input and output signals.

In the actual case of this stacker crane dynamic model the disturbance input does not exist. The state vector with the above-mentioned generalized coordinate vector is defined as:

$$x = \begin{bmatrix} \dot{q} & q \end{bmatrix}^T . \tag{3}$$

Using this definition, the state space representation of the investigated multibody model can be generated - taking notice of the above-mentioned definition of input and output signals - with the following considerations. Extending the equation (1) with the identity $M\dot{x} - M\dot{x} = 0$ the system matrices A and B can be computed by means of expressing the derivative of state vector from the extended system:

$$A = \begin{bmatrix} -M^{-1}K & -M^{-1}S \\ I & 0 \end{bmatrix}, \tag{4a}$$

$$B = \begin{bmatrix} -M^{-1}F \\ 0 \end{bmatrix}, \tag{4b}$$

where 0 is a zero matrix/vector and I is an identity matrix with the corresponding size.

As mentioned before the investigated model must fulfill the requirements of controller synthesis techniques. The multibody model introduced in this section has almost one hundred degrees of freedom, thus the order of state space representation of this model is near two hundred. This complicated, high order model is not suitable for controller design since it causes numerical problems in controller synthesis methods of modern control theory, e.g. $\mathcal{H}_\infty$ method. A smaller size model also can speed up the simulation process during the design validation phase. Because of the above-mentioned reasons our investigated model is reduced with a suitable model order reduction method, see [10].

3 Mixed-Sensitivity Loop Shaping Control of the Stacker Crane

A frequently applied and well-known control design approach in $\mathcal{H}_\infty$ control theory is the so-called loop shaping procedure presented in [12]. In this section, the $\mathcal{H}_\infty$ control design method of stacker cranes using the mixed-sensitivity loop shaping approach is presented. The aim of this section is to analyze the influence of several loop shaping weighting strategies on the main control objectives (i.e. the reference signal tracking and the mast vibration attenuation). This may help later to generate more complex and advanced weighting strategies in order to improve the control performances. For the purpose of control design the nominal model of stacker crane - with the lifted load in the highest position - is used, thus in this section the nominal performances are investigated without model uncertainties.

The augmented plant for mixed-sensitivity loop shaping is presented in Figure 1. As shown in Figure 1 the weighting functions W_1, W_2 and W_3 penalize the error signal, control signal, and output signal respectively. The weighting functions W_1, W_2 and W_3 must be proper and stable transfer functions. In the actual control design $W_2 = 0$, while W_1 and W_3 have the following general form:

$$W_1 = \frac{s/M_1 + \omega_1}{s + \omega_1 A_1},\ W_3 = \frac{s/M_3 + \omega_3}{s + \omega_3 A_3}. \tag{5}$$

This way the low-frequency asymptote (A_i), the high-frequency asymptote (M_i) as well as the bandwidth (ω_i) of weighting functions can be adjusted. These parameters have a fundamental role in the loop shaping procedure.

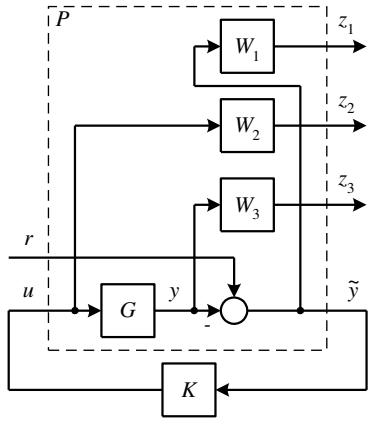


Figure 1
Augmented plant for mixed-sensitivity loop shaping

The disturbance input and the controlled output of the augmented plant are defined as: $\tilde{w} = r$ and $\tilde{z} = \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix}^T$ respectively. The measured output is equal to: $\tilde{y} = r - y$. Using the above-mentioned definitions of input- and output signals it is easy to verify that the closed loop transfer function matrix $T_{\tilde{z}\tilde{w}}$ from $\tilde{w}$ to $\tilde{z}$ can be expressed as:

$$T_{\tilde{z}\tilde{w}} = \begin{bmatrix} W_1 S \\ W_2 KS \\ W_3 T \end{bmatrix}, \tag{6}$$

where $S = (I + PK)^{-1}$ and $T = PK(I + PK)^{-1}$ are the sensitivity function and complementary sensitivity function of closed loop system respectively.

As mentioned before in the actual design cases the weighting function W_2 is equal to zero, thus the performance objective of $\mathcal{H}_\infty$ control design implies the following conditions:

$$|W_1 S| \le \gamma, \; |W_3 T| \le \gamma. \tag{7}$$

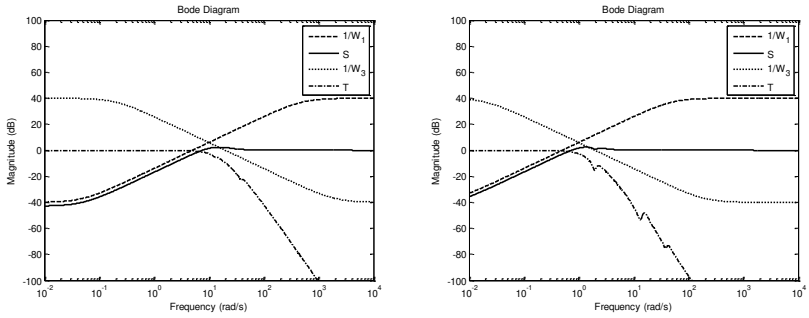
Therefore, the weighting functions W_1 and W_3 determine the shapes of sensitivity function S and complementary sensitivity function T. Typically, the inverse of W_1 is chosen to be small inside the desired control bandwidth to achieve proper performance (e.g. disturbance attenuation or tracking), and the inverse of W_3 is chosen to be small outside the control bandwidth, which helps to ensure proper stability margin (i.e. robustness).

Table 1
Parameters of loop shaping

Case #1	Case #2
$A_1 = 100$	$A_1 = 100$
$M_1 = 0.01$	$M_1 = 0.01$
$\omega_1 = 5.0 \; rad/s$	$\omega_1 = 0.5 \; rad/s$
$A_3 = 0.01$	$A_3 = 0.01$
$M_3 = 100$	$M_3 = 100$
$\omega_3 = 20 \; rad/s$	$\omega_3 = 2.0 \; rad/s$
$\gamma = 0.9004$	$\gamma = 0.9012$

By the variation of the parameters of these weighting functions two kinds of controllers are designed. In these controller design cases the desired control bandwidth is adjusted to 1 rad/s and 10 rad/s respectively. The parameters of performance weighting functions according to the above-mentioned design cases are summarized in Table 1.

The calculations of designed controllers can be carried out, e.g. by means of the solution method presented in [7]. The achieved performance levels for each design cases are also presented in Table1.

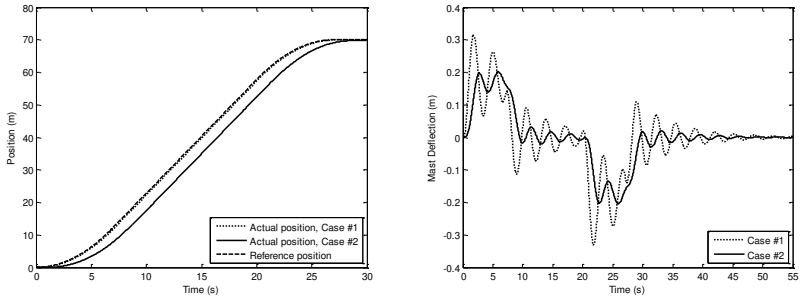


(a) Case #1 (b) Case #2

Figure 2
Performance objectives for loop shaping

The performance objectives for the closed-loop system in both design cases can be analyzed by means of Figure 2. As shown in the figure by means of the weighting function W_1 the sensitivity function is shaped so that its gain is below -40 dB in the low-frequency range. This ensures a low (practically under 1%) steady-state tracking error. The minimum control bandwidth is adjusted by the 0 dB crossover frequency of weighting function W_1, while the upper limit of control bandwidth is given by the 0 dB crossover frequency of W_3.

The simulation results, i.e. diagrams of stacker crane position and mast deflection are shown in Figure 3. During simulations the position signal of a general stacker crane moving cycle is used as the reference signal. In the first session of moving cycle the stacker crane has constant 0.5 m/s^2 desired acceleration. In the second session the desired velocity is 3.5 m/s and the deceleration value of the third session is -0.5 m/s^2. Distance covered of the moving cycle is 70 m while the total cycle time is 27 seconds.



(a) Stacker crane position (b) Mast deflection

Figure 3
Simulation results of loop shaping

Analyzing the simulations above it can be concluded that the reference signal tracking and the vibration attenuation properties can be adjusted by means of the proposed method. However, better performances can be achieved by means of more advanced weighting strategies. Another interesting observation about the simulation results is that the magnitude of mast vibrations is inversely proportional to the control error. Thus, a trade-off between mast vibration attenuation and control error can be determined. Additionally, the modeling uncertainties also must be taken into consideration in the control design method.

4 Robust Control Design for the Stacker Crane

The aim of this section is the presentation of a robust controller design method which can handle the uncertainties in the dynamic model and at the same time have proper reference signal tracking and mast vibration attenuation properties. For applying the $\mathcal{H}_\infty$ robust control approach first the control objectives must be formulated. In this section, the essential requirements for the closed-loop system (i.e. the proper reference signal tracking property and the mast-vibration attenuation) are defined, a more sophisticated way, by means of advanced weighting strategies in the generalized plant. Similar to the loop shaping case here the reference signal of investigated model is also the horizontal position demand of the stacker crane. The augmented plant for robust control design is shown in Figure 4. Since in this augmented plant both output signals of the stacker crane dynamic model are used, the vector-valued signals are denoted by thick lines. This way the diagram of the augmented plant can be simplified.

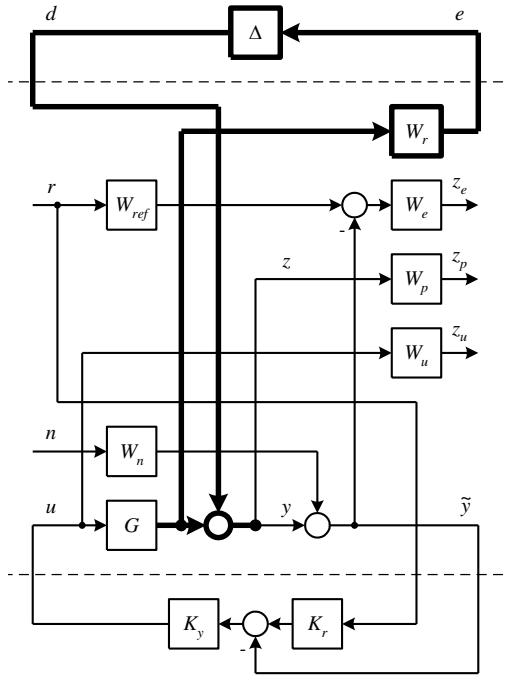


Figure 4
Augmented plant for robust control

As shown in Figure 4 here the controller K is partitioned into a feedback part K_y and a pre-filter part K_r. This controller structure is connected to the second output of the stacker crane model (i.e. the position output y). The aim of this structure is to provide for proper reference signal tracking properties in the positioning control of stacker crane.

The purpose of the transfer function W_{ref} is to represent the desired behavior of the closed loop system. It is usually a second-order transfer function with free parameters ω_r and ζ, i.e. $W_{ref} = \dfrac{\omega_r^2}{s^2 + 2\zeta\omega_r s + \omega_r^2}$.

By means of the free parameters of W_{ref} the bandwidth and damping of the ideal closed-loop transfer function can be adjusted. The difference between W_{ref} and the actual closed-loop transfer function is penalized by the transfer function W_e. The value of this penalty function should be large in the frequency range where small errors are desired and are small where larger errors can be tolerated. In most cases, the more accurate model is required in the low-frequency range thus W_e is a low pass filter.

The aim of the weighting function W_p is to penalize the harmful mast vibrations. Therefore, this weighting function is connected to the first output of the stacker crane model (i.e. the mast-inclination output z). Since penalizing the final, steady-state value of mast inclination (which depends on the acceleration of stacker crane motion) is unnecessary, the W_p transfer function is a high pass filter.

Some further performance specifications are also added to the control design augmented plant. In the high-frequency range the control input is limited by using the performance weighting function W_u, as well as the purpose of the weighting function W_n is to reflect the sensor noises. Finally, the weighing function matrix W_r reflects the amount of uncertainty and it can be determined by the procedure mentioned in [9] and [10].

The transfer function matrix of the generalized plant can be expressed as follows.

$$\begin{bmatrix} e \\ \hline z_e \\ z_p \\ z_u \\ \hline r \\ \tilde{y} \end{bmatrix} = \left[\begin{array}{c|ccc} 0 & 0 & 0 & W_r G \\ \hline -W_e & W_e W_{ref} & -W_e W_n & -W_e G_{yu} \\ W_p & 0 & 0 & W_p G_{zu} \\ 0 & 0 & 0 & W_u \\ \hline 0 & I & 0 & 0 \\ I & 0 & W_n & G_{yu} \end{array}\right] \begin{bmatrix} d \\ \hline r \\ n \\ \hline u \end{bmatrix}. \tag{8}$$

Due to the two degrees of freedom controller structure the corresponding feedback relation is: $u = K\begin{bmatrix} r & \tilde{y} \end{bmatrix}^T$.

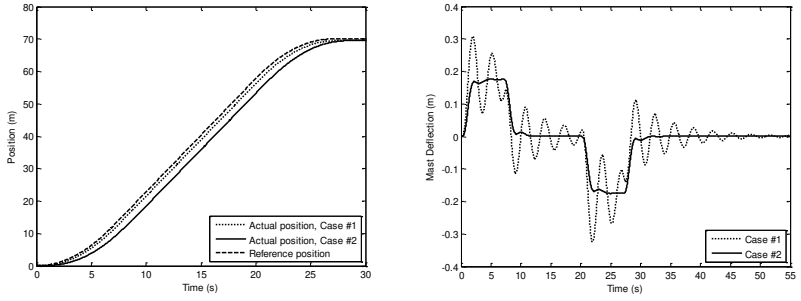
During the actual investigations the lifted load position varies in position range from 41 to 44 m which generates the model uncertainty. This helps to keep the amount of uncertainty sufficiently small. The nominal model of the model set that generates the uncertainty is the model with lifted load position in the middle of position range, i.e. 42.5 m.

In order to analyze the proposed robust control design method two kinds of control design cases are generated. The first weighting strategy (Case #1) focuses on the adequate reference signal tracking rather than mast-vibration attenuation. While in the second strategy (Case #2) the mast-vibrations are penalized more. In the control design cases for the model matching function W_{ref} the following parameter values are applied: $\omega_r = 8 \ \ rad/s$, $\zeta = 1$. The performance weighting functions according to the above-mentioned design cases are summarized in Table 2. As shown in the table the weighting functions of control input and sensor noises are permanent for both design cases.

Table 2
Weighting functions for robust control design

Case #1	Case #2
$W_e = 100\dfrac{1+0.1s}{1+10s}$	$W_e = 80\dfrac{1+s}{1+100s}$
$W_p = 0.1\dfrac{1+0.1s}{1+0.001s}$	$W_p = 0.1\dfrac{1+s}{1+0.01s}$
$W_u = 4\cdot 10^{-6}\dfrac{1+0.01s}{1+0.001s}$	$W_u = 4\cdot 10^{-6}\dfrac{1+0.01s}{1+0.001s}$
$W_n = 0.01\dfrac{1+0.1s}{1+0.01s}$	$W_n = 0.01\dfrac{1+0.1s}{1+0.01s}$

The investigation of the properties of designed controllers can be carried out by means of time-domain analysis. In this simulation, as a reference signal, the same position signal is used as in the case of loop shaping control design, see in Section 3. The simulation results (i.e. the stacker crane position and mast deflection functions) are shown in Figure 5.



(a) Stacker crane position (b) Mast deflection

Figure 5
Simulation results of robust control

For the comparison of above mentioned time-domain results the following quantities are defined. The rate of mast vibrations is measured by the overshoot of mast deflection signal in the acceleration phase of movement:

$$\sigma_a = \frac{\max_t \left| z(t) - z(\infty) \right|}{\left| z(\infty) \right|} . \tag{9}$$

The reference signal tracking properties can be investigated by means of the steady-state tracking error e_r as well as the actual cycle time t_c (which is the total time necessary to reach the final position of stacker crane). The steady-state tracking error can be defined as:

$$e_r = \frac{|y(\infty) - r(\infty)|}{|r(\infty)|} . \tag{10}$$

These time-domain quantities according to the two design cases are shown in Table 3.

Table 3
Time-domain analysis results of design cases

Case #1	Case #2
$\sigma_a = 71.6\ \%$	$\sigma_a = 0\ \%$
$e_r = 0.70\ \%$	$e_r = 0.40\ \%$
$t_c = 27.4\ s$	$t_c = 29.3\ s$

As can be seen in the presented simulation results the inverse proportionality between the magnitude of mast vibrations and control error here also exists. Therefore, in controller design the trade-off between mast-vibration attenuation and cycle time of stacker crane motion can be found. To explore this trade-off a series of controller designs and time-domain analyses are carried out again with several W_e and W_p weighting functions. In these investigations the control input and sensor noises weighting functions were permanent and identical to the functions presented in Table 2. During the investigations the weighting strategy has changed from the cycle time focusing cases to the vibration attenuation focusing cases. In the presented eleven design cases the 0 dB crossover frequencies of weighting functions W_e and W_p are modified evenly between its extreme values. In the case of W_e this crossover frequency is modified from 2 rad/s to 1 rad/s, while in the case of W_p function this value is changed from 20 rad/s to 10 rad/s.

In order to find an ideal design case the overshoot and the cycle time values of every design case are plotted in Figure 6. Analyzing the data of Figure 6 it can be observed that the overshoot of mast deflection signal vanishes sharply before the cycle time of stacker crane motion considerable starts to increase. Therefore, a sufficient trade-off between conflicting performances can be found.

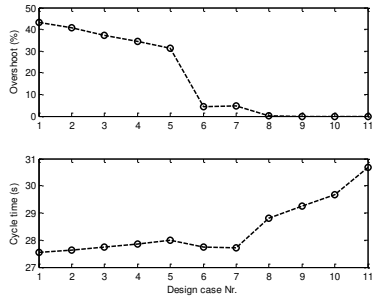


Figure 6
Trade-off between mast-vibration and cycle time

The designed controllers are calculated by means of the so-called μ-synthesis method presented in [19]. The achieved structured singular values μ for each the design cases are also plotted in the diagram of Figure 7. As shown in the figure, although the robust stability and nominal performance is achieved, guaranteeing the robust performance is a challenging task due to the strict performance specifications. However, as can be seen in Figure 7 the proposed method guarantees robust performance in the interesting region of the design cases where the vibrations are sufficiently damped.

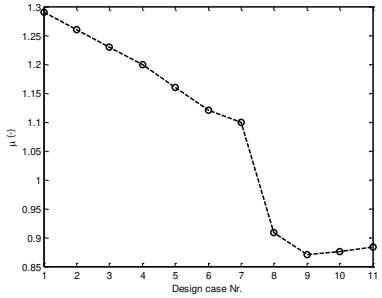


Figure 7
Achieved μ values of design cases

Conclusions

In the paper a robust controller design method was developed which is able to handle the uncertainties in the dynamic model of single-mast stacker cranes and at the same time has excellent reference signal tracking and mast-vibration attenuation properties. In the first part of the paper the dynamic modeling of single-mast stacker cranes by means of multibody modeling approach was briefly summarized. The unstructured uncertainty approach was applied to handle the varying dynamical behavior due to varying lifted load position. A robust control method was developed which is suitable for positioning control of stacker cranes

with reduced mast-vibrations in the presence of model uncertainties. By means of a controller design example the trade-off between mast-vibration attenuation and cycle time of stacker crane motion was also presented. The developed designing method is suitable for finding the controller which produces the desired motion cycle time and mast-vibration free stacker crane motion.

References

[1] Jorge Angeles and Andrés Kecskeméthy. *Kinematics and Dynamics of Multi-body Systems.* Springer-Verlag, 1995

[2] Dieter Arnold and Michael Dietzel. Aktive Schwingungsdämpfung von Regalbediengeräten. *F+H Fördern und Heben,* 50(1-2):50-52, 2000

[3] Javier García de Jalón and Eduardo Bayo. *Kinematic and Dynamic Simulation of Multibody Systems - The Real-Time Challenge.* Springer-Verlag, 1994

[4] Michael Dietzel. Beeinflussung des Schwingungsverhaltens von Regalbediengeräten durch Regelung des Fahrantriebs. Dissertation, Institut für Fördertechnik Karlsruhe, 1999

[5] John C. Doyle et al. State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems. *IEEE Transactions on Automatic Control,* 34(8):831-847, 1989

[6] B. A. Francis, J. W. Helton, and G. Zames. $\mathcal{H}_\infty$ optimal feedback controllers for linear multivariable systems. *IEEE Transactions on Automatic Control,* 29(10):888-900, 1984

[7] Pascal Gahinet and Pierre Apkarian. A linear matrix inequality approach to $\mathcal{H}_\infty$ control. *International Journal of Robust and Nonlinear Control,* 4:421-448, 1994

[8] Keith Glover and John C. Doyle. State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relations to risk sensitivity. *Systems and Control Letters,* 11:167-172, 1988

[9] Sándor Hajdu and Péter Gáspár. Investigation of the influence of lifted load on dynamical behavior of stacker cranes through unstructured uncertainties. In *CINTI 2013 : Proceeding of the 14th IEEE International Symposium on Computational Intelligence and Informatics, Budapest: IEEE Hungary Section,* pages 179.184, 2013

[10] Sándor Hajdu and Péter Gáspár. From modeling to robust control design of single-mast stacker cranes. *Acta Polytechnica Hungarica,* 11(10):135-149, 2014

[11] Sándor Hajdu and Péter Gáspár. Multi-body modeling of single-mast stacker cranes. *International Journal of Engineering Systems Modelling and Simulation,* 8(3):218-226, 2016

[12] D. McFarlane and K. Glover. A loop shaping design procedure using $\mathcal{H}_\infty$ synthesis. *IEEE Transactions on Automatic Control,* 37(6):759-769, 1992

[13] József Polák and István Lakatos. Hajtómű matematikai modell elemzése, In: Péter Tamás (szerk.), *Innováció és fenntartható felszíni közlekedés: IFFK 2015. Konferencia helye, ideje: Budapest, Magyarország, 2015.10.15-2015.10.16.*, Budapest: Magyar Mérnökakadémia (MMA), 2015

[14] József Polák and István Lakatos. Examination of drive line mathematical model. *Machine design* 8(1):33-36, 2016

[15] József Polák and István Lakatos. Efficiency optimization of electric permanent magnet motor driven vehicle. *Machine design* 7(1):11-14, 2015

[16] József Polák and István Lakatos. Analysis of propulsion unit mathematical model. *Machine design* 7(4):137-140, 2015

[17] Karl-Heinz Reisinger. Schwingungssimulation von Regalförderzeugen. Dissertation, Technische Universität Graz, 1998

[18] Meinhard Schumacher. Untersuchung des Schwingungsverhaltens von Einmast-Regalbediengeräten. Dissertation, Institut für Fördertechnik Karlsruhe, 1994

[19] Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control.* Prentice Hall, New Jersey, 1996

The Design of the Personal Enemy - MIMLeBot as an Intelligent Agent in a Game-based Learning Environment

Kristijan V. Kuk

Academy of Criminalistic and Police Studies, Cara Dusana 196, 11080 Belgrade, Serbia, kristijan.kuk@kpa.edu.rs

Ivan Z. Milentijević

Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia, ivan.milentijevic@elfak.ni.ac.rs

Dragan M. Ranđelović, Brankica M. Popović, Petar Čisar

Academy of Criminalistic and Police Studies, Cara Dusana 196, 11080 Belgrade, Serbia, dragan.randjelovic@kpa.edu.rs, brankica.popovic@kpa.edu.rs, petar.cisar@kpa.edu.rs

Abstract: A bot is one of the main elements of all computer video games, frequently used for the creation of various opponent characters within a game. Opponent modeling is the problem of predicting the agent actions in a gaming environment. This paper proposes and describes the implementation of a bot as a personal opponent in a small educational game. In order to increase the efficiency when using such a small educational application/module, artificial intelligence was added in the form of a bot competing with the students. Pedagogical elements of the intelligent learning system are introduced through the pedagogical model and the student model. This paper demonstrates the use of the student model to present the player model built by the experience of a human teacher, with true/false questions incorporated with the bot strategy into the opponent model. The authors use the Monte Carlo approach in this implementation, known as artificial intelligence technique and a best-first search method used in most video games, but to the best of their knowledge, it has not been used for prediction in educational games based on bot strategy. The results highlight that the Monte Carlo approach presented via the BFTree classifier provides the best classification accuracy compared with other predictive models based on data mining classifiers. It was shown that the training data from the human player can help in creating a bot strategy for a personalized game-based learning system. The Help option can be used for the assessment of the students' current knowledge by counting the number of Help option accesses, the player relies on Help as a 'source of

knowledge' needed to complete the game task successfully. The obtained results show that the bot (personal opponent) stimulated players to replay the game multiple times, which may contribute to the increase of the students' knowledge.

Keywords: knowledge personalization and customization; educational games; intelligent tutoring systems; personalized e-learning

1 Introduction

Depending on the type of game playing, video games may be associated with positive cognitive outcomes [1]. Simple games are mainly intended for one player. Therefore, the personal opponent developed as a game bot might have an important role in them. Game bots as 'automated programs with or without artificial intelligence that help players enhance, accelerate, or bypass some routines in the game' [4] are generally approved by the gaming community. Artificial intelligence (AI) depends both on knowledge about the world, and algorithms to intelligently process that knowledge. Three key understandings about the world, called 'models', used by intelligent systems in education, are the pedagogical model, the domain model, and the learner/student model. The application of artificial intelligence to education (AIED) has been the subject of many academic studies focusing on pedagogical agent realization [5, 6] or personalized educational game [7]. They are mostly based on a student model and teaching strategies implemented by agents technology including pedagogical features (pedagogical model) in an intelligent e-learning systems [8].

For the effective integration of educational games into the teaching material, Singh and Sivaswamy [2] proposed concentration on developing simple and small games instead of creating highly complex game systems where the student is then left alone to figure out the relations. Multimedia Interactive Modules for Learning – MIMLE is an E-learning system with many small education modules that comprises game elements [3]. It is a set of 2D interactive modules which showed significant success with teenage students. Through the Help window in the MIMLE system, the student is provided with textual messages in the form of a theorem or a definition, vital for successfully solving the problem given in the current module level. From a pedagogical aspect, the MIMLE system stimulates students to draw conclusions based on the accuracy of the achieved steps in solving the entire task on their own. Adding time as a game element contributing to the player's higher ranking in the game, also results in an increase of knowledge, as well as the improvement of the skills acquired through the easier levels. In order to keep the player as long as possible in the game, there is a need for some stimulation of his/her competitive spirit (in competition with themselves or with other players). The MIMLE system evaluates the player through his interaction with the system in the form of keeping count of the correct or incorrect

answers and tracking the Help option activation. Appropriate changes are made in the student model based on the recorded player performance. The conversion of the student model with teaching strategies so as to create a player model with agent strategies is not an easy task in the process of designing and developing educational games. The bot or opponent in the given e-learning system should assume the role of an intelligent agent aiming not to demoralize the student as a player by playing better or worse than the player does. Instead, the bot should closely match the student's performance, thereby continuously forcing him to compete.

The bot as a personal opponent in the modules of the MIMLE system, follows the interaction of the student with the system, and subsequently compiles its own set of answers requested in the game. Therefore, the bot and the student compete with each other. Based on the detected player performance, the appropriate changes are then made in the student model. In order to achieve better results than his bot, the student will start the game, over and over again, thus getting a chance to open the Help option several times and, in that manner, increase his knowledge. In this paper, the authors propose to design a personal opponent approach for the intelligent agent – bot in an MIMLE system. Machine learning is a technique by which the computer 'learns' from the set of given training data, and then the set is able to predict the result of new data. The machine learning algorithms are designed to identify patterns based on different characteristics or 'features' and then make predictions about the new, unclassified data based on the patterns 'learned' earlier. Many classification techniques are implemented on the educational datasets used to build a player model in most video games. The Monte Carlo tree search is a method for making optimal decisions in numerous artificial intelligence problems, especially for the personal opponent approach into combinatorial games. The Monte Carlo approach presented as a form of data mining classifier can give very good classification accuracy in predictive models for designing the bot strategy.

The current state-of-the-art options in teaching strategies for evaluating students' knowledge and existing opponent techniques in playing are shown in Section 2. Section 3 describes the conceptual framework of small game-based modules, addressing the problem of a player modeling based on student model in small e-learning module [3]. The knowledge discovery techniques presented in Section 4 are used for developing the bot strategy based on the data analysis of the player's answers and player's interaction with the Help option. Also, the methodology of searching for hidden connections between the gaming data is presented in the form of data mining algorithms and classifiers to create the best bot strategy. The experimental results, presented in Section 5, confirmed the validity of the described bot strategy (called MIMLeBot) and the authors' policy in setting up possible states in the player model that the player/student could experience in a special class of educational games, justifying its implementation in such a type of e-learning system.

2 State of the Art

Exams are commonly used assessment and evaluation tools in universities and there are numerous types of exam questions, generally categorized into 7 types [9]. One of them is the true-false (T/F) question type. With this type there are only two options for answering: “True” and “False”. This question provides students with a 50% chance of guessing the correct answer. Due to this fact T/F questions are suited for evaluating students' knowledge of specific facts and concepts, therefore true/false games are usually used as educational games. The student, as a player, is offered a task sequence (as a picture or text) and is expected to select the T/F item. For a simple T/F item, each student has a 50/50 chance of correctly answering the item even without any knowledge of the item's content [10, 11]. The teaching strategy for game completion is created, by monitoring the student’s response to the T/F items. In this paper, a simple educational game is developed in the described T/F manner, where a sequence of rectangles, i.e. ‘boxes’ represents T/F items.

Intelligent tutoring systems and agent-based learning environments also provide students with individualized practice rather than static sets of tasks. Many Intelligent Virtual Teaching Environments (IVTEs) include pedagogical features that are based on the student model and teaching strategies [12]. A student model (which reflects the state of knowledge), is applicable if its present state can be utilized by a certain interpreter to simulate the behavior of the modeled student when the student is solving training problems. The three-tier architecture of a typical agent IVTE system consists of a domain model, a student model and a pedagogical model [13]. Teaching an opponent model can be approached as a pattern recognition task, where the model is taught based on the history of the players’ previous actions [14]. However, this paper proposes a different approach. The authors determined a posterior distribution (using Bayes’ rule) for two specific states when the student’s answer is an accidental hit or miss, to gain appropriate (defined) bot action for those states. With this approach for T/F items, the use of the Help option plays a significant role in successfully completing the game. It may increase the students' knowledge of specific facts and concepts. For example, if the first answer is correct, and the second answer incorrect, then the probability that third answer will be correct is > 50% if the student opts for using the Help option.

In the game proposed here the player selects a color box out of a choice of four so as to complete a task, in this case it is a trick-taking game with players selecting a card sequentially. In addition, in trick-taking games, the player has one optimal strategy and should play the cards in a specific order. In many games the optimal opponent strategy is based on different search methods. Monte Carlo Tree Search (MCTS) is a best-first search method that builds a search tree iteratively. MCTS has been used in several single player games [15] such as the puzzles SameGame or Morpion Solitaire. The authors in [16] applied MCTS in the context of mini-

max game trees, since they are encountered in incomplete information games such as Poker, where the opponent's actions cannot simply be predicted as value minimization actions. Using a back propagation strategy, they are evaluated as a part of a complete Poker bot (kind of MCTS bots). MCTS is a search algorithm based on random play-outs. It has been observed that MCTS is successful for trick-taking card games, but rather not suitable for poker-like card games. The question the present authors are set to answer is whether or not MCTS can be applied in an educational game, as well, e.g. by creating the bot strategy as a personal opponent.

3 Game-based Learning Modules for Computer Science

The authors created a module as special modules into the MIMLE system, which combined different types of the existing interactive multimedia environments for learning mathematics, physics and electronics. The game concept is based on two components: a) students have to obtain the course information through its interpretation in the game world; b) students have to see the results of this algorithm in a game context. Also, apart from placing a game interface into the learning environment, the authors also applied basic game elements, such as: result, time and difficulty levels. These new modules, named game-based modules [17], which include game elements, represent research multimedia learning applications and are intended for Computer Science students. This module was deployed as a learning environment for topics in the field of Computer Science. To increase the engagement and interest of students for this type of teaching material, the authors included game characteristics in this environment. Students need to have appropriate knowledge to successfully solve the tasks given as part of the game, and to advance from one game level to the next. The *MIMLE* system is presented as an e-learning system based small two-level game-based modules for a single player.

When experiencing any difficulties in task solving on a certain level, the Help option can accelerate the student's finding the right solution. Help option activation enables the students to learn the rules and apply them in practice on the presented examples. Therefore, the formulation of definitions and theorems within the Help is a crucial moment in designing the entire application.

3.1 Simple Module for Teaching Z-buffer Algorithm

The possibility of a visual representation of the task solving method for practicing the material in the Computer graphics module enabled their implementation as a small game-based module into the MIMLE system. The 'Z-buffer' module in the

MIMLE system was designed to help students learn basic terms referring to the principle of the Z-buffer algorithm operations and apply them practically through solving the given examples with a randomly generated content of buffer registers. This game contains interactive tasks implemented in the graphic environment that are visually directly associated with the learning material.

While analyzing the techniques for the first-class innovative testing select/recognize [18], the authors opted for giving the answers in MIMLE as a series of boxes to be clicked. The player must choose one out of 5 colors to represent a bit in the buffer registry the moment when the scanning line moves across the screen. Since the implemented buffer in the 'Z-buffer' algorithm uses 15 bits, the task of this module is to determine the value of each bit (i.e. contents of the registry) in various situational polygons shown on the screen. By theoretical definition, the frame buffer is used to store the intensity value of color value at each position (x, y). The goal of the game is for the player to test the z - depth of each surface (sequence of box in one color) so as to determine the closest (visible) surface (sequence of box in another color). The player determines the Z-value presented in the game as T/F items in the screen buffer registry in the moment by moving the scanning line. The complete buffer registry should display the one (pixel by pixel) that has the smallest value from the camera.

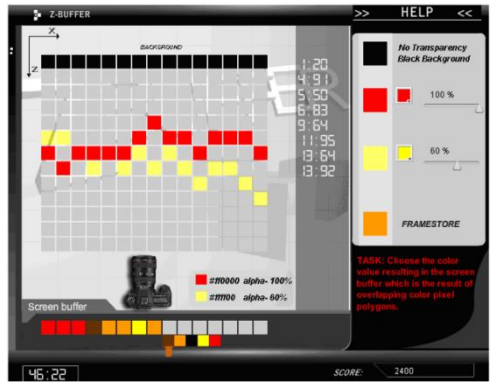


Figure 1
Interface of the 'Z-buffer' module

The correct answer which fills the content of one bit is one of the proposed answers presented to the students in the form of a box to be selected. These cubes (i.e. the offered answers), are presented in two ways. In Level 1 of the module, the answers are offered in the form of a 15-box column. When a certain buffer box is selected from the scanning line (moving from left to right), the box falls down and fills the content of the active bit (one sequence) in the buffer. Determining the buffer register contents is a task in Level 2, only during this time, does the player see polygons having some level of transparency. In comparison to Level 1, Level 2 is more challenging because it requires the player to determine the final color of

the buffer cube which presumes additional knowledge of mixing two colors with different percentage of visibility.

3.2 Player Modeling Based on Teaching Strategies

A common characteristic of many intelligent tutoring e-learning systems is that they can recognize whether or not the students understand the given lessons and thus adjust the learning process in accordance to the students' needs. This kind of reasoning is also known as setting a diagnosis, revealing the students' level of knowledge [19]. The students' current level is represented by the student model. The teachers do not simply consider the students' answers in the exam, but also the sequence of the players' actions during the exam (e.g. the students make a mistake, then correct themselves and give the right answer). They remember the number and type of the students' mistakes, as well as the number of the questions or some other type of help used to move them away from a "deadlock". The teachers grade the students' knowledge based on all these details.

The player's success in the learning system is often seen as a sequence of the given answers [20]. Based on the sequence of correct and incorrect answers, the system can make the decision about whether or not the player was successful. In the 'Z-buffer' module the authors used a diagnostic technique called model tracing [21]. Monitoring the player's interaction with the system has been reduced to the player's last three steps and the system keeps track of the values of their combinations (i.e. the situations that result from the last three events are observed). As stated earlier in the paper, the player has only three options: the correct answer, the incorrect answer and asking for help. The states that the player can find within 'Z-buffer' module are shown in the last column of Table 1. Their determination is based on tracking the last two answers and the appearance of a Help window that was opened by the player before providing the second answer. The possible states of the player dependent on the events are shown in Table 1.

Table 1
Possible states of a player, depending on the events

$Answer_{t-1}$/ A_{t-1}	$Answer_t$/ A_t	$Help_t$ / H_t	$State_t$/ S_t
Ic	Ic	N	Does not know
Ic	C	N	Accidental hit
Ic	Ic	Y	Does not know
Ic	C	Y	Knows
C	Ic	N	Does not know
C	C	N	Knows
C	Ic	Y	Accidental miss
C	C	Y	Knows

Abbreviations used in Table 1 are: Correct – C, Incorrect – Ic, Yes – Y andNo – N

The intelligent agent – bot exclusively decides its own actions based on the action of the state in which a player is. Depending on his interaction with the 'Z-buffer' module (Figure 1), the player can find himself in one of the four states: knows, does not know, accidental hit, accidental miss, therefore, S = *(Knows, Does not know, Accidental hit, Accidental miss*).

4 Implementation of the Bot in Game-based Modules of the MIMLE System

The authors chose modules of the MIMLE system as the environment where the bot will be implemented as a personal opponent. Possible decisions are presented, the bot's answers, as 'Correct' or 'Incorrect'. An MIMLE bot is a kind of a reflex agent which plays fully automatically according to the player model built by the human teacher (in the present case, by the authors of this paper). Playing against human players is the best way to test the player's knowledge. The first issue needed to model a strategy, is to obtain data from the games between human players. Intelligent agents in combination with fuzzy logic can help increase the quality and amount of interaction in a computer game.

The acquired knowledge is usually represented in the form of 'if-then' prediction rules. This representation is preferable for being a high-level, symbolic knowledge representation, contributing to the comprehensibility of the knowledge acquired. These decision rules can be placed in a decision queue (DQ) [22], which entails that they must be applied in a specific order. With this policy, the number of rules may be reduced because the rules could be one inside or another. The knowledge manager decides which representation provides the highest level of accuracy, while producing the smallest number of rules. DQ presents the following structure [23]:

If conditions ***then*** class ***Else if*** conditions ***then*** class ***Else if*** conditions ***then*** class ***Else*** 'unknown class' (1)

The discovered rules can be evaluated according to several criteria, such as the degree of confidence in the prediction, classification accuracy rate on unknown-class examples, comprehensibility, cost, etc. Since each feature, used as part of the classification procedure, can increase the cost and running time of a classifier system, as well as reduce the accuracy of the result, there is a strong motivation to design and implement systems using small feature sets.

A decision or classification tree can be described as 'a tool representing an algorithm in the form a graph or binary, tertiary or n-ary tree with each node and branch having a certain associated outcome, weight in terms of outcome, and probability' [24]. It is easy to implement, understand and customize. It is considered as one of the fastest in terms of learning and classification among

machine learning techniques, where it can be seen as a predictive model which makes decisions on a branching series of Boolean tests. In computer logic terms, it can be viewed as a series of nested 'if-else'.

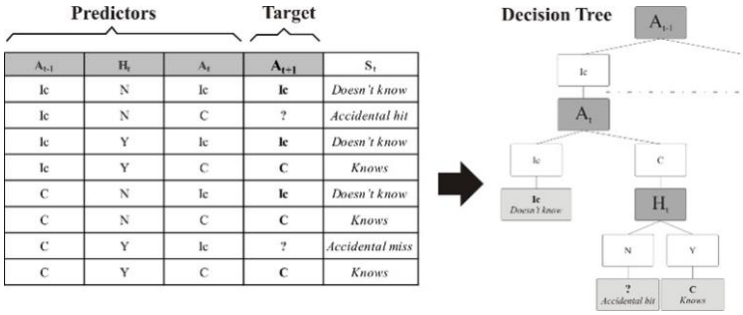


Figure 2
A decision tree for the bot's answers problem

Figure 2 shows a sample decision tree of the bot's answers. Here, the variable to be predicted or the 'target variable' is the $Answer_{t+1}$ –'A_{t+1}' variable. Given the set instances, each with same values for $Answer_{t-1}$ –'A_{t-1}', $Answer_t$ –'A_t' and $Help_t$ – 'H_t' and the resulting value is Knows, a tree is constructed as shown. Further, when given an instance with values for A_{t-1}, A_t and H_t, the machine learning system ought to be able to predict the value for the bot's answer ($Answer_{t+1}$) variable by traversing through the tree, based on the conditions in the given instance. The target variable A_{t+1} may have accidental states with the same values of predictors. Usually, one accidental variable is identified as independent (x), and the other one as dependable accidental variable (y). The set of statistical methods that explores interlinks of statistical labels and appearances (direction, strength, shape) is called the theory of correlation, and the basic indicators of correlation links are equation of regression and correlation coefficient.

In the domain of machine learning, a classification problem involves machine learning attempting to predict to which class/group/category a new observation would most likely belong to, based on the training data set, which contains the already classified data. The classification of a given object is based on finding similarities with previously determined objects belonging to different classes, whereas the similarity of two objects is determined by analyzing their characteristics [25].

4.1 The Built Predictive Model

Classification is performed through supervised learning. After extracting the game variables, the lists were used along with Weka (a collection of machine learning algorithms for data mining tasks) to train different classifiers [26]. The classifiers

were evaluated and it was determined which one was more suitable for this domain.

A total of 240 records were extracted from the 'Z-buffer' MIMLE database for analysis. The discarded records were related to the players who had only started the module, but then failed to perform any further activities. The analysis included 25 representative players and their activities. The original data vector contain 30 typical activities for each player (15 answers and 15 registered contacts with the Help option), and they represent attributes - independent input variables for a predictive model. By applying a discretional filter, numerical values of the A_{t+1} attribute were transformed to nominal and two intervals were determined ('*Correct*-C', '*Incorrect*-Ic'). A_{t+1} attribute was labeled as a class attribute, representing the dependent variable for a predictive model.

There are various techniques to test/estimate the performance of a predictive model. For the case described in this paper, the MIMLE dataset divides the training set into two parts (usually 1/3 and 2/3) where the larger part is used for training the model and the other one for validating it. Having the data ready, the next step is to use classifiers so as to learn the tactics presented on them. The following classifiers were used: Bayesian Networks (BayesNet), Best First Search Tree (BFTree), C4.5 Search Tree (J48), Multilayer Perceptron Neural Network (MultilayerPerceptron), Naive Bayes (NaiveBayes), Random Forest Search Tree (RandomForest) and SMO (SMO).

Table 2 presents the classification of the MIMLE dataset in which the BFTree, J48, Multilayer Perceptron, SMO algorithms show 75 % of correctly classified instances, followed by Bayes Net, Navie Bayes and Random Forest with 74.17% of correctly classified instances.

Table 2
Classification of MIMLE dataset using Weka

	Bayes Net	*Naïve Bayes*	*BFTree*	*J48*	*Random Forest*	*Multilayer Perceptron*	*SMO*
Correctly Classified Instances	178 (74.17%)	178 (74.17%)	180 (75%)	180 (75%)	178 (74.17%)	180 (75%)	180 (75%)
Incorrectly Classified Instances	62 (25.83%)	62 (25.83%)	60 (25%)	60 (25%)	62 (25.83%)	60 (25%)	60 (25%)
Kappa statistic	0.3393	0.3393	0.4143	0.4184	0.3342	0.4184	0.4184
Mean absolute error	0.3268	0.3377	0.3348	0.3594	0.3395	0.3383	0.25
Root mean squared error	0.4187	0.4163	0.4091	0.4239	0.4097	0.4106	0.5

Relative absolute error	74.90%	77.39%	76.72%	82.37%	77.80%	77.54%	57.30%
Root relative squared error	89.70%	89.19%	87.64%	90.81%	87.76%	87.95%	107.11%
Total Number of Instances	240	240	240	240	240	240	240
Confusion Matrix	a b 147 16 | 46 31 |	a b 147 16 | 46 31 |	a b 136 27 | 33 44 |	a b 135 28 | 32 45 |	a b 148 15 | 47 30 |	a b 135 28 | 32 45 |	a b 135 28 | 32 45 |

The measure of the dataset between the categorization of the predicted and observed is called Kappa Statistic. If the predicted and observed values are identical, then the Kappa Statistic value equals 1. The classifiers were evaluated through one of two error rates. The average values of Root mean squared error (RMSE) and Mean Absolute error (MAE) are determined in order to select the best predictive model. If both error values are higher, the accuracy is lower and vice versa. Kappa statistics for J48, Multilayer Perceptron and SMO showed the maximum. The MAE is low for SMO. The RMSE is low for BFTree. It is not possible to claim that MAE is a better indicator for model performance than RMSE because it is smaller. The chart in Figure 3 demonstrates the average error rate for each classifier.

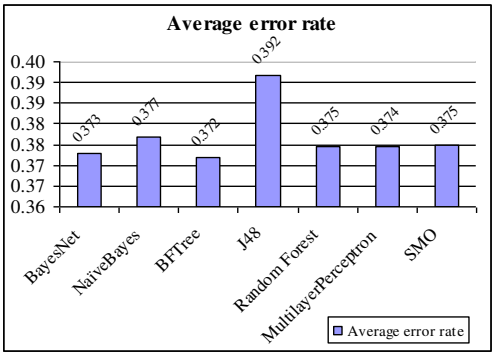


Figure 3
Average error rate

As can be seen both in Table 2 and Figure 3, the classifiers that performed better for this predictive model were the search trees, more specifically, the BFTrees with an average error of approximately 32.7%. The errors are relatively high, as was to be expected since the MIMLE bot tends to change tactic during the game, making it difficult to find a pattern with a small error.

4.2 Creating the Bot Strategy

In order to determine the likeliness of the possible states the player could fall into after three interactions with the game, the authors compared the results acquired through the predictive model with the parameters set up by the authors (on the basis of teaching experience) in Table 1. This paper shows that deploying classifiers for the prediction model design can be used as a method for detecting undefined players conditions, for which the target variables do not have values in the players model table ('Accidental miss/hit').

After training the selected predictive model on the basis of the game, with 240 players, the acquired results confirm the validity of the players' model (shown in Table 1) set up by the teachers. In the re-evaluation of the model with the data test set where all instances contained the values '*?*' for all possible players conditions (Table 3), the same values were obtained as assumed and set up for the class attribute – target during the modeling process by the authors .

Table 3

Predictions on test set

Instance#	Actual	Predicted	Probability	Distribution
0,0,n	?	2:Ic	0.359	*0.641
0,1,n	?	**1:C**	***0.563**	**0.438**
0,0,y	?	2:Ic	0.167	*0.833
0,1,y	?	1:C	*0.857	0.143
1,0,n	?	2:Ic	0.462	*0.538
1,1,n	?	1:C	*0.863	0.137
1,0,y	?	**1:C**	***0.5**	**0.5**
1,1,y	?	1:C	*1	0

Abbreviations used in Table 3 are: Correct – 1, Incorrect – 0, Yes – y andNo – n

Bot strategy creation is based on a set of *if-then-else* decision rules, as well as on decision tables. The question is what action the bot should perform if the target remains in an unknown state. Once the model was trained based on the BFTree classifier, it can be used to classify as yet unseen MIMLE data as part player model, presented in Figure 4. First, the file with the cases to be predicted must have the same structure as the file with the training set used to teach the model. Prior to training the classifiers, it was necessary to create an 'arff' file so that Weka can recognize the data. Assuming that one has trained the decision tree using the MIMLE datasets with 8 instances, as shown in the Table 1 column called $State_t/S_t$, one receives the result of predictions obtained on test data. The predicted column contains '*Correct'* or '*Incorrect'* for each of the lines in the test file. The instances with data which show player's states: '*Accidental miss'* and '*Accidental hit'*, predicted columns are marked as '*Correct'* value.

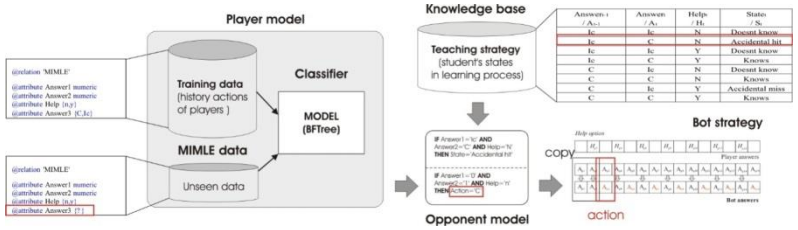


Figure 4

Generalizable opponent strategy approach in MIMLE modules

After teaching the tactics, the next stage is to implement the bot strategy for its answers in the game. The idea behind the strategy is that the bot will follow the combinations that the player accomplishes with the MIMLE module and thereby, will fill in the answers in the game. The first and second answer is copied from the player, while the third answer is filled in on the basis of the assumed position the player currently holds (Table 1). The subsequent set of the answers is filled in by moving one step back so that the fourth answer is copied from the player while the fifth answer is given on the basis of the player's condition caused by its third answer interpreted as the player's third answer, the player's fourth answer and option of asking for help during that time-frame. At the end of the game, the player can see the final set of answers given by the bot after each level (Figure 5). Further, the player can see the general success that he or she achieved, as portrayed by the accomplished scores (Figure 6).

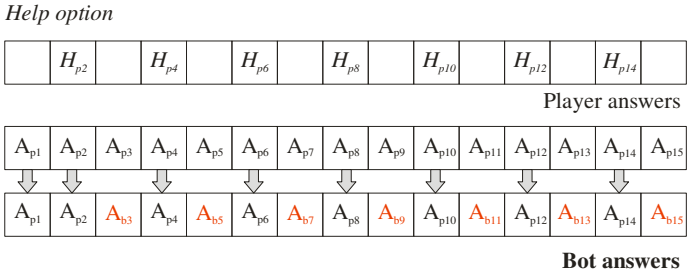


Figure 5

Bot strategy for the student's answers in modules

Figure 6
MIMLe Bot answers at game end of a level in the module

5 Results and Discussion

Decision trees are among the most popular classification techniques in data mining [27]. The classifier, whose average error value with MIMLE player data was shown as the lowest, was BFTrees. The performance of the BFTrees using different test options is presented and interpreted in Table 4. The performance is interpreted in terms of accuracy and error rate.

Table 4
Performance of BFTrees under different test options

Test Options	Accuracy	Error Rate	Kappa Statistic	MAE
UTS	75%	25%	0.4143	0.3348
CV (10 folds)	71.7%	28.3%	0.3454	0.3444
PS (66%)	74.4%	25.6%	0.357	0.3392

Table 4 demonstrates that the performance is the best when tested with "*use training set - UTS*" followed by "*cross validation-CV*" with 10 folds (it would apply training on the first 9 parts and testing on the last part), then with "*percentage split-PS*" option (random percentage split of the dataset is going to be 66% training data and 34% test data). In the BFTrees the selection of the best split is based on boosting algorithms [28] which are used to expand nodes in the best-first order instead of a fixed order. The results of this study confirmed that the Monte Carlo approach presented via BFTree classifiers provides the best classification performances in a small educational game, which supports the

authors' assumption that it can be successfully used in these kinds of educational games.

The authors performed the analysis of the number of game activations with and without the bot. The efficiency of using the bot and its role in game-based modules by the *MIMLE* system was studied by analyzing the interaction of players with the system, as well as having an actual insight into their achievements at the end of the game. The analysis included 32 typical players for the Z-buffer module. The number of game activations was followed up for each player, as were their highest scores when the system did not contain a bot, and in the case when player competed with bot. It was established that the number of repetitions of the game increased in the case when the player competed with his personal opponent. The number of repetitions of the module the player performed with and without the use of the bot can be seen in Figure 7. The average value of module repetition without a bot was 3.78, while for the same player, the value increased to 6.47 for when he or she was playing against the personal opponent. It was also determined that the player made progress in his or her average value of the points.

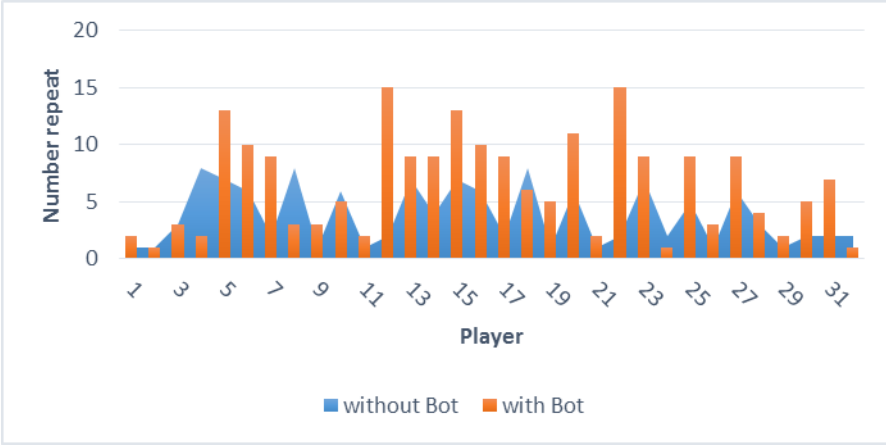


Figure 7

The interaction of players with the game-based module Z-buffer

In order to test whether or not a bot, personal opponent, has an impact on the number of game replays by the player, the authors formed two groups (experimental and control group), as shown in Figure 7, consisting of 32 players each. The results obtained using the (non-parametric) Mann-Whitney's test, by conventional criteria, indicate a statistically significant means difference in the number of game activations with and without the bot. The realized U-value is 309.5. The distribution is approximately normal, which indicated that the Z-value below should be used. The Z-Score is -2.71229. The p-value is .00672. The result is significant at $p < .01$. This result confirms the authors' right attitude regarding the importance of using a personal opponent in the game since its existence

triggers a greater interest in the player, and motivates him or her to compete and achieve better results.

Conclusions

Many incorporated AIED and educational data mining (EDM) techniques are used for 'tracking' student's behavior – for example, collecting data on class attendance and assignment submission in order to identify (and provide support for) students who are at risk of dropping out from their studies. Data mining methods provide automated predictions of the proposed solutions on the basis of well-known behavior patterns of the past, as well as the identification of previously unknown relationships, patterns and trends in very large databases.

This paper outlines a part of a bot strategy based on EDM, and its implementation in a small game-based learning system. Machine learning algorithms in EDM are applied to the task of detecting a bot's strategy before it is executed and predicting when a player will perform strategic actions. As researches in gaming try to draw conclusions about player characteristics from their actions in open-ended gaming environments, understanding players' goals can help provide an interpretive lens for those actions. The idea behind the bot strategy is that the bot follows the combinations that player accomplishes with the MIMLE module: correct answer, incorrect answer and asking for help. Based on the present authors' analysis, it can be concluded that in the case of a small training data set, the BFTree algorithm provides good prediction results if the data are a combination of numerical (ANSWER1 and ANSWER2) and category types (Help), where a class attribute can have one of two class values (ANSWER3). The performance of the BFTree algorithm on this dataset was better than C4.5 algorithm.

Future work will focus on improving the precision of the bot's strategy and implementing artificial intelligence in educational networking. For example, bots and students might be able compete with each other in social networks, thus maximizing the effectiveness of the learning process.

Acknowledgement

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under the project no. III 47016.

References

[1] T. Greitemeyer, D. Mügge: Video games do affect social outcomes a meta-analytic review of the effects of violent and prosocial video game play, *Personality and Social Psychology Bulletin*, Vol. 40, No. 5, pp. 578-589, 2014

[2] J. Singh, J. Sivaswamy: Creating Educational Game by Authoring Simulations, *Proceedings of the 17th International Conference on*

Computers in Education, Hong Kong, Asia-Pacific Society for Computers in Education, 2009.

[3] K. Kuk, I. Milentijević, D. Rančić, P. Spalević: Pedagogical agent in multimedia interactive modules for learning–MIMLE, *Expert Systems with Applications*, Vol. 39, No. 9, pp. 8051-8058, 2012.

[4] K. Chen, H. K. Pao, H. C. Chang: Game bot identification based on manifold learning, *Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games*, ACM, pp. 21-26, 2008.

[5] C. Conati, A. Gertner, K. Vanlehn: Using Bayesian networks to manage uncertainty in student modeling. *User modeling and user-adapted interaction*, Vol. 12, No. 4, pp. 371-417, 2002.

[6] J. Sabourin, B. Mott, J. C. Lester: Modeling learner affect with theoretically grounded dynamic Bayesian networks, *International Conference on Affective Computing and Intelligent Interaction*, Springer, Berlin Heidelberg, pp. 286-295, 2011.

[7] M. M. Weng, I. Fakinlede, F. Lin, T. K. Shih, M. Chang: A conceptual design of multi-agent based personalized quiz game, *Advanced Learning Technologies (ICALT)*, 11th IEEE International Conference on Advanced Learning Technologies, pp. 19-21, 2011.

[8] W. L. Johnson, J. C. Lester: Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later, *International Journal of Artificial Intelligence in Education*, Vol. 26, No. 1, pp. 25-36, 2016.

[9] D. M. Zimmaro: Writing good multiple-choice exams, Measurement and Evaluation Center, *University of Texas at Austin*, https://facultyinnovate.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-04-28-10_0.pdf

[10] E. Ferrándiz, C. Puentes, P. J. Moreno, E. Flores: Engaging and assessing students through their electronic devices and real time quizzes, *Multidisciplinary Journal for Education, Social and Technological Sciences*, Vol. 3, No. 2, pp. 173-184, 2016.

[11] S. M. Čisar, D. Radosav, B. Markoski, R. Pinter, P. Čisar: Computer adaptive testing of student knowledge, *Acta Polytechnica Hungarica*, 7(4): pp. 139-152, 2010.

[12] M. A. Nunes, L. L. Dihl, L. M. Fraga, C. R. Woszezenki, L. Oliveira, D. J. Francisco, G. J. C. Machado, C. R. D. Nogueira, M. G. Notargiacomo: Animated pedagogical agent in the intelligent virtual teaching environment, *Digital Education Review*, Vol. 4, pp. 53-61, 2002.

[13] X. Luo, M. Spaniol, L. Wang, Q. Li, W. Nejdl, W. Zhang (eds.): Advances in Web-Based Learning-ICWL 2010, *9th International Conference*,

Shanghai, China, Proceedings Lecture Notes in Computer Science. Vol. 6483. Springer, 2010.

[14] M. J. V. Ponsen, G. Gerritsen, G. Chaslot: Integrating opponent models with Monte-Carlo tree search in poker, *Proc. Conf. Assoc. Adv. Artif. Intell. Inter. Decision Theory Game Theory Workshop*, pp. 37-42, 2010.

[15] S. Matsumoto, N. Hirosue, K. Itonaga, K. Yokoo, H. Futahashi: Evaluation of simulation strategy on single-player Monte-Carlo tree search and its discussion for a practical scheduling problem, *Proceedings of the International MultiConference of Engineers and Computer Scientists*, IMECS 2010, Hong Kong. Vol. 3, pp. 2086-2091, 2010.

[16] G. Van den Broeck, K. Driessens, J. Ramon: Monte-Carlo tree search in poker using expected reward distributions, *Asian Conference on Machine Learning*, ACML 2009, Nanjing, China, Springer Berlin Heidelberg, pp. 367-381, 2009.

[17] K. Kuk, I. Milentijević, D. Rančić, P. Spalević: Designing Intelligent Agent in Multilevel Game-Based Modules for E-Learning Computer Science Course, *E-Learning Paradigms and Applications*. Springer Berlin Heidelberg, pp. 39-63, 2014.

[18] R. Felder: Reaching the second tier: Learning and teaching styles in college science education, *College Science Teaching*, Vol.23, No.5, pp. 286-290, 1993.

[19] M. Chi, P. W. Jordan, K. VanLehn, M. Hall: Reinforcement Learning-based Feature Seleciton For Developing Pedagogically Effective Tutorial Dialogue Tactics, *Proceedings of 1st International Conference on Educational Data Mining*, EDM'08, Montreal, Quebec, Canada, pp. 258-265, 2008.

[20] I. Varlamis, S. Bersimis: Providing shortcuts to the learning process, Recent Progress in Computational Sciences and Engineering, Edited by Th. Simos & G. Maroulis, Lecture Series on Computer and Computational Sciences 7, 2006.

[21] K. Kuk: Artificial intelligence in process of collecting and analyzing data within police works, *Nauka, bezbednost, policija*, 20(3):131-48, 2015.

[22] J. C. Riquelme, J. S. Aguilar, M. Toro: A decision queue based on genetic algorithms: axis-paralle classifier versus rotated hyperboxes, *Computational Intelligence and Applications*, pp. 123-128, 1999.

[23] S. O. Danso: *An Exploration of Classification prediction techniques in data mining: the insurance domain*, Master Degree Thesis, Bournemouth University, 2006.

[24] C. Bhawe: Big data classification using decision trees on the cloud, Master's Projects Paper 317, http://scholarworks.sjsu.edu/etd_projects/317, 2013.

[25] G. Šimić, Z. Jeremić, E. Kajan, D. Randjelović, A. Presnall: A Framework for Delivering e-Government Support, *Acta Polytechnica Hungarica*, 11(1), pp. 79-96, 2014.

[26] S. M. Weiss, N. Indurkhya: *Predictive data mining: a practical guide*, Morgan Kaufmann, 1998.

[27] M. A. Hall, I. H. Witten, E. Frank: *Data Mining: Practical machine learning tools and techniques*, Kaufmann, Burlington, 2011.

[28] R. E. Schapire*: The boosting approach to machine learning: An overview*, In Denison DD, Hansen MH, Holmes C, Mallick B and Yu B, eds. Nonlinear estimation and classification, Springer New York, pp. 149-171, 2003

Land Consolidation based on Cluster Analysis

János Katona[1], Kornél Czimber[2], Andrea Pődör[1]

[1]Alba Regia Technical Faculty, Óbuda University, Budai út 45, H-8000 Székesfehérvár, Hungary, {katona.janos, podor.andrea}@amk.uni-obuda.hu

[2]Faculty of Forestry, University of Sopron, Bajcsy-Zsilinszky u. 4, H-9400 Sopron, Hungary, czimber.kornel@uni-sopron.hu

Abstract: The optimisation of land use structure is crucial to have a competitive agricultural production. In Hungary land consolidation lacks some important conditions such as a reasonable decision making support system based on Geoinformatics. In the paper, we optimize the land structure based on landownership. The present paper analyses the DigiTerra software that operates on the basis of Cluster Analysis and it provides a solution for the development. The improved software is introduced on a sample area. The efficiency of the allocation is proven by the internationally accepted fragmentation (Simmons, Januszewski, Igozurike) indices for parcels.

Keywords: land consolidation; cluster analysis; land valuation

1 Introduction

One of the main challenges of our time is to provide sustenance and clean water for people. The primary aim of the agriculture is to produce raw food materials to provide that sustenance. On the other hand, it is well known that monocultural production of thousands of hectares is not sustainable due to the stresses on the ecosystem and lowering of employment. It follows that adequate land politics can be both globally and locally provided that can support a land structure serving both personal and social interests.

The availability of agricultural assets in Hungary is favourable. However, the present fragmented land structure – resulting from the compensation and distribution of shareholdings in the 1990s – is not appropriate for competitive agricultural production. Therefore, it is necessary to consolidate the land.

According to the 2010 land use registry, on the average [1]:

- the number of land parcels used by a private person is 4.44 pieces
- the size of land parcels used by a private person is 9.54 hectares
- the number of land parcels used by professional farmers is 39.42 pieces
- the size of land parcels used by professional farmers is 296.38 hectares

The above-mentioned land sizes are not available for the maintenance of viable farms. The lease of land can somewhat help this situation, although it cannot provide a real solution because it removes capital from the production process. Since 2012 the policy to abolish the undivided joint ownership has also increased the extent of fragmentation. Consequently, the fragmented land structure cannot be maintained anymore. According to spatial data, the basis of land consolidation can be Geoinformatics.

2 The Background of Land Consolidation

2.1 The Measurement of Land Fragmentation

The efficiency of land consolidation methods can be measured by analysing land fragmentation, although land fragmentation has no objective measure. The number of parameters taken into the measurement is high and the scalability of parameters is not trivial. In this section, we introduce the index calculation accepted by international literature.

The index number by Simmons [2] gives a relevant value for a farm. This value is made up from the number of parcels *(n)*, the size of parcels (*a)* and the size of the whole farm (*A)*.

$$\mathrm{FI} = \frac{\sum_{i=1}^{n} a_i^2}{A^2} \tag{1}$$

Dorving [3] additionally, uses the distance that is taken by the farmer to reach one of his parcels. According to some critics [4], it would be more realistic to use both the back and forth distances and the annual frequency in the calculation.

A similar process was carried out by Januszewski [5], as well. He combined the number of land parcels belonging to a certain farm and their size distribution into a *K* factor. The value of this factor can change between 0 and 1, the nearer the value of K is to 0, the bigger the fragmentation is.

$$\mathrm{K} = \frac{\sqrt{\sum_{i=1}^{n} a_i}}{\sum_{i=1}^{n} \sqrt{a_i}} \tag{2}$$

The following consequences can be determined:

- the extent of fragmentation proportionally grows with the number of parcels,
- the fragmentation grows if the parcels are small,
- the fragmentation decreases if the size of big parcels grows and the area of small parcels falls off at the same time.

According to Igozurike [6], the average size of land parcels and the back and forth distances between them should also be taken into account.

$$P_i = \frac{1}{S_i / 100} Dt \tag{3}$$

, where P_i = fragmentation of the farm; S_i = the size of parcels; Dt = the whole back and forth distance.

The above-mentioned indices have three significant disadvantages:

- they neglect some spatial factors such as the index of land parcel per owner, the shape of parcels and some nonspatial factors like the type of the ownership or the accessibility of parcels,
- they are not flexible since the complex mathematical equation does not allow the separate handling of each and every member,
- they are not problem-oriented since they take the factors equally into account.

According to the literature [7, 8], each and every criterion (factor) and its entirety should fulfill some preconditions. Each and every criterion should be comprehensive to be measured objectively and reach its aim. The set of criterion should be complete, important aspects cannot be neglected. The criterion system should be flexible so that the problem can be divided into small parts such as economic, environmental, societal, etc. The final criterion system should be determined in a way to avoid the duplication of consequences of the decision. Duplication can be avoided in case of an additive result. If the correlation coefficiency of a criterion couple is near to 0, the two criteria are independent and not redundant. Therefore, six variables can be used [4]:

- the spatial location of land parcels,
- the size of land parcels,
- the shape of land parcels,
- the accessibility of land parcels,
- the type of the ownership,
- diversification of the ownership.

The so-called LandFragmentS Model [4] provides an index per land parcels by using the above-mentioned aspects and factors. The factors (f_{ij}) are determined with different weights (w_j). The weighted factors are contracted land parcel by land parcel (LFIi - land fragmentation index). The index referring to the whole area (GLFI - global land fragmentation index) comes from the summation of the land parcel index (LFI_i) / land parcel number (n) quotients.

$$LFI_i = \sum_{j=1}^{m} f_{ij} \cdot w_j \tag{4}$$

$$GLFI = \sum_{i=1}^{n} LFI_i / n \tag{5}$$

This complex evaluation method can provide an objective index for land proportion before and after the modification.

In the course of the DigiTerra software development, an evaluation method has been worked out on the basis of spatial data [9]. Besides land quality – that can be found in the real estate registry – , the method takes into account the shape, the size, the location, the accessibility, the relief and the slope conditions, the drainage, the irrigational conditions and environmental protection. Land quality found in the real estate registry can be modified with the above-mentioned factors, so we can get a modified Golden Crown value (mGC). In the present paper, we deal with the method of the allocation on the basis of previous studies.

2.2 Mathematical and Informatical Methods for Land Consolidation

The spread of information technology in the 1960s provided a new prospective for land consolidation, as well. In Germany, graphical data processing (e.g. David) and database management (e.g. Oracle) programs became the means of computer-aided planning. In 1984 the first CONEF (COmputerunterstützte NEuverteilung in der Flurbereinigung) land consolidation program package was developed. The program automatically processed alphanumerical data; however, the graphical data were processed manually. In 1990 at the Technical University of Munich the Chair of Land Management, the further development of CONEF, the CARE (Computer Aided Reallotment) was improved, which could manage both the regional data and the numerical demands of the owners.

Computer-aided land consolidation has antecedents in Hungary, as well. The task has already been approached by mathematical programming [10], combinatorial modelling [11] and Cluster Analysis [12]. According to the present operative law, the land consolidation module of the DigiTerra Map software is the most appropriate, which is based on voluntary land exchange institution. However, the original land structure is unchanged; the software re-allocates the ownership rights in a way that they are distributed to the nearest centroid of the previously owned land parcels. To implement the approach, it is essential to weigh land parcels in advance. “During scalability, three parameters are analysed: the owners’ previous land property, the distance of the nearest centroid, the rate of the distance of the second nearest and the furthest centroid. Therefore, the program provides a classification order, on the basis of which it rates parcels to the nearest centroid.” [12]

3 The DigiTerra Map Land Consolidation Module

3.1 The Work of the Module

The module carries out the planning based on three parameters (own land area, the distance of the first district, the distance of the second district) that can be weighted one by one. The weighting of the parameters can be fulfilled in various variables and the operation can be iteral. The coherence among the weighting of parameters, the number of iterations, the number of the parcels and owners of the planning should be defined numerically. The following analysis has been performed with the use of a generated test area containing 100 land parcels.

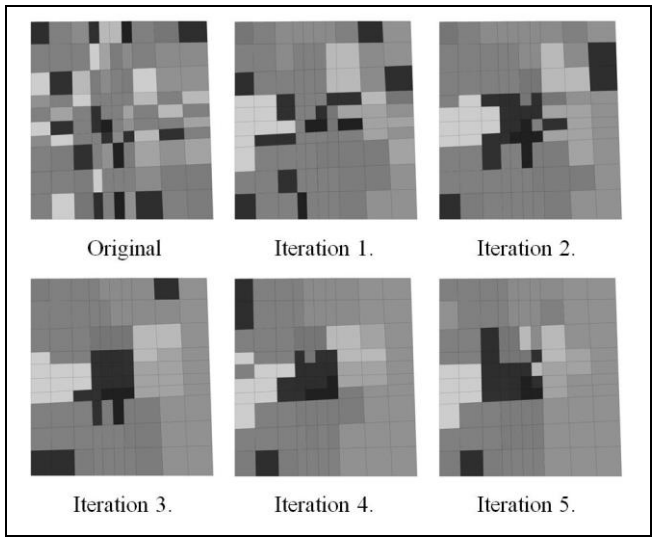


Figure 1
Distribution in 5 iterations with the help of the DigiTerra land consolidation module

Simmon's land parcel index, c.f. eq. (1) can show the efficiency of the planning. Fig. 1 displays the analysis of the sample with 20 owners in 5 iterations. The weighting of parameters: 1.0, 10.0 and 1.0, respectively for the own land area, the distance of the first district, the distance of the second district, and the district distance is 10000 m. During the planning, Simmons's land parcel index was determined by the combination of land parcels distributed next to each other. If an owner has not been given an area, the index cannot be counted, so its value is 1. The analysis shows that the highest degree can be weighted during the very first allocation. Some improvement can be observed, but not consequently, since its value is below some of the previous values (Fig. 2).

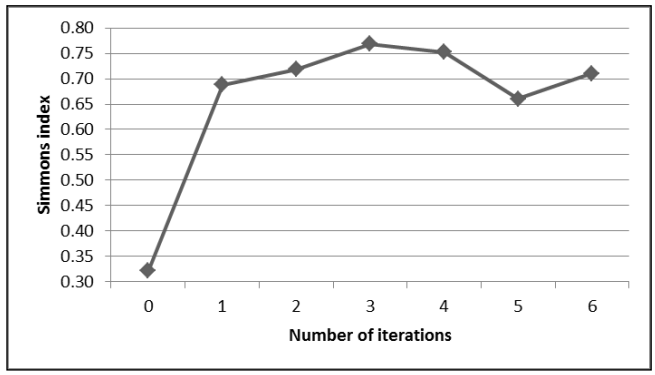


Figure 2
The coherence between the iterations and the fragmentations

There is no unambiguous connection between the shape factor of the land parcels and the number of the iterations. It is proven by Student's t-test, which says that the connection between the two data sets is $t = 4.209 > t_{0.05} = 4$, so they are found to be different from each other with 95% probability.

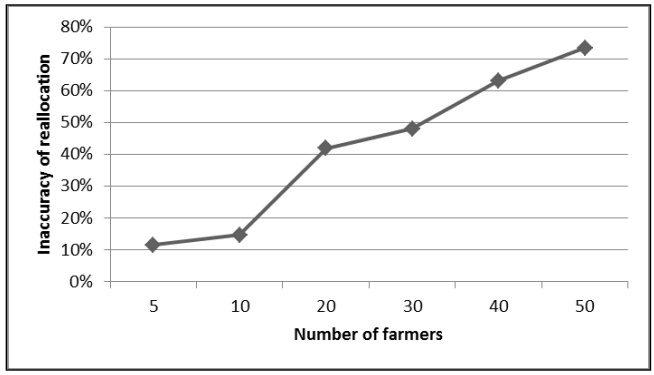


Figure 3
The inaccuracy of the allocation according to the number of the owners

As for the new land structure, an important aspect is value correspondence, in other words, the accurate allocation of the determined values. This accuracy is influenced by the number of the owners (Fig. 3) among other factors. The percentage accuracy of the distribution has been analysed in an area containing 100 parcels according to 6 different owners.

The positive correlation can be observed between the inaccuracy of the allocation and the number of the owners, the correlation coefficiency is r=0.984. A similar coherency can be observed between the number of the iterations and the inaccuracy of allocations (Fig. 4).

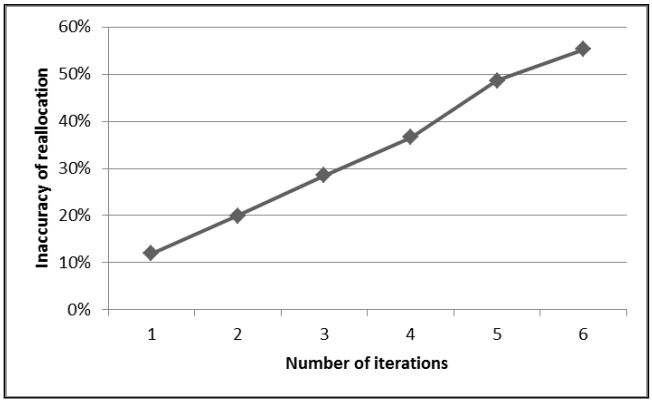


Figure 4
The accuracy of the allocation according to the iterations

The weighting of parameters can significantly influence the result (Fig. 5). To prove this, three different weighting variables have been analysed: 1) parameter: own land area, 2) parameter: the distance of the first district, 3) parameter: the distance of the second district). Noticeably, among the analysed variables, the bigger-than-average weighting of the own land area is the worst, while the most favourable result was given by the above-average weighting of the second district.

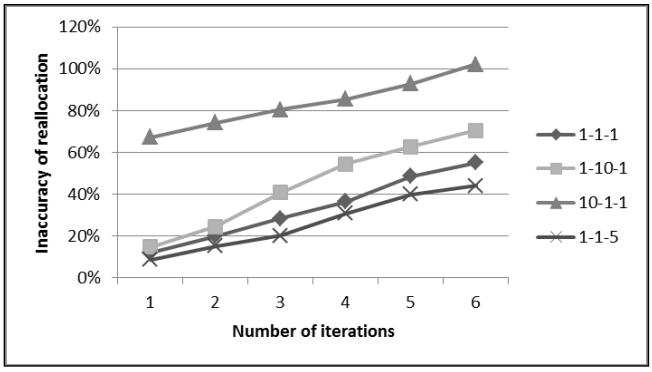


Figure 5
The inaccuracy of the allocation according to the weight of parameters

To plan optimally, it is not practical to carry out a huge number of the iterations since it can reduce the accuracy of allocation. However, it does not raise the fragmentation and the shape factor value of the parcels considerably. Since the sample data is not articulated by linear infrastructure, the results are true in general, as well.

3.2 The Development of the Module

Applicable conditions of the development:

- to keep the exact rate of ownership in the course of allocation

  The software does not modify the land structure but it assigns new owners to the existing land parcels. The accuracy of reallocation can have a maximum value that is correlated to the exchange value of the land parcel with the smallest value. Therefore, owners with few land parcels can be given areas whose exchange value is significantly different from the original. The number of iterations worsens the value protection as well. The present work of the module allows the repetition of the allocation. It is possible to boost land concentration by growing the number of iterations, on the contrary, the difference between the initial and the final value in exchange will be extensive. The accuracy of re-allocation can be grown by parcelling the area for units. This operation can be carried out in land parcels or tables. The latter can be supported only in the case of institutionalised land consolidation since new land structure and land parcel boarders come into existence.

- to give the demands of the owner

  Based on the present work of the module, the coordinates belonging to the owners are determined so that the module counts the centroid of the areas that belong to the same owner. This method is objective; however, it does not necessarily give an optimal solution. For instance, if earlier an owner lived further from the place of production, this problem is still present in the new land structure. To solve this problem, instead of centroid dispersion, the farming premises and the place of living may be defined as a preliminary condition. This method would be in line with the operative land law. According to this law, residents would be preferred for the purchase of land.

As a summary, the following hypothesis can be formulated: based on the owners' aspects, giving the farming premises and lamellating the area would be more accessible and favourable.

4 Planning on the Sample Area

4.1 Sample Data

We would like to carry out the analysis both in mountainous and plain areas, but the Ministry of Agriculture has merely provided the data of Mesterszállás for the

analysis. This village is on the Great Hungarian Plain. It is 42.92 km2 and its population is 702 people. 508 land parcels have been involved in the planning. Out of these 508 land parcels 182 owners own 335 land parcels with 1/1 ownership proportion; 245 owners own 173 land parcels with undivided joint ownership. On the sample area 424 natural people and 3 legal people are registered into Land Registry. According to the modified Golden Crown (mGC) [9], land size distribution is displayed in Table 1. The size of the area is 67578.05 mGC, the average size of the area per person is 158.26 mGC.

The relevant authority has given the digital and certified cadastral maps of the sample area. The Department of the Land and Geoinformatics of the Ministry of Agriculture and Regional Development has provided the data of the owners and land users free of charge with the condition that the personal data can be accessed only in encoded format. The encoding of owners and users was carried out with a 6-digit-tag. Since the encoding was fulfilled independently, they cannot correlate with each other. The owners' data are formed into an .xls format with 11columns and 2716 rows. The attributes of the statement: location, profile number, the owner's tag, legal status, counters of the ownership interest and denominator of the ownership interest. The land user data table contains the following information: location, profile number, land usage, quality class, used area and the user's tag.

According to the licence of the Ministry, the Institute of Geodesy, Cartography and Remote Sensing has provided the following mapping data:

- the settlement boundary of the municipality
- parcel boundary in the municipality
- profile numbers of the parcels
- the boundary of land usage
- building boundary
- labels of the cadastral map
- 1:10 000 topographical map
- aerial photograph without deformation
- relief model
- 1:50 000 land cover
- 1:100 000 land cover
- MePAR 2012 block map
- MePAR 2012 thematic layers

4.2 Planning

In the course of planning, more alternatives have been carried out. The number of parcels and the owners are almost equal, so planning based on parcels and ownership proportion cannot provide an appropriate solution. According to the present land policy, it is legitimate to establish a claim to a planning based on land usage. Land lease – along the farmers' demand – basically means use-based land

concentration in the sample area as well. The number of land users in the sample area (83 pieces; abbreviated as "pcs") is much lower than the number of the owners (427 pcs), so the situation is conspicuous.

The planning was carried out based on two variables, in four ways. The very first variable is the base unit of the allocation, which can be the land parcel or the lamella (1 ha unit area). The second variable is the starting point of the allocation which can be the centroid of the area belonging to the owner or a freely-given premise coordinate. Dividing the land parcels into unit areas has resulted in increasing the number of the areas (508 to 3728) that can be used in planning. Furthermore, the accuracy of allocation can get better. Lamella based planning is displayed in Fig.6.

Figure 6
Lamella based planning on the sample area

During land parcel based planning, the sample area was allocated according to the mGC values determined in a previous study [9]. During lamella based planning, the evaluation was carried out again, but the shape and size factors were not present due to the artificially produced shape of the lamellas.

4.3 Results

To compare the planning versions, the land parcel indices (Simmons, Januszewski and Igozurike) and the accuracy index of the allocation were used (Table 1).

It turns out from the analysis that all of the planning versions have more favourable index numbers compared to the preliminary situation. (While for Simmons and Januszewski index 1 stands for the best value; according to Igozurike, index 0 means the best one.) According to the indices, there is no unambiguous difference between the land parcel based planning and lamella based

planning. According to the Simmons-index, lamella based planning is favourable, but Igozurike finds land parcel based planning to be better. The accuracy of allocation is more favourable especially in lamella based planning, but the number of parcels per farmer is lower in the case of land parcel based planning.

Table 1
Indices of the planning versions carried out on the sample are

	Starting point	Planning based on parcel		Planning based on lamella	
		with centroid	with coord. of estate	with centroid	with coord. of estate
Number of parcels	508	309	243	497	325
Parcel per capita	6.120	3.723	2.928	5.988	3.916
Index according to Simmons	0.678	0.720	0.750	0.725	0.756
Index according to Januszewski	0.749	0.785	0.814	0.764	0.791
Index according to Igozurike	1.081	0.600	0.299	0.942	0.531
Accuracy of distribution		0.994	0.919	0.999	0.996

The hypothesis, which says that lamella based planning using premise coordinates is more optimal than land parcel based planning with the use of centroid coordinates, has been proven by the analysis. Considering the accuracy of allocation and the indices of fragmentation, the lamella based planning has been accepted.

In terms of the evaluation of the developed method it would have been useful to test different kinds of methods on the sample area. However, there were not any possibilities for it due to the lack of IT support.

In order to take the personal demands of the farmers and the importance of their participation in land consolidation into account, the accepted planning version has been worked on with the use of shape and size factor and mGC. The data was normalised in a way that all shape and size factors vary between 0 and 1. The need of normalization arose from the different number of parcels. The area calculation under the normalised function was carried out by numeric integration the results of which:

$$T_{bLC} = \int_0^1 f_{bLC}(x) = 0{,}504 \tag{6}$$

$$T_{aLC} = \int_0^1 f_{aLC}(x) = 0{,}529 \tag{7}$$

It turns out from the analysis that after the change the shape factor of the areas is cumulatively more favourable. However, the narrow parcels have come into existence in limited number as well. The formation of such parcels can be

improved by the optimal spatial distribution of premises (e.: minimalizing the centroid of the areas distributed into one block).

The mGC values counted before and after the change can contribute to the persuasion of the farmers and the effectiveness of land consolidation. The grade mGC values are shown in Fig. 7, in logarithmic scale. To compare the different data, it is not necessary to normalize the values because mGC values are not specific. The analysis needs cumulated summation.

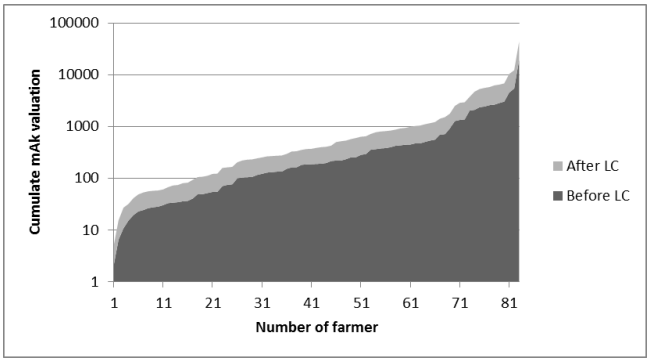


Figure 7
Lamella based planning on the sample area

The results of the summation:

$$\sum_{i=1}^{n} mGC_{bLC} = 67578.05 \tag{8}$$

$$\sum_{i=1}^{n} mGC_{aLC} = 81771.07 \tag{9}$$

It turns out from the analysis that the cumulated mGC value of the sample area has grown by 21% which is mainly as a consequence of the shape factor. Based on the land fragmentation indices and the mGC values (cumulated by shape and size factors), it could be said that the land consolidation was successful.

Conclusion

The present paper has dealt with the problems and solutions of land consolidation, prioritizing Geoinformatics. The method of land fragmentation calculation, the necessary data of planning, and the analysis of the DigiTerra land consolidation module and further development opportunities have been introduced. Land consolidation has been carried out on a sample area; the results have been proven by land fragmentation indices, the shape factor and mGC calculations. According to the analysis, the planning on the basis of lamella with coordinate determination has been proposed. The results have proved the efficiency of the proposed method

since it resulted in a 21% increase of the cumulative mGC of the sample area. The validation of the method can be achieved by further tests on differently scattered areas. The methodology can be flexibly shaped according to the participants' demands since the aim is their satisfaction. Therefore, land consolidation can contribute to the competitive production, the sustainable production, and to the correction of the quality of life in rural areas.

Acknowledgement

The research was supported by the Department of the Land and Geoinformatics of the Ministry of Agriculture and Regional Development Hungary. We thankfully, acknowledge the Óbuda University Alba Regia Technical Faculty.

References

[1] J. Ripka: Az agrártárca hosszú távú földbirtok-politikája, Háttéranyag, Vidékfejlesztési minisztérium Földügyi Főosztály, FF/1744/2011, 11 p

[2] A. J. Simmons: An index of farm structure, with a Nottingahmsire example, East Midlands Geographer, 3. 1964, pp. 255–261

[3] F. Dovring: Land and labour in Europe in the twentieth century (3rd ed.) Hague: The Nijhoff. 1965

[4] D. Demetriou: The Development of an Integrated Planning and Decision Support System (IPDSS) for Land Consolidation, Doctoral Thesis accepted by the University of Leeds, UK, ISBN 978-3-319-02346-5, 2014, 351 p

[5] J. Januszewski: Index of land consolidation as a criterion of the degree of concentration. Geographia Polonica 14, 1968, pp. 291–296

[6] M. U. Igozurike: Land tenure, social relations and the analysis of spatial discontinuity. Area, 6, 1974, 132–135

[7] R. Keeney, H. Raiffa: Decision with multiple objectives: Preferences and nvalue trade-offs. Cambridge: Cambridge University Press. 1993

[8] A. Sharifi, M. Herwijnen, W. Toorn: Spatial Decision Support Systems. Lecture Notes. ITC, International Institute for Geo-Information Science and Earth Observation, The Netherlands. 2004

[9] J. Katona: The Application of Fuzzy Logic in the Field of Land Consolidation, 10th International Symposium on Applied Informatics and Related Areas (ISBN:978-615-5460-49-4), P15. 2015, 4 p

[10] P. Gáspár: Birtokrendezési feladatok megoldása matematikai programozással. Budapesti Műszaki és Gazdaságtudományi Egyetem Általános- és Felsőgeodézia Tanszék, 2003, 12 p

[11] M. M. Csordásné: Matematikai modell a birtokrendezés támogatására, Geodézia és Kartográfia, Budapest, 57. évf. 2. szám. 2005, pp. 24-30

[12] K. Czimber: Képfeldolgozási és geoinformatikai algoritmusokon alapuló birtokrendezési eljárás kifejlesztése, Geodézia és Kartográfia 65. évf. 2013/11-12. pp. 15-18

Gearless Micro Hydropower Plant for Small Water-Course

Yury Dementyev[1], Roman Kuzmin[2], Aleksandr Serikov[2], Viktor Suzdorf[2], Kirill Negodin[1], Istvan Vajda[3]

[1]Tomsk Polytechnic University, Institute of Power Engineering, av. Lenina 30, 634050 Tomsk, Russian Federation , e-mail dementev@tpu.ru; knn1@tpu.ru

[2]Komsomolsk-na-Amure State Technical University, Electrotechnical Faculty, av. Lenina 27, 681013 Komsomolsk-na-Amure, Khabarovsk region, Russian Federation, e-mail epapu@knastu.ru; kepapu@knastu.ru; em@knastu.ru

[3]Óbuda University, Kandó Kálmán Polytechnic, Bécsi út 96/b, 1034 Budapest, Hungary, e-mail vajda@uni-obuda.hu

Abstract: The paper focuses on problem of development of autonomous power-supply systems based on micro hydropower plants, which are using small watercourse power. The design and development of such systems is influenced by a number of conflicting objectives. The power source has to generate ac voltage with steady-state magnitude and frequency and, at the same time, it has to be fairly simple and inexpensive. One of the future-proof designs that provides fulfillment of the above mentioned requirements is a gearless micro hydropower plant with a combined impeller of axial-flow turbine and an electric arc-shape inductor generator. The authors have identified how geometrical parameters of the arc-shape inductor generator influences the machine operation factors. In addition, they have found that the air gap impacts the ripple factor significantly. Finally the paper shows functional dependence of the slot chamfer factor on chamfer angle, which simplifies the problem of choosing reasonable, in terms of efficiency, design parameters of the generator for the micro hydropower plant

Keywords: micro hydropower plant; arc-shape inductor generator; form factor; design solutions; parameter optimization; ripple factor; chamfer angle

1 Introduction

During the 18th, 19th and the first half of the 20th Century, water wheels were important hydraulic energy converters. It is estimated that in England 25,000-30,000 wheels were in operation around 1850; in Germany 33,500 water wheels were recorded as late as 1925. Today, only very few water wheels are still in use.

Low power hydropower is seldom exploited since cost-effective energy converters for these conditions are not available [1]. Design of autonomous power-supply systems for lowland rivers with small watercourse power is carried out by solving the whole range of conflicting problems. The power source has to generate ac voltage with steady-state magnitude and frequency and, at the same time, it has to be fairly simple and inexpensive. One of future-proof designs that meets the above-mentioned requirements is gearless micro hydropower plant with combined impeller of axial-flow turbine and electric arc-shape inductor generator [1]. An advantage of propeller-type axial flow turbines is maximal specific speed for low heads, which allows for the development of a gearless micro hydropower plant. Hydroturbine in a river with low flow rate is placed on floats in order to be able to adjust the depth of the impeller immersion into water, so it does not have negative impact on the environment, including on spawning rivers. Thereby, the problem of designing electric power supply systems based on gearless micro hydropower plants for lowland rivers is topical [2].

Simplified design of hydroturbine in lowland river is shown in Fig. 1.

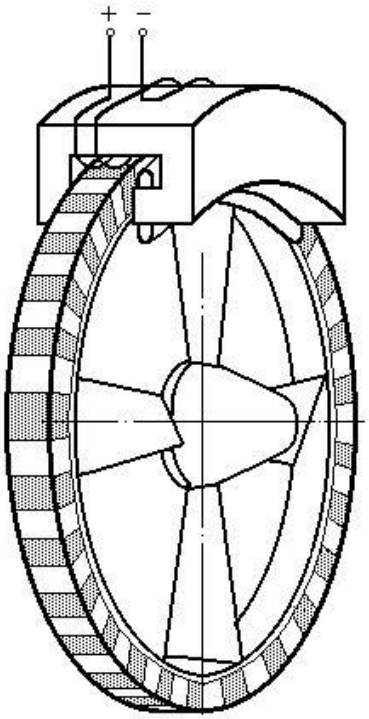


Figure 1
Hydroturbine for lowland river

Simulation of such a complex technical object as a micro hydropower plant is carried out based on generally accepted assumptions. The simulation outcomes should indicate characteristics of efficiency and other parameters of the device performance quality. Initial parameters that determine all the simulation factors are the turbine diameter, blade angle, water course velocity [3].

Simulation model studies have shown the main relationships of the design parameters on parameters of the water course. Functions shown in Fig. 2 compose 3D characteristic «Power of hydroturbine, P_G – water course velocity V_w - turbine wheel diameter D_w».

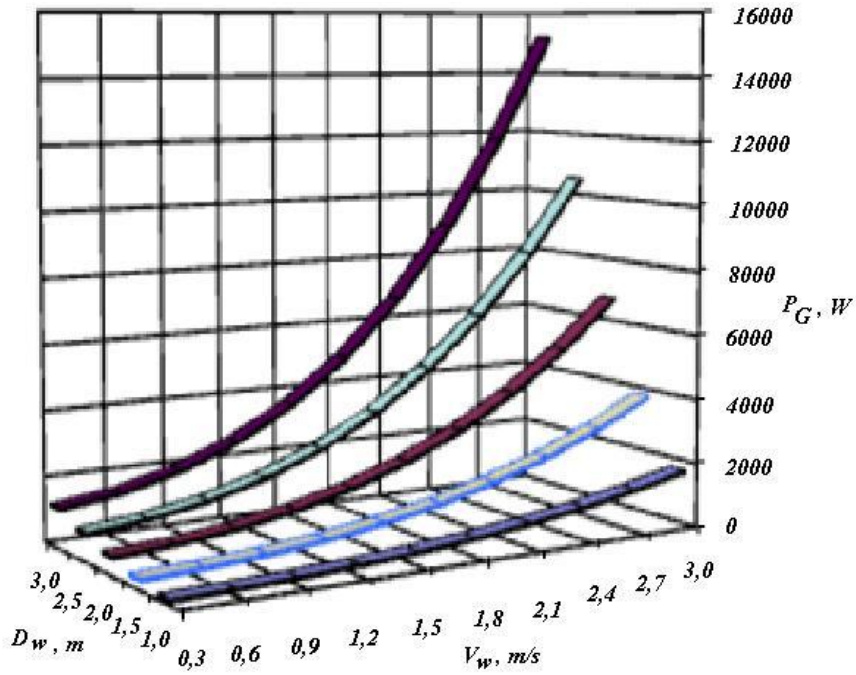


Figure 2
Functions «Power of hydroturbine – water course velocity - turbine wheel diameter»

2 Design and Calculation of the Arc-Shape Generator

The source of electric energy is a generator of special developed design, which determines all other parameters of the system. Therefore, it is important to predetermine static and dynamic characteristics of the source based on the generator in the designing phase [3]. So, we need to develop adequate

mathematical model of an electric arc-shape inductor generator of a special developed arc-shape design.

Structure features of the inductor generator with an arc-shape stator influence on the form of magnetic field in the air gap demands corresponding analysis to determine dependencies of parameters of the generator and the field harmonic composition as well as the losses on higher harmonics.

Magnetic induction distribution in the air gap of synchronous electric arc-shape inductor generator with electromagnetic excitation is described by an equation set of the stationary magnetic field [4]. One of main approaches to its solution is finite element method (FEM).

Constructively the magnetic core is made of laminations, that is why at the stage of mathematical description of the generator magnetic circuit it is convenient to use the projections of magnetic permeability on two axes (Y,X) that correspond to longitudinal and transversal lines of the iron rolling.

Considering non-saturated magnetic circuit of the generator, the following equations [5] can be used:

- Magnetic permeability in Y-axis (along rolled sheet), H/m, is determined in terms of formula:

$$\mu_Y = \mu_{ir} \cdot K_L ,$$

where μ_{ir} is relative permeability of iron; K_L is lamination factor.

- Magnetic permeability in X-axis (across rolled sheet), H/m, is determined in terms of formula:

$$\mu_X = \frac{(2 - K_L) \cdot \Delta_{ir}}{\frac{\Delta_{ir}}{\mu_{ir}} + \frac{\Delta_{ir} \cdot (1 - K_L)}{\mu^{ë}_0}} ,$$

where Δ_{ir} is thickness of rolled sheet, m;

$\mu^{ë}_{ir} = 1$ is value of relative permeability of iron.

- Magnetizing force in the air gap of the generator is determined by:

$$F_\delta = \frac{B_\delta}{\mu_0} \cdot \delta ,$$

where δ is value of air gap, m.

- Magnetic potential difference in the air gap between stator and rotor is given by:

$$U_m = F_\delta .$$

Boundary conditions, which are taken into account to solve the field problem, are the following [4] :

- boundary conditions of the first kind on external (upper and lower) borders of the simulated area (homogeneous) (see Fig. 3)are given by:

$$U_{m\lfloor B \rfloor} = const\,.$$

- boundary conditions of the second kind on external (left and right) borders of the simulated area (see Fig. 3) are determined in terms of:

$$\frac{\partial U_m}{\partial n}_{\lfloor B \rfloor} = 0\,.$$

The last condition is true, when moving away the borders of the area for a considerable distance from the field source.

It is necessary to maintain the continuity condition of magnetic scalar potential and equality of normal and tangential derivatives on the interfacial area. In finite element method these conditions are met automatically. Computational area of the magnetic field studies with boundary conditions is shown in Fig. 3.

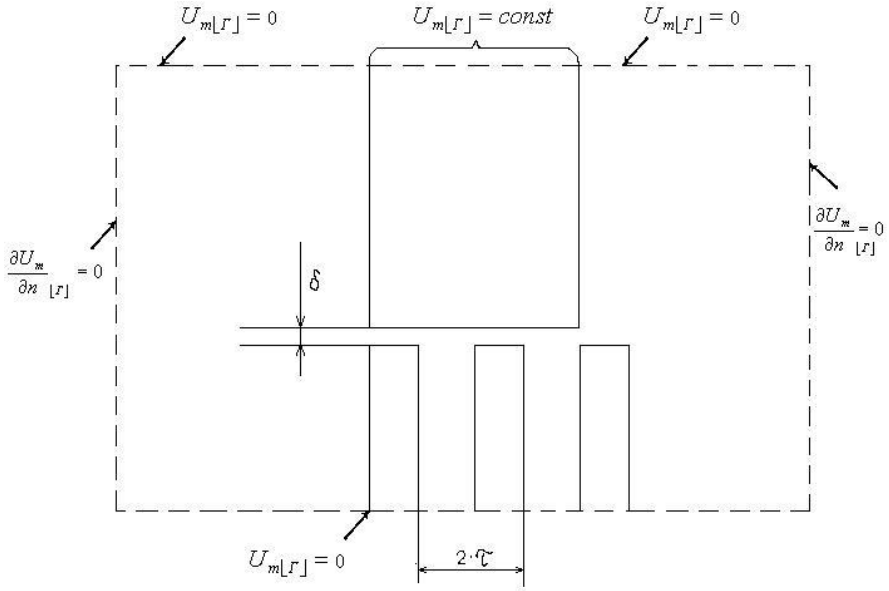


Figure 3
Research area of the magnetic field

3 Results of Simulation Studies

3.1 Geometrical Parameters Impact on Machine Performance

Curves of induction distribution, which have been obtained as a result of computational simulation of field in the air gap (the rotor tooth shape is assumed to be rectangular), are shown in Fig. 4.

Simulation of the magnetic field parameters in an electric machine in order to find out the qualitative and quantitative evaluation of the induction distribution in the air gap also allows to carry out its harmonic analysis.

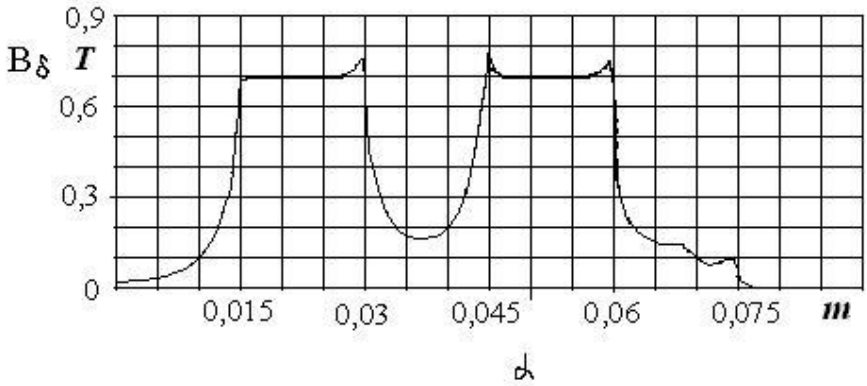


Figure 4
Curve of magnetic induction in the air gap (for rectangular rotor tooth shape)

Variable algorithm of searching the optimum shape of the curve of magnetic induction distribution in the air gap allows to define the following:

- Step-by-step synthesis of the pole shape, as it is shown in Fig. 5;
- Variation of the pole factor $\alpha_{\ddot{I}} = b_p / \tau$,

  where b_P is pole span, m; τ is pole pitch, m.

Estimation method of shape factor K_S of the air gap under stator pole is illustrated by Fig. 5 and relationship given by:

$$K_S = S_M / S_0 .$$

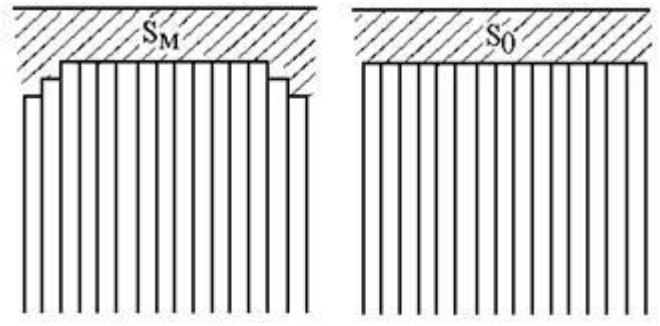


Figure 5
Estimation method of shape factor K_S of the air gap under stator pole

Harmonic composition of the induction distribution curve along the rotor surface is determined by the following factors:

1. Form factor of variable component of the magnetic field of excitation for ν- th harmonic is determined in terms of formula:

$$\hat{E}_{f\nu} = \frac{B_{\delta m\nu}}{B_m},$$

where $B_{\delta m\nu}$ is peak value of the magnetic induction harmonic with number ν in the air gap, T;

B_m is peak value of magnetic induction in the air gap on the axis of the rotor pole, T.

2. Utilization factor of the magnetic field is given by [5]:

$$K_{\grave{e}} = \frac{B_{\delta 1m}}{A(0)/2},$$

where: A(0)/2 is zero harmonic of the magnetic field in the air gap of the machine, T;

$B_{\delta 1m}$ is peak value of first harmonic of magnetic induction in the air gap, T

Based on outcomes of computational simulation the influence of the required factors on the magnetic induction distribution can be estimated to vary the pole shape of the machine. One of the variants of the form factor and the utilization factor subject to geometry of the tooth zone is shown as graphs in Fig. 6-7.

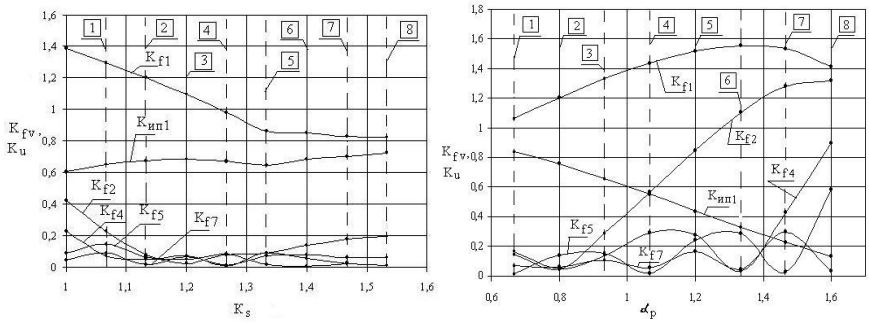


Figure 6

Functions of form factor and the utilization factor subject to geometry of the tooth zone

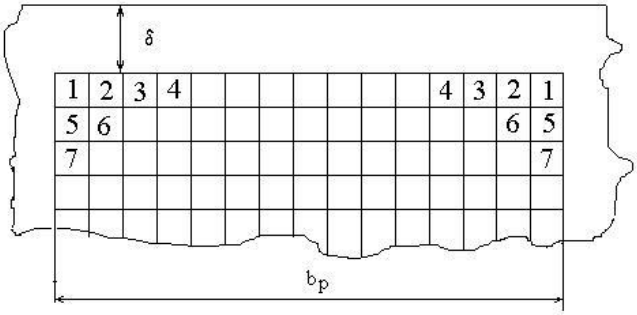


Figure 7

Diagram of shape variation sequence in order, indicated by numbers, in which the pole sheets are cut out

The simulation model study findings have shown significant influence of the air gap value on ripple of the generator magnetic field. Numerical values of ripple factor K_P, subject to air gap, are shown in table 1.

Table 1

Dependance of ripple factor on air gap

Air gap, m.	Ripple factor Kp
0.002	1.1
0.003	1.141
0.004	1.294

3.2 Chamfer Factor Influence on the Generator Operation

Design of the generator, where stator is made up in the form of an arc, has both the rotor tooth zone and stator slots with arc-shape geometry. When the rotor teeth are located radially, the stator slot axes are in parallel [6]. The axial matching of

stator and rotor teeth, which are located in the middle of the arc, should be noted. When moving along the rotor tooth axis to the edge of the arc, the slot chamfer angle is increasing (Fig. 8, 9). It means that the slot chamfer angle is not a constant, as in standard ac machines, but a variable that is altered 0 to its maximum value [7]. Hereby the influence of the chamfer on the EMF of the armature winding should be studied.

The study of the dependence of the chamfer factor on fundamental harmonic of magnetic induction shown in Fig. 8 allows to obtain localized zone of permissible ratios between number of poles and pole arc angle for edge slots of stator for $K_C \geq 0,7$.

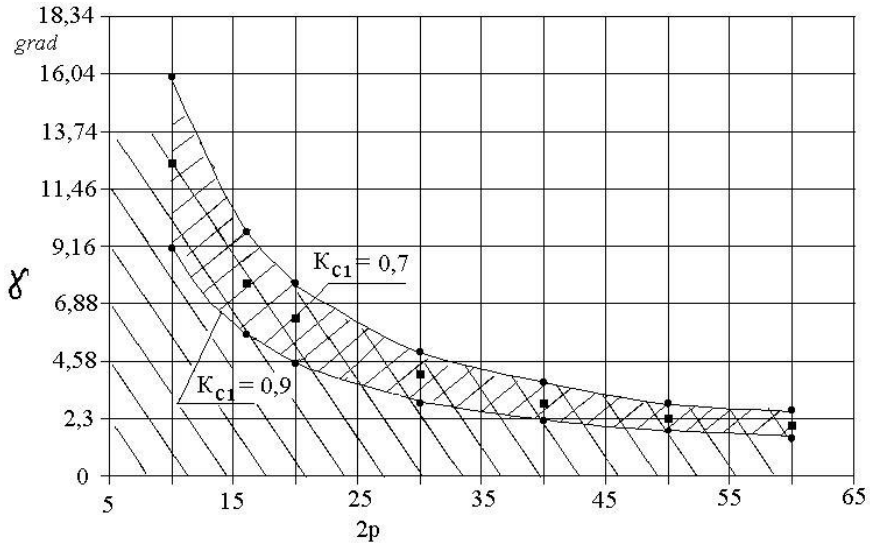


Figure 8
Zone of permissible ratios between number of poles and pole arc angle

EMF $\mathring{A}_q$ of a coil group is determined by adding together the EMF vectors $\Delta\mathring{A}_{\hat{e}}$ of the coils that are shifted in space for angle $\gamma_{\tilde{n}\hat{e}_i}$ (Fig. 10).

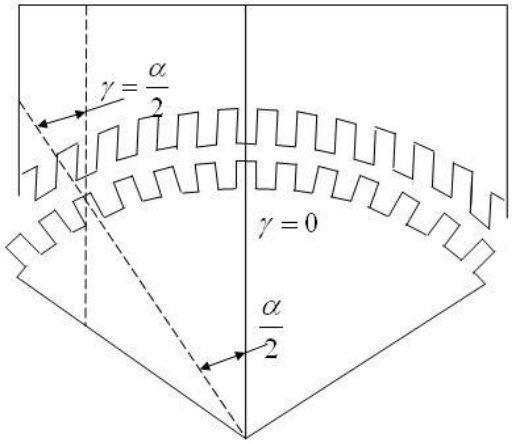


Figure 9

Geometrical interpretation of slot chamfer factor

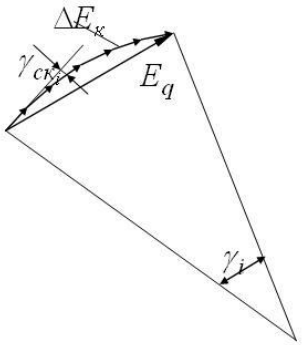


Figure 10

EMF vector of a coil with slot chamfer

Expressions to calculate the slot chamfer factor illustrated by Fig. 9, 10 are given in terms of formulae below:

$$\gamma_{ci} = \frac{\alpha}{2z_i}, \quad 0 \le \gamma_i \le \frac{\alpha}{2}, \quad K_c = \left. \frac{\sin\frac{\alpha}{2}}{\sin\frac{\gamma_i}{2}} \right|_{\gamma_i = \frac{\alpha}{2z_i}}$$

$$K_c = \int_0^{\alpha/2} \frac{\sin\frac{\alpha}{2}}{\sin\frac{\gamma_i}{2}} d\gamma_i = \sin\frac{\alpha}{2} \int_0^{\alpha/2} \frac{dx}{\sin\frac{x}{2}} = 2\sin\frac{\alpha}{2} \int_0^{\alpha/2} \frac{dz}{\sin z}$$

$$K_c = 2\sin\frac{\alpha}{2} \ln\left(\cos ec\, z - ctg\, z\right)\Big|_0^{\frac{\alpha}{2}}$$

The achieved dependences of the slot chamfer factor on chamfer angle are shown in Fig. 11. Approximating function can be specified by 3-rd order poly-nomial determined by:

$$K_{ci} = -0,0607\,\gamma_i^3 - 11,996\,\gamma_i^2 + 0,0039\,\gamma_i + 0,7839.$$

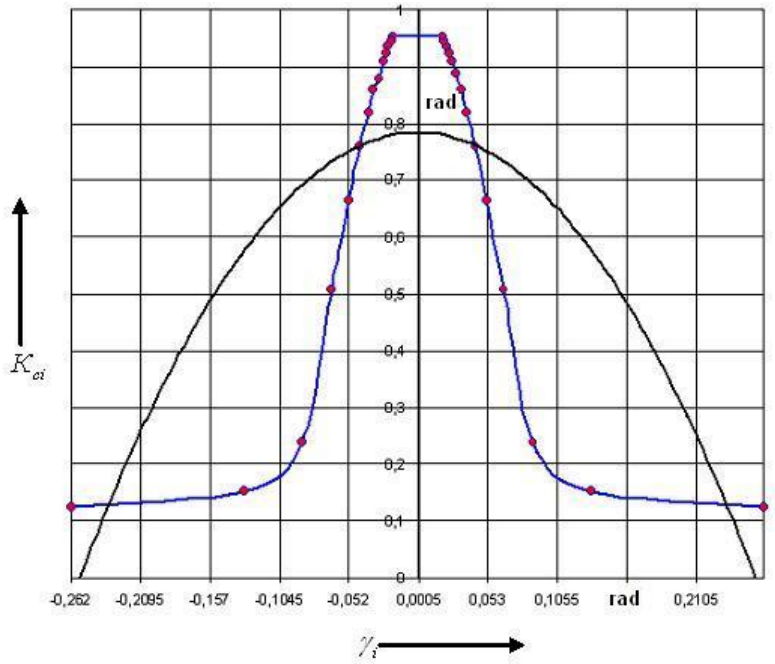


Figure 11

Dependences of the slot chamfer factor K_{ci} on chamfer angle γ_i

Conclusion

Research shows qualitive and quantative influence of design features of an arc-shape stator, rotor poles, geometry parameters of the air gap on spectral composition and power efficiency of the magnetic field. It allows determining optimal range of variation of the slot chamfer and dependence of first harmonic of EMF on it.

Analysis of the obtained calculated and simulation data shows that the generator output parameters (form factor, utilization factor, ripple factor) are influenced by geometrical parameters of the machine. In addition, the impact of the the air gap on the ripple factor is found to be significant.

Functional dependence of the slot chamfer factor on chamfer angle has been found, which simplifies the problem of choosing reasonable, in terms of efficiency, design parameters of the generator of a micro hydropower plant.

Acknowledgments

The research is funded from Tomsk Polytechnic University Competitiveness Enhancement Program grant, Project Number TPU CEP_IPE_97\2017.

References

[1] Müller, G.a , Kauppert, K.b Performance characteristics of water wheels. - Journal of Hydraulic Research. Volume 42, Issue 5, 2004, Pages 451-460.

[2] Anagnostopoulos, J.S. , Papantonis, D.E. Optimal sizing of a run-of-river small hydropower plant. - Energy Conversion and Management. Volume 48, Issue 10, October 2007, Pages 2663-2670.

[3] Müller, G., Denchfield, S., Marth, R., Shelmerdine, R. Stream wheels for applications in shallow and deep water (2007) Proc 32nd IAHR Congress, Venice C (2c, Paper 291).

[4] Generator for micro power plant// V. M. Kuzmin, G. A. Sedov, V. I. Suzdorf /Patent 39918 of Russian Federation, MPK H02P7/29. publ. in bull. №23, 2004 (in Russian).

[5] Suzdorf V.I. Simulation of voltage sources for autonomous power supply system. Education and science: actual state and outlook// Proceedings of conference NTK 31st of July 2014: in 6 vol. Vol.3. Tambov: «Consulting company Yukom», 2014.- p.128-130 (in Russian).

[6] Victor I. Suzdorf, Aleksandr S. Meshkov, Yuri N. Dementyev and Dmitriy A. Kaftasyev Energy efficiency improvement of medical electric tools and devices// The 2nd International Youth Forum "Smart Grids", MATEC Web of Conferences, Volume 19, 2014. -Published online: 15 December 2014.

[7] Quaranta, E. , Revelli, R. Performance characteristics, power losses and mechanical power estimation for a breastshot water wheel. – Energy. Volume 87, 1 July 2015, Pages 315-325.

The Comparison of Data Envelopment Analysis (DEA) and Financial Analysis Results in a Production Simulation Game

Tamás Koltai, Judit Uzonyi-Kecskés

Department of Management and Corporate Economics
Budapest University of Technology and Economics
Magyar tudósok körútja 2, 1117 Budapest, Hungary
koltai@mvt.bme.hu, uzonyi-kecskes@mvt.bme.hu

Abstract: Production and service systems are generally evaluated based on financial information. The financial approach looks for opportunities to boost profits in two main ways: by decreasing operating costs and/or by increasing production quantity. Consequently, the cost of operation is evaluated and cost reduction possibilities are explored with proper cost analysis methods. Scoring methods extend the frontiers of performance evaluation by also employing non-financial information, although these methods generally contain several subjective elements. Data Envelopment Analysis (DEA) aims to integrate several performance measures into an aggregate output measure and several resource usage characteristics into an aggregate input measure. Based on the inputs applied and on the outputs generated, an efficiency score is calculated using linear programming. The objective of this paper is to illustrate the differences between performance evaluations, based on financial information, versus the DEA results. The results of a production simulation game are used to show how a DEA based performance evaluation can be carried out. The additional information provided by DEA may help to identify the causes of inefficient operation and to explore ways of improving efficiency.

Keywords: Data envelopment analysis (DEA); Performance evaluation; Production management; Simulation games; Linear programming

1 Introduction

Evaluating the performance of production systems is one of the most important tasks for managers. Performance evaluation is especially complicated when several conflicting evaluation criteria must be considered at the same time. Profit, for example, is one of the major objectives of a production system. High customer satisfaction favorably influences profit in the long term. Spending on customer satisfaction improvement may, however, decrease profit in the short term. It is

difficult to evaluate these two conflicting criteria at any particular point in time. This is because the integration of the value of profit and the measure of customer satisfaction into a single score is subjective. Many other similar evaluation issues arise, in practice.

A specific example of performance evaluation is the analysis of the results of business simulation games used in management education and training programs. Simulation games are especially popular in the field of production and operations management. The beer game has been used for many years to study the bullwhip effect in supply chains, and a range of other logistic and/or manufacturing related games are in common use (see for example, Sterman, 1989; Ammar and Wright, 1999; Holweg and Bicheno, 2002; Battini et al., 2009). In the case of simulation games, the evaluation of the performance of the participating teams (or individuals) must be completed using special evaluation criteria related to the learning process which occurs during the game (Voss, 2015).

Generally, when evaluating the performance of systems on the basis of several evaluation criteria, scoring methods are employed. Scoring methods transform performance data into a common scale and an aggregate score is calculated with subjective weights. Data envelopment analysis (DEA) is a special type of scoring method. In DEA, weights are determined by means of linear programming (LP). Hence, the subjective judgment of the decision maker is eliminated when the efficiency scores are calculated. Data envelopment analysis evaluates the performance of decision making units (DMU) based on the outputs provided and on the inputs used by the DMUs. Thus, DEA determines the relative efficiency of DMUs based on the observed input and output values. A single efficiency score is calculated, and improvement policies are explored for non-efficient DMUs. Many DEA models exist in the literature, which aim to capture different real life operation and decision making environments (Cooper, Seiford and Tone, 2007).

The objective of this paper is to show how a slack-based DEA model can be used to analyze the performance of student teams in a production simulation game. The paper compares the financial results and the DEA efficiency scores of the participating teams. A correlation of the two results is analyzed and the differences are explained. The additional information provided by the DEA results are illustrated with several examples.

The structure of the paper is as follows. In Section 2, relevant literature related to DEA in financial analysis is reviewed. In Section 3, the main DEA concept and the basic models related to the presented research are reviewed. Section 4 introduces the objective and the basic conditions of the production simulation game used in the study. Section 5 compares and explains the financial and DEA results obtained. Finally, in Section 6, some general conclusions are drawn and further research possibilities are suggested.

2 Literature Review Related to the Application of DEA for Financial Evaluation

The three major sources of financial analysis are generally the Income Statement, the Balance Sheet, and the Statement of Cash Flow. When the characteristics of some economic sectors are analyzed or the performance of companies, are compared, financial ratios, using data of these three sources, are calculated. Sometimes simple ratios are used, such as return on assets or return on investment, but sophisticated systems of ratios are also found in practice (Harrison and Rouse, 2016).

The application of ratios is very widespread and accepted by practitioners, but there are also several criticisms (see for example Smith, 1990; Thanassoulis, Boussofiane and Dyson, 1999 or Harrison and Rouse, 2016). First, those ratios consider only two dimensions of operation, namely those which are described by the numerator and those by the denominator. It is possible to aggregate several ratios to incorporate more dimensions of the analyzed problem, but in this case the weights used for aggregation are subjective. Second, ratios generally provide an indication of efficiency problems, but a further analysis is required to trace the causes of inefficiencies. Both problems can be solved using DEA, which calculates an aggregate measure of efficiency and provides information about efficiency improvement possibilities.

Smith (1990) was one of the first to suggest the application of DEA to evaluate Financial Statements. He studied the efficiency of 47 pharmaceutical firms using average equity, average debt as inputs, and earnings available for shareholders, interest payments and tax payments as outputs, taken from their accounting system. The efficiency scores, calculated with an input oriented variable return to scale model, were compared with the return on capital ratios.

The efficiency of bank branches, belonging to a Turkish bank, were analyzed by Oral and Yolalan (1990). They applied two DEA models, one for analyzing profitability using financial information and one for analyzing service efficiency using operational information. They showed that DEA is not only complementary to the traditionally used financial ratio analysis, but also a useful tool for operations management decision making.

A similar study was conducted by Bowlin (1999), who compared the efficiency of the defense and non-defense related segments of the defense industry. Accounting information of 18 randomly sampled firms was used for the analysis of the trends of efficiency change between 1983 and 1992. The trends indicated by DEA were very similar to the trends indicated by some classic financial ratios.

Thanassoulis, Boussofiane and Dyson (1996) analyzed the perinatal care system in the United Kingdom. The efficiency of 189 units providing perinatal care was calculated with a radial model applying five inputs and five outputs, and the

results were contrasted with several officially used performance indicators. In this case not traditional financial ratios, but official performance indicators of the District Health Authorities were used for comparison.

These four examples are among the first published cases which intended to use accounting information for DEA. Since then, several papers were published with the objective of highlighting the possibilities of DEA for financial evaluation. (for example, Ferro, Kim and Raab, 2003; Fenyves, Tarnoczi and Zsidó, 2015; Ederer, 2015; or Hosseinzadeh et al., 2016).

The general conclusion of these applications is that if inputs and outputs are carefully selected, then DEA results generally do not contradict financial results, and furthermore, DEA provides direct information for improvement possibilities. Two problems, however, must be considered when DEA is applied for the comparison of financial performance of different organizations.

The first problem is the violation of homogeneity assumption (Dyson et al., 2001). When DEA is applied, it is important that DMUs undertake similar activities, produce comparable products and/or services, apply a common set of inputs and use similar technology. This assumption is partly overlooked in the previously mentioned cases. The pharmaceutical companies, analyzed by Smith (1990), do not produce identical products. In the defense industry analysis, even the separation of defense and non-defense segments of the activity is ambiguous, according to the author (Bowlin, 1999). The prenatal care units analyzed by Thanassoulis, Boussofiane and Dyson (1996) do not provide exactly the same services, and finally the services offered by the bank branches may also differ (Oral and Yolalan, 1990).

The second problem is related to the application of DEA results. In the previously mentioned cases the information of DEA concerning improvement possibilities are not generally traced back directly to the analyzed units. Only Oral and Yolalan (1990) mentioned that DEA can also serve as a bank management tool if the results are used for decision making related to future operation.

The main novelty of this paper is that in the presented application these two problems are solved. All participating teams in the production simulation game produce the same products, use the same inputs, and apply the same production technology. Only marketing, financial and operation decisions are different. Consequently, homogeneity assumption is perfectly satisfied. The results of DEA are used for performance evaluation of the students, that is, the results and improvement possibilities are directly traced back to the decision makers.

3 Basic Concepts and Models of DEA

Charnes, Cooper and Rhodes (1978) suggested a linear programming model for the comparison of Decision Making Units (DMUs) using relative efficiency measures. Based on the model they suggested, relative efficiency analysis, or data envelopment analysis (DEA), became an important research area and a useful tool for performance evaluation. Several applications of DEA models are reported in the literature in both the service and the production sectors (Doyle and Green, 1991; Panayotis, 1992; Sherman and Ladino, 1995). A frequently applied area of DEA is higher education. Johnes (2006) compared more than 100 higher educational institutions in England using a nested DEA model. Sinuany-Stern, Mehrez and Barboy (1994) analyzed the relative efficiency of several departments within the same university.

The model suggested by Charnes, Cooper and Rhodes (1978) can be explained by an intuitive analogy taken from engineering. According to the law of energy conservation while energy can be transformed from one form into another, energy cannot be created. In power plants, for example, it is not possible to produce more energy than the energy content of the fuel used, or, to expressed differently, the technical efficiency of a power station is always lower than 1. Applying this engineering analogy to the area of performance evaluation in operations management, it can be stated that the measure of output is always smaller than the measure of input. In the best possible case, the ratio of output measure and input measure is equal to 1. The output and input measures are calculated as weighted outputs and weighted inputs, and the best possible weight values are sought for a reference DMU R. Let us assume that J number of DMUs are evaluated, when K different outputs are observed and I different inputs are used. Notations applied in this paper are listed in Table 1. If y_{kj} (k=1,…,K; j=1,…,J) are the observed output values of output k, and x_{ij} (i=1,…,I; j=1,…,J) are the observed input values of input i for DMU j, while v_k (k=1,…,K) and u_i (i=1,…,I) denote the output and input weights, then the linear programming formulation for finding the most favorable weights for DMU R is as follows:

$$
\begin{aligned}
&Max\left(\sum_{k=1}^{K} v_k y_{kR} \Big/ \sum_{i=1}^{I} u_i x_{iR}\right) \\
&\sum_{k=1}^{K} v_k y_{kj} \Big/ \sum_{i=1}^{I} u_i x_{ij} \le 1 \qquad j = 1,\ldots,J \\
&u_i, v_k \ge 0 \qquad\qquad i = 1,\ldots,I; \quad k = 1,\ldots,K.
\end{aligned}
\tag{1}
$$

If problem (1) is transformed in order to eliminate the ratio of variables, and the weighted input is fixed (equal to 1) in order to obtain a unique solution for LP problem (1), then the primal version of the input oriented, constant return to scale (CRS) model is obtained, that is:

$$
\begin{aligned}
&Max\left(\sum_{k=1}^{K} v_k y_{kR}\right) \\
&\sum_{i=1}^{I} u_i x_{iR} = 1 \\
&\sum_{k=1}^{K} v_k y_{kj} - \sum_{i=1}^{I} u_i x_{ij} \le 0 \qquad j = 1,\ldots,J \\
&u_i, v_k \ge 0 \qquad\qquad i = 1,\ldots,I; \quad k = 1,\ldots,K
\end{aligned}
\tag{2}
$$

The dual version of problem (2), however, has more practical relevance and leads to another interpretation of DEA. According to the dual interpretation, any linear combination of the observed output and input values leads to a new and feasible DMU, which may exist in practice. The production possibility set is determined by all possible linear combinations of the observed outputs and inputs. If λ_j (j=1,…,J) are the coefficients of the linear combination of output and input values, then the production possibility set of DMU R can be defined as follows,

$$
\begin{aligned}
&y_{kR} \le \sum_{j=1}^{J} y_{kj}\lambda_j \qquad k = 1,\ldots,K \\
&x_{iR} \ge \sum_{j=1}^{J} x_{ij}\lambda_j \qquad i = 1,\ldots,I
\end{aligned}
\tag{3}
$$

If we consider the λ_j (j=1,…,J) coefficients as variables, and a proper objective function is used to get an optimal combination of the output and input values, then the distance of any existing DMU from the optimal DMUs can be the basis of the efficiency score. The dual version of the input oriented CRS model assumes that all inputs must be decreased to the same proportion (θ), and efficiency is given by the smallest value of this proportion. Consequently, the smallest amount of input necessary to produce the observed output must be determined. The corresponding dual LP model is as follows:

$$
\begin{aligned}
&Min(\theta) \\
&\sum_{j=1}^{J} \lambda_j y_{kj} \ge y_{kR} \qquad k = 1,\ldots,K \\
&\sum_{j=1}^{J} \lambda_j x_{ij} \le \theta x_{iR} \qquad i = 1,\ldots,I \\
&\lambda_j \ge 0 \qquad\qquad j = 1,\ldots,J
\end{aligned}
\tag{4}
$$

Models (2), (3), and (4) are based on a radial measure of efficiency, where all inputs are decreased proportionally by the same ratio. The slack based model (SBM) proposed by Tone (2001) uses the difference of the observed values and

the best possible linear combination of inputs and outputs. The difference between the actual value and the best possible value is called slack. All possible slack values of DMU R can be determined if (3) is completed with slack variables. In (5), s_k^+ indicates the degree to which output k can be increased and s_i^- indicates the degree to which input i can be decreased, thus:

$$\begin{aligned} s_k^+ &= \sum_{j=1}^{J} \lambda_j y_{kj} - y_{kR} \qquad k = 1,\ldots,K \\ s_i^- &= x_{iR} - \sum_{j=1}^{J} \lambda_j x_{ij} \qquad i = 1,\ldots,I \end{aligned} \tag{5}$$

The slack values express the distance of a DMU from the best possible DMU. Based on the slack values the following efficiency measure can be used,

$$\mu_R = \frac{1 - \sum_{i=1}^{I} w_i^- s_i^- / x_{iR}}{1 + \sum_{k=1}^{K} w_k^+ s_k^+ / y_{kR}} \tag{6}$$

The slack-based measure of efficiency proposed by Tone (2001) can take any value between 0 and 1, and it is based on the weighted average of the normalized input and output slacks. Depending on the orientation of the analysis, either the nominator or the denominator can be ignored in the objective function. The input-oriented approach applied in this paper uses the following objective function:

$$Min\left(1 - \sum_{i=1}^{I} w_i^- s_i^- / x_{iR}\right) \tag{7}$$

In Section 4, an input oriented, slack-based DEA model, using objective function (7) and production possibility set (5), is applied to evaluate the results of student teams in a production simulation game. Since all DMUs in the game start with the same initial conditions, and considerable size differences cannot be achieved during the game, a constant return to scale (CRS) model is appropriate.

Table 1
Notation

Indices:		
j	-	index of decision making units (DMUs), j=1, …, J
i	-	index of inputs, i=1, … I
k	-	index of outputs, k=1, …, K
R	-	index of the reference DMU
Parameters:		
J	-	number of DMUs
I	-	number of inputs
K	-	number of outputs
x_{ij}	-	quantity of input i of DMU j
y_{kj}	-	quantity of output k of DMU j
w_i^-	-	weight of input slack i
w_k^+	-	weight of output slack k
Variables:		
u_i	-	weight of input i
v_k	-	weight of output k
λ_j	-	dual variable of DMU j
θ	-	radial efficiency score
μ_R	-	slack based measure efficiency score of DMU R
s_i^-	-	vector containing the input surplus values of each DMU
s_k^+	-	vector containing the output shortage values of each DMU

4 Introduction of the Simulation Game Applied in the Experiment

The production simulation game applied in this paper was developed by EcoSim Ltd. to support education and training in the field of production management. This simulation game is used in a module entitled *Decision Making in Production and Service Systems,* on the Production and Operations Management Master's degree program at the Budapest University of Technology and Economics. The objective of the game is to simulate production management decision making in a car engine manufacturing factory. The factory produces three different car engines for five different markets. Each market has its own demand characteristics. The car engines are assembled from parts on assembly lines operated by workers. Decisions must be made by each student team for the next production period (year) in the following areas:

– *The production quantities of the three car engines*. Forecasts must be prepared of expected demand based on the known demand of several previous periods. The

expected demand, the available production capacity and the final product inventory information are used to determine the production quantities for the next year.

– Prices and payment conditions. Demand can be stimulated by changes to the selling price and by offering favorable payment conditions. Decisions must be made on the purchase price in the next production period and on the payment delay percentages offered to customers.

– *Quantities of parts to be ordered*. The order quantities of the various parts groups must be determined based on the planned production quantities, on the bill-of-material of the car engines and on inventory and financial information.

– *Number of workers, number of shifts, and quantity of overtime*. Production quantity is determined by the machine capacity and by the number of workers. In the short term, capacity can be changed by hiring or firing workers and by changing the number of production shifts, or by applying overtime. Decisions must be taken about the number of the workforce, about the number of shifts and about the quantity of overtime in the next production period.

– *Investments in the production line and in space*. In the long term, production capacity can be increased by investing in new production lines and in making more space available for production and for inventory. Decisions must be made in each production period about the number of new production line installations and about the number of square meters of space extensions.

– *Launch of efficiency improvement projects*. It is possible to launch projects which may improve production conditions. The predefined projects have different effects and different launch and maintenance costs. Decisions must be made on which projects to launch in a production period.

– *Application for loans*. Three different types of loan are available for financing the operation of the factory. Each type of loan has different conditions. Decisions must be made about the amount used of each loan type and about the repayment of earlier loans.

After the decisions are submitted, the simulation program generates the results of the current production period. The results are summarized in two reports:

– *Production report*. The production report summarizes the decisions made by the student teams for the current production period and the current state of the production system. The quantity of engines produced and sold, the quantity of parts used and the engine and part inventories at the end of the production period are given in details. The number of workers, machine capacities, number of production lines, and space, available for the next production period are also listed.

– *Financial report*. The financial report contains the balance sheet, the revenue report and the cash flow report valid at the end of the current production period.

When students evaluate the production and financial reports, and take decision on the next production period they need to apply their knowledge of several study areas taught on the Master's program. Awareness of marketing methods is required to estimate the behavior of customers when prices and payment conditions changes. A familiarity with forecasting models is needed to evaluate future demand possibilities. Inventory control and materials requirement planning techniques must be used to determine and control the inflow of raw materials and parts. Capacity planning techniques are needed to determine the workforce level, the number of assembly lines operating and the amount of space required. Cash flow analysis methods are required to evaluate the potential effects of efficiency improvement projects. Finally, managerial accounting and corporate finance knowledge is needed to properly understand balance sheets, cash flow reports and revenue reports.

At the end of the seventh production period the student teams are evaluated. This evaluation is very difficult even if only the financial situation of the plants is considered. Furthermore, purely financial analysis can be misleading. Some of the possible traps of narrow minded financial evaluation include:

– Short term success may not necessarily lead to long term success. The plant may make large profits in the first seven periods, but if production resources (production lines, production space, improvement projects) do not support production increases in the future financial performance may later decrease.

– A group may follow a cautious strategy. They may decide on a low production quantity, financed solely by their own financial sources. In these cases small profits and slow but steady growth can characterize the plant.

– Long term strategic thinking may provide unfavorable financial results in the short run. Heavy investments can be made at the beginning using loans in order to secure capacity for future growth. If all this is paired with a demand- stimulating marketing policy and with efficiency-improvement projects, profit will be low at the beginning, but steep growth can be expected in the future.

5 Comparison of Financial and DEA Results

Financial data (revenue and profit) are provided by the simulation game automatically in the output report at the end of each production period. The efficiency scores are determined with an SBM, input oriented, constant return to scale DEA model using an objective function (7) and a production possibility set (5). Cumulated output and input data are applied to evaluate the overall efficiency of operation over the course of seven production periods. Four inputs are used, which represent the four main resource groups used for production (workers, machines, material and money). The cumulated number of workers, the cumulated

number of machine hours, the cumulated sum of money spent on raw materials and the cumulated value of credits represent the resources used in the production process, and the corresponding values are taken from the production and financial reports. The cumulated revenue is used as a single output of the DEA model. Comparison of the financial results and the DEA results is based on a comparison of the cumulated profit and of the SBM efficiency scores of each team.

The results of the simulation game are summarized in Table 2. Column (2) of the table lists the revenue of the student teams while column (5) shows the ranking of the teams based on revenue. Column (3) shows the net profit value of the teams and column (6) shows the ranking of teams based on net profit. Finally, column (4) shows the SBM efficiency score of each team and column (7) shows the ranking of teams based on SBM efficiency scores.

Intuitively, it may be assumed that if high revenue is paired with good financial and production decisions then the profit and the efficiency score will also be high. Consequently, a team with a good profit rank will also have a good efficiency rank. A rank correlation analysis shows that the Spearman rho value is equal to 0.588 with a p-value equal to 0.008. This result indicates a strong correlation between efficiency scores and profit. Tied ranks of efficiency scores are substituted with rank averages in the calculation of the Spearman rho value (Iman and Conover, 1989).

Relatively high differences in ranks, however, not necessarily express very different results. Figures 1, 2 and 3 show the revenue, net profit and SBM efficiency scores of the teams in decreasing order and in the form of column diagrams. Figure 1 shows that there are very small differences among the revenues of the first 9 teams. Figure 2, however indicates, that teams with similar levels of revenue may have very different operations, as illustrated by the much wider spread of profit values. The spread of efficiency values is less marked, and several teams have identical or very similar efficiency scores. Similar efficiency scores, however, may be attained with very different operations, as indicated by the slack values in Table 3. The slack values of the optimal solution of the SBM DEA model can be used to explore operational shortcomings and possibilities for improvement.

Detailed analysis of the net profit ranks and efficiency ranks shows that high profit ranks do not always correlate with high efficiency ranks. Figure 4 shows the net profit ranks, revenue ranks and SMB efficiency ranks in a column diagram. Teams are ordered in increasing order of profit rank. It can be seen that in many cases that if revenue is high but it is not paired with efficient operation then profit is low, and the efficiency score is also low (see, for example Team 3, 5 and 16). Sometimes, however, low revenue and low profit co-occur with high SBM efficiency, showing a modest but efficient operation (see for example Team 14). Some typical cases are presented below to illustrate the additional insight provided by DEA results.

Table 2
Result of the simulation game

Team (1)	Revenue (WCU) (2)	Net profit (WCU) (3)	SBM Efficiency (4)	Rank Revenue (5)	Rank Net profit (6)	Rank Efficiency (7)
1	10,661,696	176,318	0.9907	13	18	8
2	10,577,446	1,148,395	0.9902	14	11	9
3	12,033,234	991,530	0.9784	8	12	13
4	12,061,476	1,538,008	0.9816	7	6	12
5	11,637,708	793,093	0.9653	10	14	17
6	12,101,500	1,800,132	0.9948	4	4	6
7	12,075,614	1,936,412	**1.0000**	6	3	1
8	10,232,003	1,370,187	0.9923	15	9	7
9	10,145,111	346,178	0.9771	16	17	15
10	12,213,420	2,048,193	**1.0000**	2	2	1
11	11,341,162	1,609,319	0.9875	12	5	10
12	10,102,317	120,912	0.9388	17	19	19
13	12,336,201	2,130,817	**1.0000**	1	1	1
14	9,633,311	726,334	0.9962	18	15	5
15	12,211,651	1,353,441	**1.0000**	3	10	1
16	12,099,778	1,511,871	0.9780	5	7	14
17	11,608,198	855,236	0.9823	11	13	11
18	12,012,255	1,479,598	0.9756	9	8	16
19	8,029,665	507,042	0.9649	19	16	18

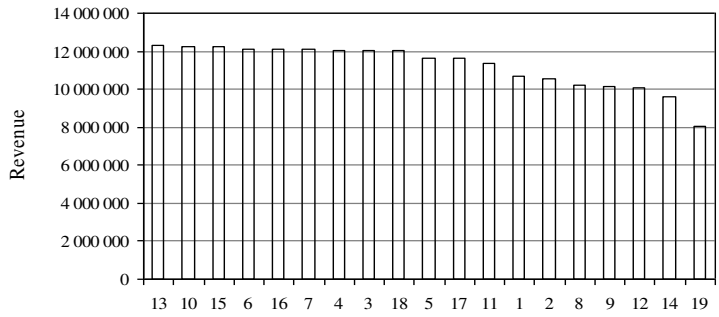


Figure 1
Revenue of team in decreasing order

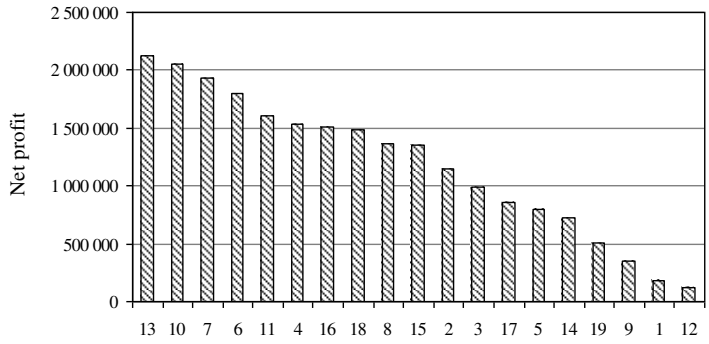


Figure 2
Net profit of team in decreasing order

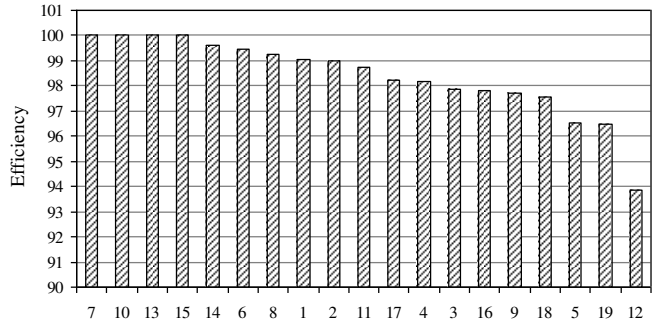


Figure 3
SBM score of teams in decreasing order

Team (1)	Revenue (2)	Workers (3)	Capacity (4)	Raw material (5)	Debt (6)
1	0	0	0	1.25	1.44
2	0	0	0.37	0.31	0.22
3	0	0	0	0.54	1.27
4	0	1.97	0.46	0	0
5	0	0.15	0	0.55	0
6	0	0	0	0.03	0
7	0	0	0	0	0
8	0	1.23	0.17	0	0
9	0	2.35	0	0.86	0.62
10	0	0	0	0	0

11	0	0.77	0.59	0	0
12	0	1.43	0	0.48	0.62
13	0	0	0	0	0
14	0	0.16	0	0.41	0
15	0	0	0	0	0
16	0	1.01	0.5	0	0
17	0	0	0	0.84	0.19
18	0	0.55	0.33	0	0
19	0	1.32	0.48	0	1.09

Table 3
Optimal slack values of the SBM DEA model (scaled data)

The best results were obtained by Team 13. This team is ranked first by all criteria. Three other teams were also efficient (Teams 7, 10, 15), but they have very different revenue and profit results. Consequently, a different operation policy may lead to different financial results, but still result in efficient operations. Based on net profit and on efficiency, Team 12 is ranked last.

Team 7 is ranked 6th in terms of revenue, but 3rd by profit. This shows that despite its relatively low revenue this team operated efficiently, which is reflected in the efficiency score. Team 14 had similar results, with low revenue and profit paired with relatively high efficiency. In this case, however, efficiency problems can be seen. Based on the slack values found in Table 3, the number of workers could be decreased and better inventory management would be required to improve the efficiency of this low revenue production.

Team 16 is ranked 5th by revenue, but the differences in revenue are insignificant for the first 9 teams, as seen in Figure 1. In terms of net profit, this team is ranked 7th and the differences in profit between the best teams is significant, as seen in Table 2. Consequently, operational shortcomings may be suspected, and indeed this is indicated by the relatively low efficiency score (rank 14). Similar results can also be observed for Team 3. Despite the similarity in results of Teams 3 and 16, they were quite different in operational terms, as indicated by the slack values in Table 3. Team 3 used excessive amounts of raw materials and ran up high levels of debt. Team 16 employed too many workers, and had excess machine capacity. Team 3's problem might be solved by improving inventory and financial management, while in the case of team 16, better capacity management may improve performance.

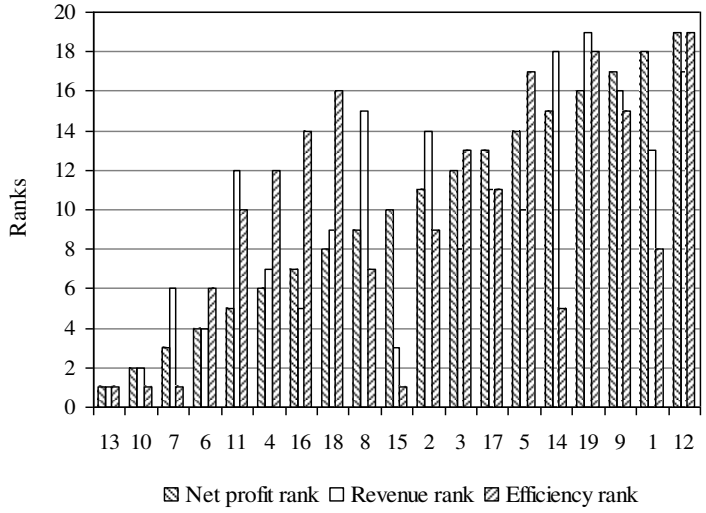


Figure 4
Comparison of the ranking of teams

Finally, an interesting case is Team 15, which is among the best with respect to revenue. This team is efficient according to the SBM score, even though its profit is relatively low (ranked 10th). In this case efficiency and profit does not correlate. This situation can be explained partly by the different marketing policy of Team 15. This team charged higher selling prices than the other teams and produced slightly smaller quantities. The joint effect of higher selling price and smaller quantity resulted in high total revenue (ranked 3rd). In this case the inefficient use of inputs had greater consequences for profit, than for SBM efficiency because the negative consequences of increasing selling prices are not considered by the DEA model. In the long run, selling price increases, may unfavorably affect market demand. This is not, however, reflected in the SBM efficiency score.

Conclusions

This paper compared different evaluation possibilities for a production simulation game. The first type of evaluation is based on financial data provided in form of traditional financial reports (balance sheet, revenue and cash-flow reports). The second type of the evaluation is made with the help of a slack based DEA model. Both results are based on the same basic principle, contrasting revenue with the resources used. Financial analysis takes profit as its major indicator, which is calculated as the difference between revenue and the cost of all resources used for operation. In DEA an efficiency score is calculated, which is the ratio of weighted outputs and weighted inputs. Since only one output is used in this paper, the

efficiency score expresses the ratio of revenue and the weighted sum of the major resources used for operation. Roughly speaking, first we take the difference between the revenue and the cost of resources and second, we take the ratio of the revenue and the resources used. Consequently, the major results of the two analyses should be similar. This is partly borne out by rank correlation analysis, which showed that profit ranks and efficiency ranks exhibited a strong correlation.

The major difference between performance evaluation based on financial data and DEA is that DEA considers only inputs which can be influenced by the decision maker, and intentionally incorporates them into the analysis. In contrast, profit related results include all the costs of operation. This difference has two major consequences:

− DEA based performance evaluation better expresses the shortcomings of operation caused by improper management decisions, since not all costs (only discretionary costs) are involved in the analysis. Ways to improve can thus easily be discovered.

− An assessment of profit includes all the costs of operation, while the inputs used in DEA are decided upon by the decision maker, which involves a subjective judgment in the calculation of the results. The advantage of this is that efficiency scores can better express the priorities of the decision maker. On the other hand, some important inputs are ignored, which may distort the result. Consequently, the selection of which inputs to employ in DEA must be very carefully considered.

Similar considerations can be made about the outputs. Financial based evaluation concentrates only on revenue, while DEA can incorporate several non-financial results of operation, such as quality indicators, customer satisfaction, speed of delivery etc. Consequently, DEA can provide a much more detailed picture of the results of operation and ways to improve.

Apart from the evaluation of results, one of the key objectives of performance evaluation is to explore ways to improve operation. Several techniques are used in financial analysis to explore avenues for improvement (see, for example, variance analysis in standard costing, or activity based costing). In DEA, however, the slack values directly show operational shortcomings and the areas of improvements.

This paper presented a special example of performance evaluation based on financial data and on the results of data envelopment analysis. Comparing the performance of student teams in a production simulation game provided information not only about the result of the game, but also about the learning process. A detailed analysis of the learning characteristics of student teams with DEA is presented by Koltai et al. (2013).

The main contribution of the results presented in this paper can be summarized as follows. The evaluation of the results of business simulation games with DEA is a new area of application. If the traditional financial information is the output of the

simulation game then an ideal environment is found for the comparison of financial results and DEA efficiency scores. First, all participating teams in the production simulation game produce the same products, use the same inputs, and apply the same production technology. Only marketing, financial and operation decisions are different. Consequently, *homogeneity assumption* is perfectly satisfied. Second, the results used for performance evaluation of the students, that is, the results and improvement possibilities are *directly traced back* to the decision makers.

In future work, the DEA-based performance evaluation presented in this paper can be extended to consider several other characteristics of operations. Non-financial outputs can easily be incorporated, the orientation of the analysis (input oriented, output oriented) can be changed, and the dynamic characteristics of the results can be analyzed with network DEA models. As a consequence of the development of DEA in the last decades, several important elements of real life operation can be easily considered. Consequently, applying DEA instead of, or in parallel with, financial analysis, is a challenging possibility for performance evaluation optimization.

References

[1] Ammar, S., Wright, R.: Experiential Learning Activities in operations Management, International Transactions in Operational Research, 1999, 6, 183-197

[2] Battini, B., Faccio, M., Persona, A., Sgarbossa, F.: Logistic GameTM: Learning by Doing and Knowledge-Sharing, Production Planning & Control, 2009, 20 (8), 724-736

[3] Bowlin, F. W.: An Analysis of the Financial Performance of Defense Business Segments using Data Envelopment Analysis, Journal of Accounting and Public Policy, 1999, 18 (4-5), 287-310

[4] Charnes, A., Cooper, W. W., Rhodes, A.: Measuring the Efficiency of Decision Making Units, European Journal of Operations Research, 1978, 2, 429-444

[5] Cooper, W. W., Seiford, L. M., Tone, K.: Data Envelopment Analysis, Springer, 2007

Doyle, J. R., Green, R. H.: Comparing Products using Data Envelopment Analysis, Omega, 1991, 19 (6), 631-638

[6] Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., Shale, E. A.: Pitfalls and Protocols in DEA, European Journal of Operational Research, 2001, 132, 245-259

[7] Ederer, N.: Evaluating Capital and Operating Cost Efficiency of Offshore Wind Farms: A DEA Approach, Renewable and Sustainable Energy Reviews, 2015, 42, 1034-1046

[8] Fenyves, V., Tarnóczi, T., Zsidó, K.: Financial Performance Evaluation of Agricultural Enterprises with DEA Method, Procedia Economics and Finance, 2015, 32, 423-431

[9] Feroz, E. H., Kim, S., Raab, R. L.: Financial Statement Analysis: A Data Envelopment Analysis Approach, Journal of the Operational Research Society, 2003, 54(1), 48-58

[10] Harrison, J., Rouse, P.: DEA and Accounting Performance Measurement, International Series in Operations Research and Management Science, 2016, 239, 385-412

[11] Holweg, M., Bicheno, J.: Supply Chain Simulation-a Tool for Education, Enhancement and Endeavor, International Journal of Production Economics, 2002, 78, 163-175

[12] Hosseinzadeh, A., Smyth, R., Valadkhani, A., Le, V.: Analyzing the Efficiency Performance of Major Australian mining Companies using Bootstrap Data Envelopment Analysis, Economic Modelling, 2016, 57, 26-35

[13] Iman, R. L., Conover, W. J.: Modern Business Statistics, John Wiley & Sons, 1989

[14] Johnes, J.: Data Envelopment Analysis and its Application to the Measurement of Efficiency in Higher Education, Economics of Education Review, 2006, 25 (3), 273-288

[15] Koltai, T., Lozano, S., Uzonyi-Kecskés, J., Moreno, P.: Evaluation of the Results of a Production Simulation Game using a Dynamic DEA Approach, Computers and Industrial Engineering, 2017, 105, 1-11

[16] Oral, M., Yolalan, R.: An Empirical Study on Measuring Operating Efficiency and Profitability of Bank Branches, European Journal of Operational Research, 1990, 46, 282-294

[17] Panayotis, A. M.: Data Envelopment Analysis Applied to Electricity Distribution Districts, Journal of the Operations Research Society, 1992, 43 (5), 549-555

[18] Sherman, H. D., Ladino G.: Managing Bank Productivity using Data Envelopment Analysis (DEA), Interfaces, 1995, 25 (2), 60-73

[19] Sinuany-Stern, Z., Mehrez, A., Barboy, A.: Academic Department's Efficiency via DEA, Computers & Operations Research, 1994, 21 (5), 543-556

[20] Smith, P.: Data Envelopment Analysis Applied to Financial Statements, Omega, 1990, 18 (2), 131-138

[21] Sterman, J.: Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment, Management Science, 1989, 35 (3), 321-339

[22] Thanassoulis, E., Boussofiane, A., Dyson, R. G.: A Comparison of Data Envelopment Analysis and Ratio Analysis as Tools for Performance Assessment, Omega, 1996, 24 (3), 229-244

[23] Tone, K.: A Slacks-based Measure of Efficiency in Data Envelopment Analysis, European Journal of Operational Research, 2001, 130, 498-509

[24] Vos, L.: Simulation Games in Business and Marketing Education: How Educators Assess Student Learning from Simulations, The International Journal of Management Education, 2015, 13 (1), 57-74

Modeling, Development and Control of Linear Twisted-String Actuator

Djordje Urukalo, Milos D Jovanovic, Aleksandar Rodic

Mihailo Pupin Institute, Volgina 15, 11000 Belgrade, Serbia
djordje.urukalo@pupin.rs, milos.jovanovic@pupin.rs, aleksandar.rodic@pupin.rs

Abstract: For the scientific community worldwide, developing a new actuator is a challenging task. New types of actuators are needed, especially in humanoid robotics in order to replace real human muscle. There are several approaches for how to obtain this goal. One approach is to realize real muscle using new synthetic materials such as piezoelectric components or pneumatic polymer materials. A second approach is to improve standard electromotor-gear actuators. Another unconventional approach is to use standard electromotor together with a tendon-based driving system. This paper presents a successful realization and control model for a proposed twisted-string actuator. Controller design is based on the National Instruments Single Board RIO driving a MAXON motor type tendon driven muscle. A Powerful Spartan FPGA is a key element for the presented hardware implementation. To program the whole system, LabVIEW software is used. Theoretically explained simulation results for adopted model design, as well as real measured experimental movement under the load force, are presented in the paper.

Keywords: twisted string; tendon; actuator; SB-Rio; LabVIEW

1 Introduction

Bio-inspired humanoid robotics and its realization is currently a promising area of research activity. One of the main goals today in the scientific community and technology is the realization of an efficient electrically-driven actuator that is comparable with real human muscle. The musculoskeletal system of the human body is one of the most well-known sophisticated actuation systems worldwide. Since nature took thousands of years to optimize each particular muscle of the human body, it is then obvious that biological knowledge of the human body should be taken into account in order to design a bio-inspired artificial muscle. Several different types of criteria must be satisfied: approx. same mass and dimensions of the artificial muscle and the human body muscle, maximum payload fraction, satisfactory speed and payload, precision and repeatability in a range of human skill, linear actuation in order to imitate biological muscle, compliancy actuation, etc. Recently, a lot of work has been carried out in order to

create artificial muscle (Figure 1) that is similar to the characteristics of human muscle, such as: pneumatic McKibben actuator [1], electroactive polymer actuators as artificial muscles [2], piezoelectric muscle-like actuator [3], shape memory alloys [4] and many other technical solutions.

Figure 1

The world realized artificial muscle: McKibben [1] pneumatic actuator (top-left), electroactive polymer actuators [2] (top-right), shape memory alloys [4] (down-left), piezoelectric muscle-like actuator (down-right) [3]

Numerous authors have already presented technical realizations of human-like actuation in robotics, such as: tendon driven antagonistic robotic actuator at the German Aerospace Center (DLR) [5], antagonistically coupled pneumatic actuator at Osaka University [6], and the Japanese robot Kenshiro from University of Tokyo [7].

Pneumatic actuators deal with high forces and displacement, however, distribution of control signals takes a huge amount of space. They are noisy and have a significant hysteresis work ratio. Electroactive polymer actuators sustain large forces for a small displacement. A large activation voltage is necessary for these types of actuators. Shape memory alloy actuators have high energy density, easy control, compact, and good mechanical properties, but they are rather expensive and they have a slow dynamic response and poor fatigue properties. Piezoelectric muscle-like actuators are suited only for very small forces and displacements. It is extremely difficult to control such kinds of devices because of problems with a very high hysteresis and memory effect.

Another proposed approach is from researchers at Duke University in Durham, USA; they revealed that they have grown the first ever human skeletal muscle that contracts in response to external stimuli, such as electrical impulses and pharmaceuticals [8]. This is a very promising area of research because it is obvious that natural muscle is the best possible actuator regarding the power efficiency versus realized force and torque, especially when compared to some other mechanical drive systems. If it will be possible to implement such kinds of

laboratory grown muscles to mechanical systems, extreme technical improvement needs to be done regarding the realized force and torque, high dynamic of the systems and even full system control. Of course, a lot of obstacles should be solved, such as a mechanical connection between human tissue and the external links and especially the problem of how to "power" such kinds of hybrid systems Some kind of bio power circulation for tissue should be realized to preserve the functionality and lifetime of the laboratory grown muscle.

In this paper, a twisted-string linear actuator is realized and presented in the paper as a light-weight, low-noise and compact linear design with high-speed actuation and satisfactory high payload.

2 Problem Statements and Task Description

A linear twisted-string actuator is designed to produce movement in a humanoid robot arm that is close to the movement of a human arm with similar requirements of speed and force. By taking into account the human arm dimension and natural human arm movement, artificial muscle requirements are calculated and simulated.

In order to estimate the required shoulder arm torque and composite speed at the end of the hand for natural human arm movement, a simulation of 7 degrees of freedom (d.o.f.) robotic arm in MATLAB is carried out [9-11]. Robotic arm parameters in standard D-H notation as well as masses of the segments used in the simulation are listed in the Table 1.

Table 1
Required robotic arm parameters and segment mass

Link i	a_{i-1}	α_{i-1}	d_i	Θ_i	m [kg]
1	0	$-\pi/2$	0	0	0
2	0	$-\pi/2$	0	0	0.223
3	-0.3	0	0	0	2.276
4	0	$\pi/2$	0	0	0.795
5	0	$-\pi/2$	0.3	0	0.586
6	0	$-\pi/2$	0	0	0.059
Tool	0.22	$\pi/2$	0	$-\pi$	0.213

The circular movement of the robotic arm by carrying the load of 1 kg for 6.5 seconds in the frontal plane is simulated. Mass hand center trajectory for this movement is shown on Figure 2:

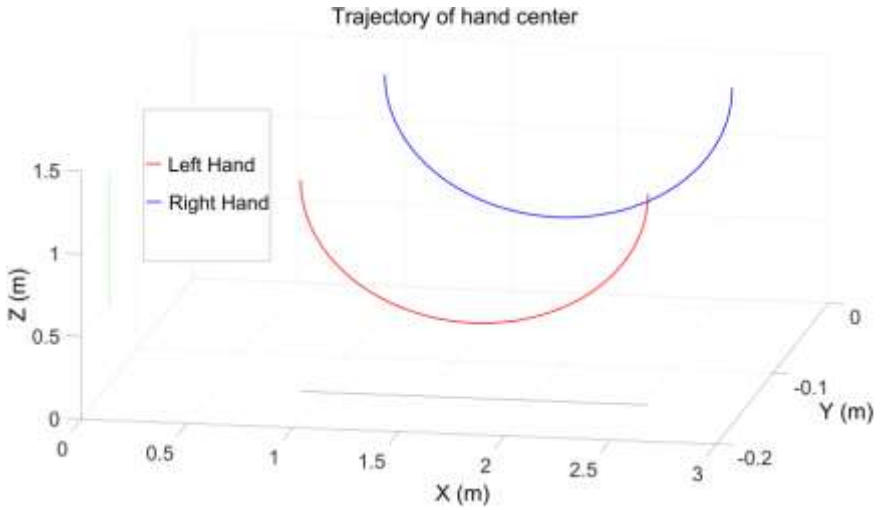


Figure 2

Mass hand center trajectory for circular movement of arms carrying 1 kg load. Duration of the movement is 6.5s.

Hand velocity v_{HAND} and acceleration a_{HAND} are calculated by using the following equations:

$$v_{HAND} = J_{HAND}\dot{Q} \tag{1}$$

$$a_{HAND} = J_{HAND}\ddot{Q} + \ddot{J}_{HAND}Q^2 \tag{2}$$

Where J_{HAND} is Jacobian matrix.

During the simulation, a maximum composite speed of 3 m/s at the center of the hand is found (Figure 3).

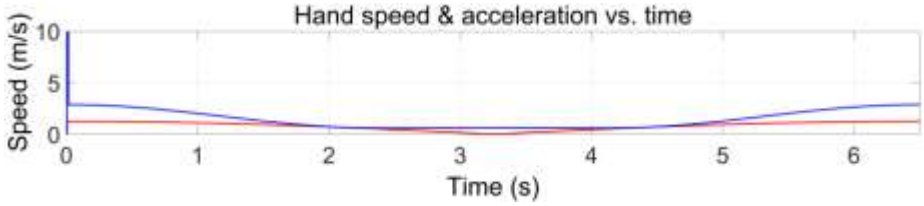


Figure 3

Calculated hand speed and acceleration in time

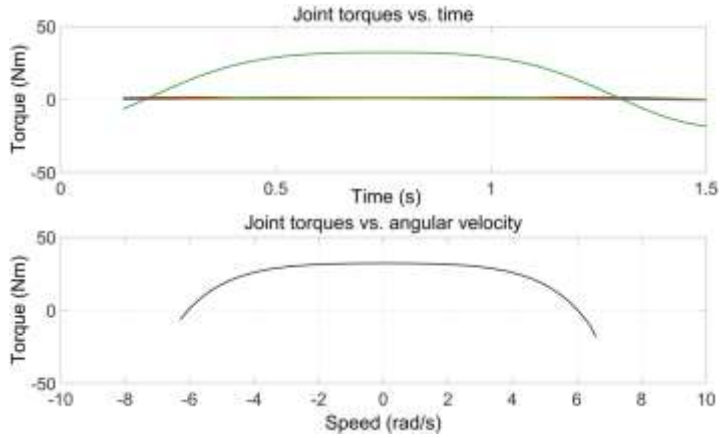


Figure 4
Joint torques versus time and angular velocity

Required shoulder arm torque (Figure 4) of maximum 20.89 Nm as well as other joint torques is determined using inverse dynamics from the simulation accordingly:

$$\tau = H(Q)\ddot{Q} + C(Q,\dot{Q})\dot{Q} + G(Q) - J_{HAND}{}^{T} F \tag{3}$$

Q, $\dot{Q}$, $\ddot{Q}$ are the vectors of generalized joint coordinates, velocities, and accelerations. H is the joint-space inertia matrix, C is the Coriolis and a centripetal coupling matrix, F is the friction force, and G is the gravity loading. The last term gives the joint forces due to a wrench F that is applied to the end effector.

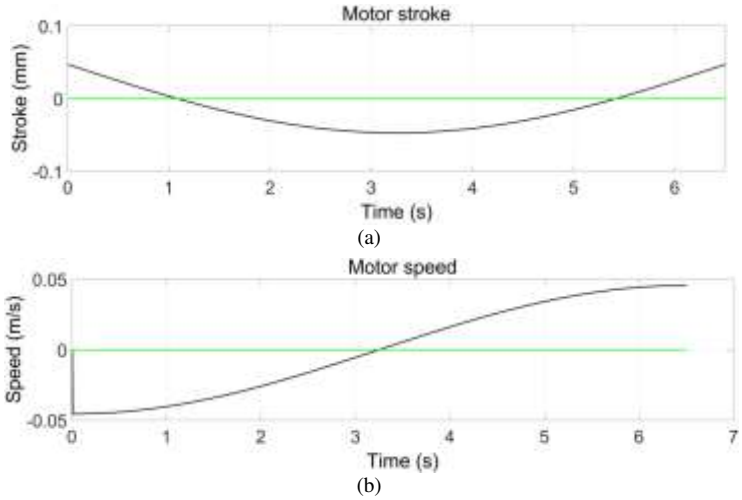


Figure 5
Calculated motor speed (a) and stroke (b) for linear actuator

Actuator linear motion in pulling direction of maximum 0.05 m is calculated as Motor stroke L_{MOTOR} accordingly (Figure 5a):

$$L_{MOTOR} = r_{MOTOR}(Q - Q_0) \tag{4}$$

Actuator pulling speed of maximum 0.05 m/s is determined as linear motor speed v_{MOTOR} (Figure 5b) accordingly:

$$v_{MOTOR} = r_{MOTOR}\dot{Q} \tag{5}$$

Actuator pulling force up to 696.2 N is found as the total tensile motor force for a certain degree of freedom, F_{MOTOR} (Figure 6), which is calculated by using the following equation:

$$F_{MOTOR} = \frac{\tau_{MOTOR}}{r_{MOTOR}} \tag{6}$$

Figure 6

Calculated motor payload for linear actuator

for which r_{MOTOR} represents the winch radius and Q_0 is zero position.

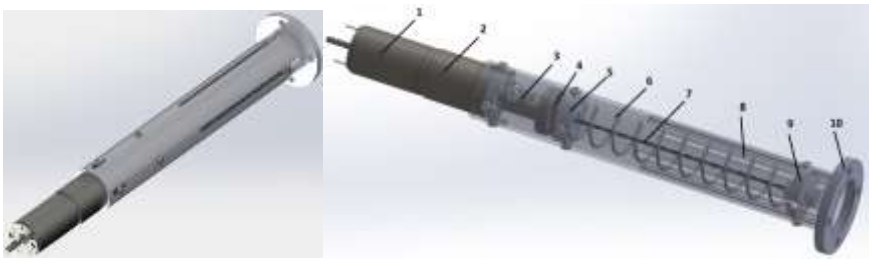


Figure 7

Design of proposed linear twisted-string acutator: 1. DC motor, 2. Gearhead, 3. Coupling, 4. Bearing, 5. Spring holder, 6. Spring, 7. Twisted-string, 8. Tube, 9. Driving part, 10. Bracket

To satisfy previously explained requirements, new approaches in mechanics should be introduced. For this purpose, the authors proposed a rather new design, twisted-string linear actuator (Figure 7), which has a tube structure with light-weight, low-noise, and compact linear design with high-speed actuation and satisfactory payload value.

This type of actuator should be used for driving humanoid robotic arms and hands. It is planned to use four twisted-string artificial muscles to actuate a humanoid arm without a hand in the following order: shoulder pitch and roll, elbow pitch, lower arm yaw (Figure 8). Actuators for shoulder pitch and roll joints will be placed inside the torso of a humanoid robot, and other two actuators will be placed in the upper arm link.

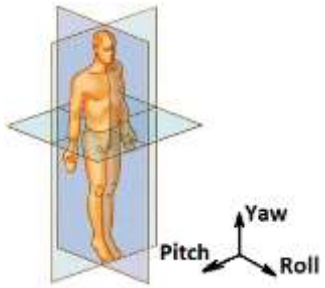


Figure 8
Possible angle of rotation of human's body [12]

Movement of a humanoid robotic hand will be realized with several twisted-string artificial muscles placed circularly in the lower-arm link. The total number for activating a robotic hand is still under investigation. Complex mechanical design as well as control of a light-weight humanoid robotic arm will be the next challengeable task in the future. This is still a promising new approach because of its complexity. There are some works concerning activation of only elbow joints with this type of actuator [13]. Here it is presented as a mechatronic design and experimental evaluation of synergy-based control for human-like grasping of robotic hand within the Dexmart Project [9, 12]. Several StMA-based hexapod walking robots are presented to the public [14].

3 Mechanical Design

Our proposed and realized twisted-string actuator is composed of one Maxon DC motor (20 W) (Figure 9, part 1) [15] that is equipped with an incremental optical encoder CPT1000 [17] and a planetary gearhead that has a ratio of 19.2:1 (Figure 9, part 2) [16]. Other mechanical parts of twisted-string actuators are: axial bearing FAG 51100 (Figure 9, part 3) [18], spring (Figure 9, part 4), coupling (Figure 9, part 5), 4 linear guides (Figure 9, part 6) with appropriate linear bearings (Figure 9, part 7), two brackets (Figure 9, part 8 and part 9), strings (Falcon Fishing Tackle, Catfish Leader, 146 kg) (Figure 9, part 10) and the stand (Figure 9, part 11). Other parts are not labeled. Complete mechanical design is presented in the Figure 9.

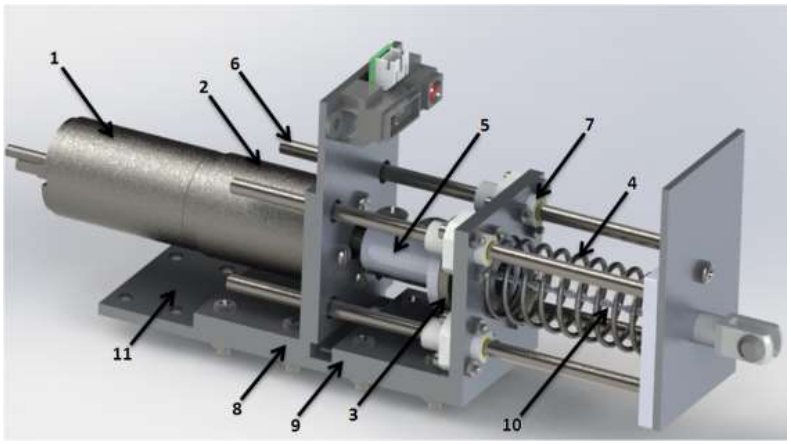


Figure 9

Mechanical design of twisted-string actuator with separate mechanical parts (1-DC motor with encoder; 2-planetary gearhead; 3-axial bearing; 4-spring; 5-coupling; 6-linear guides; 7-linear bearing; 8, 9-brackets; 10-string; 11-stand

DC motor with a gearhead is connected to the bracket, and the bracket is connected to the stand. An incremental encoder is used for actuator movement control. Another bracket is used for supporting 4 linear bearings that serve to lead 4 guides of 4mm diameter each in linear parallel motion. This bracket is also used to support the axial bearing that prevents gearhead from destruction of carrying a high axial load by twisting 4 non-tensile and high-flexible strings of 1mm diameter each. A spring is used for turning the actuator back to the initial position of maximum actuator displacement. The spring should be well-chosen; it should be powerful enough to overcome friction losses between linear bearings and guides, as well as friction losses inside twisted strings. If the actuator is mounted only in a vertical position, lifting and lowering the load, the spring is not necessary and could be removed from an actuator.

4 Controller Design

Global block diagram control scheme is shown on Figure 10. Standard PC computer is connected with Escon motor driver module [19] via USB in order to tune the current and the speed control loop; it is also used for monitoring controller's states and the DC motor's states with LabVIEW [20]. Escon module is in fact a smart DC motor driver. It consists one MOSFET H bridge together with an intelligent controller that is capable of realizing speed, current, and velocity control of the DC motor in the loop.

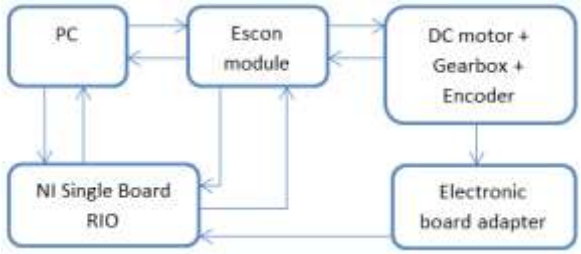


Figure 10
Global block diagram of controller design of twisted-string actuator

A PC computer is also connected to an NI Single-Board RIO 9636 [21] using ethernet cable in order to establish data acquisition of the measured signals in the system by LabWIEW. NI SbRIO 9636 is a powerful ARM based microcontroller board that operates under the real time NI OS. It consists of a 400 MHz ARM microcontroller and powerful SPARTAN FPGA running up to 40MHz. It has 48 programmable digital and analog IO pins which could be directly controlled by the FPGA or ARM microcontroller. Some IO pins could be both analog and digital according to the users.

A special electronic adapter board is realized to obtain 16-bit counter data which is received from the differential line encoder and filtered with high-speed logic circuits. NI Single-Board RIO is also connected to the ESCON module by some separate digital lines. Through this connection, the NI board can set motor current, speed, and direction of motor rotation acquiring the encoder's data in time during the movement. The Escon module also has the information about the encoder's data during the movement of the motor.

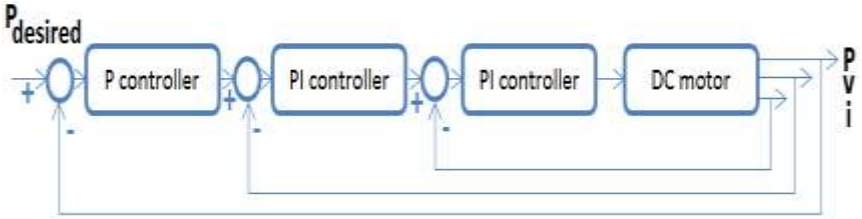


Figure 11
Cascade control structure for actuator control

A unique cascade control scheme (Figure 11) is realized to establish position and velocity control of a DC motor [22] and the whole actuator. Flexibility is a key feature of this type of controller. It consists of three distinct control loops: the innermost current loop is followed by the speed loop, and the speed loop is followed by an outermost position loop. This type of control requires increasing the response time of the controller towards the inner loop. In other words, the current loop is the fastest and the position loop is the slowest.

5 Model Analysis and Simulation Results

In order to control a twisted-string actuator, it is necessary to find a correct mathematical model that describes real physical systems. Until now, different analyses of twisted-string actuators have been carried out. Helix schematic representations of twisted-strings are used in modeling load position p (Figure 12).

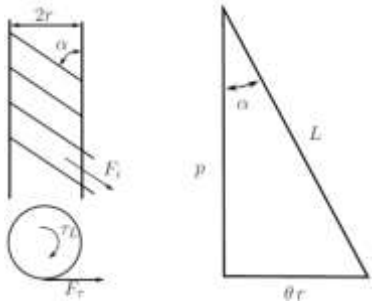


Figure 12
Helix schematic representation of twisted string

The parameters for modeling twisted-string actuators are: L-string length, θ-twisting angle, r-radius, α-helix slope, F_i -axial force of each string, τ_L-external torque, and F_τ -the tangential force. Position p is expressed by simple equations applying Pythagoras's theorem [23], where fiber tension is taken into account as well as stiffness, actuator rotation, and the radius of the string. Variations of the twisted-string radius with the twisting angle are included in [13]. The final model after consideration of the effective length of twisted string as a function of the number of turns is given in [24]. One simple model of load position is given in [9, 25-26]. Since proposed models of load position p mismatch measured load positions in real experiments, load position is expressed as a function of motor position (θ_M) and load (F_L) (7). Minimization processes between measured ($p_i^{measured}$) and obtained ($p_i^{obtained}$) load positions are taken into account by (8). It is done using the LM algorithm [27-28].

$$p_i^{obtained} = c_1 + c_2\theta_M + c_3\theta_M^2 + c_4 F_L \tag{7}$$

$$\min_{\theta_M, F_L} f(\theta_M, F_L) = \|F(\theta_M, F_L)\|_2^2 = \sum_i F_i^2(\theta_M, F_L) = \sum_i \left(p_i^{measured} - p_i^{obtained}\right)^2 \tag{8}$$

Minimization converged and residual is $1.0077 \cdot 10^{-5}$. The coefficients of the equation (7) are listed in the Table 2 and they are calculated using the simulation.

Table 2
Required robotic arm parameters and segment mass

c_1	0.0897
c_2	4.7922e-06
c_3	-9.5170e-09
c_4	3.2083e-05

Unloaded and untwisted string length L_0 is estimated (coefficient c_1 in equation (7)) and it will be used hereafter. Figure 13 represents lengths of twisted strings as a function of load and motor position. The points represent measured lengths of twisted-strings, and lines represent obtained lengths of twisted strings used (7).

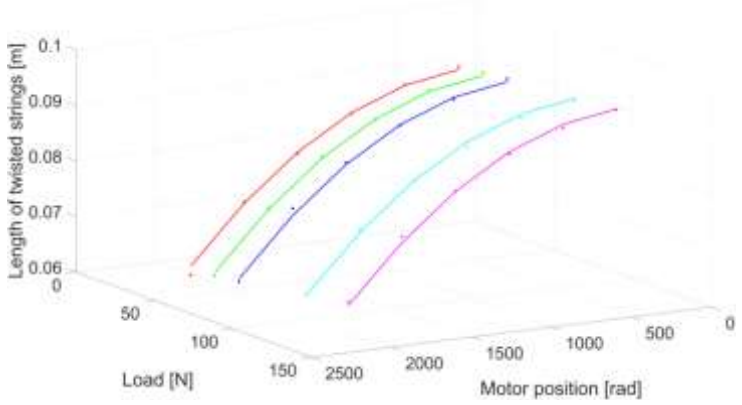


Figure 13

Lengths of twisted strings: measured points $\boldsymbol{p_i^{measured}}$ and obtained points $\boldsymbol{p_i^{obtained}}$

Hence, it is not possible to have a 100% malleable string, obtained load position is found applying Pythagoras's theorem (Figure 12), where ΔL is string elongation (9). String elongation is expressed as a function of motor position and load (10). Estimated unloaded and untwisted string's length $L_0 = 0.0897m$ is taken from Table 3. In order to find coefficients for equation (10) (listed in Table 3), a minimization process between measured $p_i^{measured}$ and obtained $p_i^{obtained}$ load positions are carried out (11).

$$p_i^{obtained} = \sqrt{(L_0 + \Delta L)^2 - (\theta_M r)^2} \tag{9}$$

$$\Delta L = f(\theta_M, F_L) = a_1\theta_M + a_2\theta_M^2 + a_3 F_L \tag{10}$$

$$\sum_i \left(p_i^{measured} - p_i^{obtained}\right)^2 \tag{11}$$

Table 3

Coefficients of equation (10)

a_1	8.1153e-06
a_2	8.4427e-09
a_3	1.6914e-05

Residual of minimization function (11) is $1.1403 \cdot 10^{-5}$.

Motor torque is estimated using the same principle as described earlier. The axial force of each string F_i can be found using Pythagoras's theorem (12):

$$F_i = \sqrt{F_L^2 + F_\tau^2} \tag{12}$$

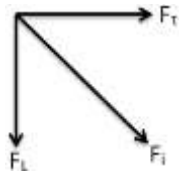


Figure 14
Decomposition of axial force of each string

Calculated axial force (12) in string F_I as a function of load and motor position is shown on Figure 14.

Transmission ratio can be found using:

$$\frac{\tau_L}{F_L} = \frac{\theta r^2}{p} \tag{13}$$

The tangential force F_τ (15) can be found transferring equation (13) and taking into account:

$$\tau_L = F_\tau r \tag{14}$$

$$F_\tau = \frac{F_L \theta_M r}{p} \tag{15}$$

Calculated motor torque (16) is ratio of gear torque τ_G and gear ratio i, where previous equations are used.

$$\tau_{M_i}^{calculated} = \frac{\tau_G}{i} = \frac{F_L \theta_G r^2}{i p_i^{measured}} = \frac{F_L r^2 \theta_M}{i^2 p_i^{measured}} \tag{16}$$

Obtained motor torque is expressed as the following:

$$\tau_{M_i}^{obtained} = b_1 + b_2\theta_M + b_3\theta_M^2 + b_4 F_L + b_5 F_L \theta_M \tag{17}$$

After the according minimization process:

$$\sum_i \left(\tau_{M_i}^{calculated} - \tau_{M_i}^{obtained}\right)^2 \tag{18}$$

Coefficients b_1 to b_5 are obtained and presented in Table 4, where the residual is $1.7743 \cdot 10^{-6}$.

Table 4
Obtained motor torque coefficients

b_1	0.0014
b_2	-3.1935e
b_3	1.4529e-09
b_4	-1.2142e-05
b_5	6.0940e-08

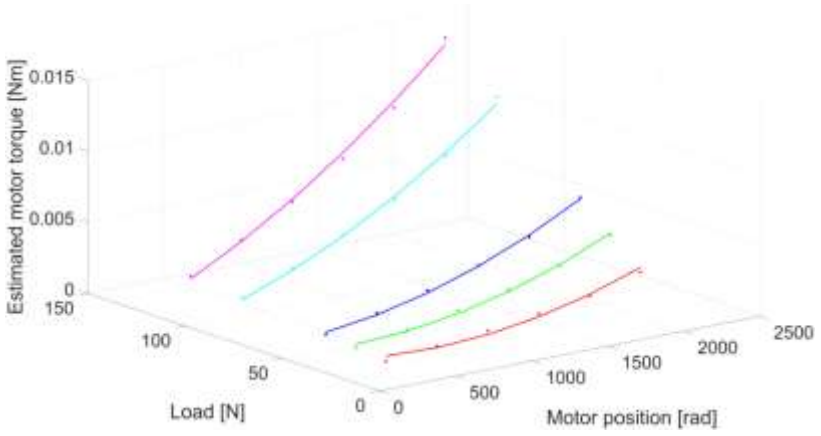


Figure 15

Estimated motor torque as a function of load and motor position

Motor position and motor load position dependency of calculated and estimated motor torque are shown on Figure 15.

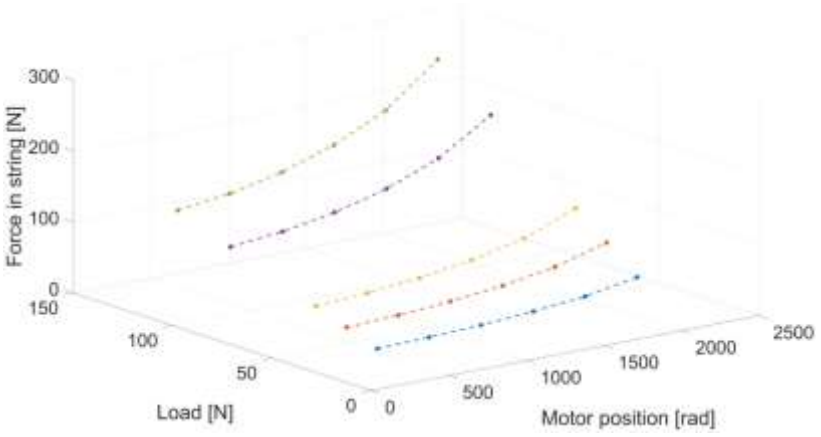


Figure 16

Estimated axial force in string as a function of load and motor position

Figure 16 presents dependency of calculated and estimated string force in relation to motor position and motor load.

Helix slope α is calculated using the following equation:

$$\alpha = \operatorname{acos}\left(\frac{F_L}{nF_i}\right) \tag{19}$$

Where n is the number of strings, which is 4 in the presented experiment. A relationship of Helix slope as a function of load and motor position is represented in Figure 17.

A small variation of angle α will produce a large load position p-variation.

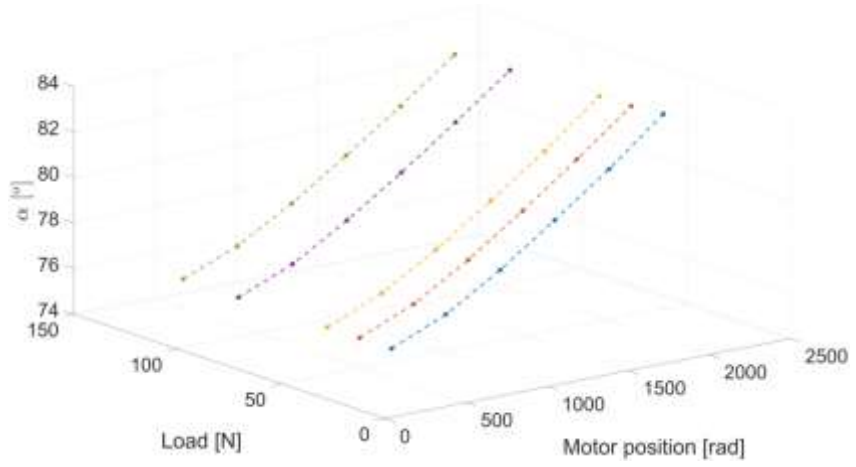


Figure 17
Helix slope as a function of load and motor position

6 Experimental Results

As mentioned before, a linear twisted-string actuator is placed in a vertical position. Different loads are applied onto the actuator in order to replicate weights of human muscle's load realistically. The loads of 23 N, 39 N, 54 N, 97 N and 124 N are applied. Strings are terminated on aluminum plates with 4 holes and twisted with the actuator from 0 to 7000 degrees with a step of 1000 degrees. A set of points from real experiment are obtained and presented onto the diagrams (Figure 13. Figure 15 and Figure 18). In Figure 18 a relationship between string length as a function of actuator rotation for each load is depicted.

There are quadratic regressions in a range from 0-6000 degrees of rotation angle where strings behave regularly, i.e. strings twist till maximum possible angle - strings pack properly. Above 6000 degrees of rotation angle, non-regular strings packaging appear. In such a way, several stress concentrations can cause strings to break.

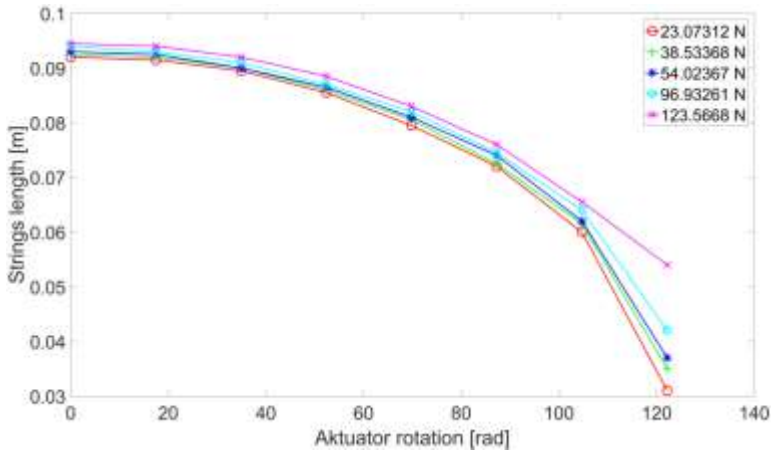


Figure 18
String's length as a function of actuator rotation

The speed and current control loops are tuned to the ESCON auto tuning process using required ESCON control software. The *P* gain of the position controller is found manually. Controller's parameters are listed in the Table 5.

Table 5
Applicated controller gains for a DC motor during the test

Controller	P gain	Integration time constant
Position	0.0035	-
Speed	963	42ms
Current	165	69µs

Maintaining stable control of a DC motor with maximum steady-state position error of 2 degrees is realized with hysteresis. An error of 2 degrees in position suits the presented application since the motor rotation is in thousands of degrees with four times multiplying encoding of optical differential encoder of 1000 CPT and the planetary gearbox with ratio 19.2:1. Stabilization of the motor position in hysteresis control is done by using a huge first order RC circuit with a time constant of 1^{10} seconds where states (20-21):

$$PV > SP - 2° \tag{20}$$

$$PV < SP + 2° \tag{21}$$

Where PV is Process Variable and SP is Set Point.

The main objective of this work is to make an actuator with similar characteristics of real human muscle. One of the characteristics of real human muscle is stiffness. Artificial twisted-string actuator has a property of stiffness that varies with motor position (Figure 19). That will give compliance in actuation of humanoid robotic arm.

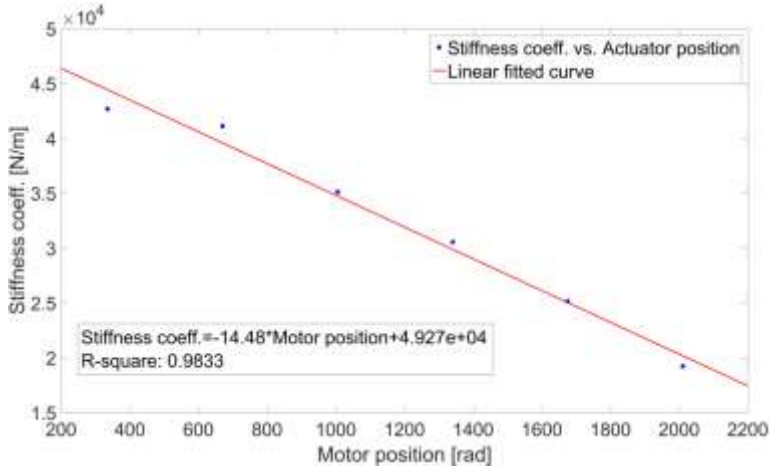


Figure 19

Stiffness coefficient as a function of motor position

Each point is obtained from a linear regression model of load versus string's length where correlation coefficients are: 0.9617, 0.9655, 0.9323, 0.9723, 0.9290, and 0.9855. It has to be noted that the first and last points are excluded to achieve a linear dependency role in decreasing manner of the twisted-strings stiffness coefficient versus actuator rotation with satisfied correlation coefficient of 0.9833.

For motor range between {335.1 – 2011} rad, string's length changes in range of {0.0315; 0.0305; 0.0305; 0.0290; 0.0285;} m, and considering a maximum rotation speed of 8000RPM is limited to a maximum permissible input speed in gearhead and maximum measured actuator speeds are successfully {1.58; 1.53; 1.53; 1.45; 1.42} cm/sec.

Conclusions

In the paper a new proposed twisted-string linear actuator is designed and realized (Figure 20). Characteristics of twisted-strings artificial muscle are found and explained. Hardware and software control design are done and described in this work. The actuator is placed in a vertical position for lifting and lowering different loads. According to presented tasks, a robust and reliable control design is accomplished.

The presented realized design and proposed control algorithm of twisted-string actuator supports current intentions for realizing artificial muscle that is very close to the characteristics of the human arm muscle. Non-linear dependency model of the actuator is observed and explained. Twisted-string actuator has the characteristic of compliance similar to that of true human muscle. Compliance is followed with a stiffness that varies depending on actuator length in a linear descending manner. There is a limitation of speed and velocity of presented twisted string actuator compared to that of human muscle.

Figure 20
Pictures of realized and tested twisted-string actuator

Certainly more attention should be paid to string choice and termination in order to have reliable and long-term activation of a humanoid robot arm. Further dynamic analysis and dynamic characteristics should be done, as well as real implementation in a robotic arm.

Acknowledgement

The research in the paper is funded by the Serbian Ministry of Education Science and technological development under the grants TR-35003, III-44008. The paper is partially supported by the project named by Research Group Linkage Program, Alexander von Humboldt Foundation, "Building attributes of artificial emotional intelligence aimed to make robots feel and sociable as humans (Emotionally Intelligent Robots - E I *robots*)", Contract no. 3.4-IP-DEU/112623, University of Kaiserslautern, Institute for informatics, Robotics department, Germany 2015-2017.

References

[1] F. Daerden, D. Lefeber, "Pneumatic Artificial Muscles: Actuators for Robotics and Automation", *European Journal of Mechanical and Environmental Engineering,* Vol. 47, pp. 10-21, 2000

[2] Y. Bar-Cohen, *Electroactive Polymer (EAP) Actuators as Artificial Muscles: Reality, Potential, and Challenges,* 2nd ed., SPIE Press, Vol. PM136, 2004

[3] T. Secord, "Design and Application of a Cellular, Piezoelectric, Artificial Muscle Actuator for Biorobotic Systems" Ph.D. dissertation, Dept. of Mechanical Engineering, Massachusetts Institute of Technology, Boston, Massachusetts, USA, 2010

[4] H. Taniguchi, "Flexible Artificial Muscle Actuator Using Coiled Shape Memory Alloy Wires", Proc. the 3rd International Conference on Biomedical Engineering and Technology - ICBET , Copenhagen, Denmark, Vol. 7, pp. 54-59, May, 2013

[5] M. Grebenstein, P. Van der Smagt, "Antagonism for a Highly Anthropomorphic Hand–Arm System", *Advanced Robotics*, Vol 22, No. 1, pp. 39-55, 2008

[6] Y. Ariga, H. Pham, M. Uemura, H. Hirai, F. Miyazaki, "Novel Equilibrium-Point Control of Agonist-Antagonist System with Pneumatic Artificial Muscles", Proc. of the IEEE International Conference on Robotics and Automation, Minnesota, USA, pp. 1470-1759, 2012

[7] Y. Nakanishi, S. Ohta, T. Shirai, Y. Asano, T. Kozuki, Y. Kakehashi, H. Mizoguchi, T. Kurotobi, Y. Motegi, K. Sasabuchi, J. Urata, K. Okada, I. Mizuuchi, M. Inaba, "Design Approach of Biologically-Inspired Musculoskeletal Humanoids", *International Journal of Advanced Robotics Systems*, Vol. 10, No. 216, pp 1-13, 2013

[8] http://www.medicalnewstoday.com/articles/288012.php

[9] A. Rodić, B. Miloradović, Đ. Urukalo, "Towards Building Of Lightweight Robot Arm Of Anthropomorphic Characteristics", International Conference On Electrical, Electronic And Computing Engineering-IcETRAN 2014, Vranjačka Banja, Srbija, 2-5.07.2014

[10] Karanovć V, Jocanović M, Jovanović V; "Review of Development Stages in the Conceptual Design of an Electro-Hydrostatic Actuator for Robotics", Acta Polytechnica Hungarica, Vol. 11, No. 5, pp. 59-79, 2014

[11] Torani F, Farahmandzad H, Aghamirsalim M; "Gray-Box Modeling of a Pneumatic Servo-Valve", Acta Polytechnica Hungarica, Vol. 7, No. 5, pp. 129-142, 2010

[12] https://en.wikipedia.org/wiki/Sagittal_plane#/media/File:BodyPlanes.jpg

[13] D. Popov, I. Gaponov, J.-H. Ryu, "Bidirectional Elbow Exoskeleton Based on Twisted-String Actuators", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5853,5858, 3-7 Nov. 2013

[14] M. Suzuki, T. Mayahara, A. Ishizaka, "Redundant Muscle Coordination of a Multi-DOF Robot Joint by Online Optimization," *IEEE/ASME international conference on Advanced intelligent mechatronics,* pp. 1,6, 4-7 Sept. 2007

[15] Maxon DC Motor, Re25, Graphite Brushes, 20W. Available data on the internet: http://www.maxonmotor.com/maxon/view/product/motor/dcmotor/re/re25/339150

[16] Maxon Planetary Gearhead GP 26A, 19.2:1, Available data on the internet: http://www.maxonmotor.com/maxon/view/product/gear/planetary/gp26/406762

[17] Maxon Encoder MR, CPT 1000. Available data on the internet: http://www.maxonmotor.com/maxon/view/product/sensor/encoder/Encoder-MR-TypML-128-1000imp-3Kanal/225780

[18] Axial bearing FAG 51100, Available data on the internet: http://www.fagbearing.com.cn/FAGbearinglist/FAG/FAG-51100.html

[19] ESCON Module, Servo controller. Available data on internet: http://www.maxonmotor.com/maxon/view/product/control/4-Q-Servokontroller/438725.

[20] LabView system design software. Available data on the internet: http://www.ni.com/labview/

[21] NI Single-Board RIO. Available data on the internet: http://www.ni.com/singleboard/

[22] N. Mohan, “Electrical Drives: An Integrative Approach”, MNPERE, USA, ISBN 0-9715292-1-3

[23] T. Würtz, C. May, B. Holz, C. Natale, G. Palli, C. Melchiorri, "The Twisted String Actuation System: Modeling and Control," *Advanced Intelligent Mechatronics (AIM), 2010 IEEE/ASME International Conference on* , vol., no., pp. 1215,1220, 6-9 July 2010

[24] J. J. Guzek, C. Petersen, S. Constantin and H. Lipson, "Mini Twist: A Study of Long-Range Linear Drive by String Twisting", ASME. *J. Mechanisms Robotics.* 2012; 4(1):014501-014501-7

[25] M. Suzuki, "Complex and Flexible Robot Motions by Strand-Muscle Actuators, Climbing and Walking Robots: towards New Applications", Houxiang Zhang (Ed.), ISBN: 978-3-902613-16-5, InTech, 2007

[26] I.-W. Park; V. SunSpiral, "Impedance Controlled Twisted String Actuators for Tensegrity Robots," *14th International Conference on Control, Automation and Systems (ICCAS),* pp. 1331-1338, 22-25 Oct. 2014

[27] K. Levenberg, "A Method for the Solution of Certain Problems in Least Squares," *Quart. Appl. Math*. Vol. 2, pp. 164-168, 1944

[28] D. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM J. Appl. Math*. Vol. 11, pp. 431-441, 1963

Application of Self-Organizing Maps for Technological Support of Droplet Epitaxy

Antal Ürmös[1], Zoltán Farkas[1], Márk Farkas[2], Tamás Sándor[3], László T. Kóczy[4], Ákos Nemcsics[1]

[1]Institute of Microelectronics and Technology, Óbuda University
Tavaszmező utca 17, H-1084 Budapest, Hungary
[2]Nexogen Ltd., Alkotás u. 53, H-1123 Budapest, Hungary
[3]Institute of Instrumentation and Automation, Óbuda University
Tavaszmező utca 17, H-1084 Budapest, Hungary
[4]Department of Automation, Széchenyi István University
Egyetem tér 1, H-9026 Győr, Hungary
E-mails: urmos.antal@kvk.uni-obuda.hu, farkas.zoltan@kvk.uni-obuda.hu, farkas.mark@nexogen.hu, sandor.tamas@kvk.uni-obuda.hu, koczy@sze.hu, nemcsics.akos@kvk.uni-obuda.hu

Abstract: The subject of this paper is the self-organized grouping of droplet epitaxial III-V-based nano-structures. For the nano-structure grouping, our developed algorithm - called Quantum Structure Analyzer 1.0 - is used. The operation of this software is based on the principles of the Kohonen Self-Organizing Network. Here, three possibilities for nano-structured groupings are shown. On one hand, we examine the classification of nano-structures with Kohonen Self-Organizing Maps, on the other hand, fuzzy inference systems are applied for the same goal. In the case of the fuzzy methods two approaches are examined in detail. According to the first fuzzy inference approach, the shape factor is calculated from the size of nanostructures. According to the second fuzzy inference approach, the shape factor calculation is based on the controllable parameters of the growth process (eg. pressure and the temperature of the substrate).

Keywords: nanostructure; classification; self-assembling; Kohonen SOM; fuzzy inference system; shape factor

1 Introduction

Recently, semiconductor nanostructures are being intensely examined in basic research as well as in applied science. The importance of epitaxially grown, low-dimensional nano-structures, is well recognizable on the yield improvement of electronic devices (eg. LEDs, lasers, solar cells). These technologies may also have impact on the development of radically new computers, such as, quantum

computers. These nano-structures are manufactured mainly with molecular beam epitaxy (MBE) technology. The features and functioning of these devices depend on the type, shape, size and spatial distribution of the contained nano-structures. For this reason, it is essential to know the impact the technological parameters have on the qualities of the above mentioned nano-structures.

A good example for the application of these nanostructures is in the manufacturing of high efficiency solar cells. There are two developing applications for solar cells. One of them is the search of the highest efficiency for solar cells (e.g. for space exploration applications). The GaAs-based solar cells containing quantum wells ($\eta > 40\%$) or quantum dots ($\eta > 60\%$) belong to this group [1] [2].

There are papers from several authors on solar cells created by using intermediate-band quantum dot (QD) structures. These QD layers are between the two usual p and n layers. The band structure is shown in (Figure 1A). The photon current on these junctions is added to the usual current between the valence band and conductance band. In this way, a very high efficiency can be achieved [3] [4] [5].

The total band gap of an optimal intermediate-band QD solar cell is 1.95 eV, which is divided into two sub band gaps. The width of the higher band gap is E_L (Figure 1B) and the width of the lower band gap E_H is 1.24 eV. As an intermediate band, the allowed energy levels of the QDs can be used. The solar cell operating this way was described in year of 2004. On Figure 1C, the layers of this cell can be seen. In that solar cell InAs QDs were grown in GaAs matrix by MBE with the Stransky-Krastanov mode.

The nanostructures mentioned in this section are prepared by droplet epitaxy (DE), which is described later. Although the DE process was developed in detail by a team led by Koguchi at the beginning of the 90s [6] [7], publications already covered the subject between 1985 and 1991. Figure 2. indicates the number of publications on this DE method. As it can be seen, the number of relevant publications exponentially grows between 1985 and 2016.

These nanostructures can be grown using MBE equipment and a DE process [6]. Several shaped nano-structures can be formed by DE such as QDs, quantum rings (QRs), double quantum rings (DQRs), and nano-holes (NHs).

The process starts with depositing a metallic component from column of III onto a substrate (eg. on GaAs) (Figure 3). The deposited material forms droplets on the substrate surface. The next stage is the crystallization of the material from the column of V. In this second step, different types of nano-structures are formed depending on the physical parameters of the process (eg. temperature of the sample and arsenic pressure) [8].

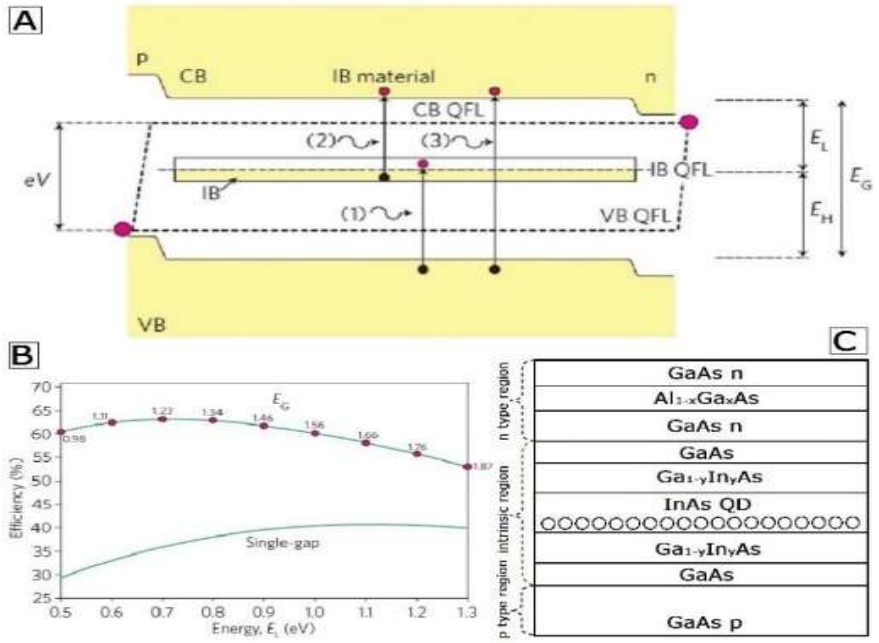


Figure 1

(A.)The band diagram of intermediate-band QD solar cell. E_G is the band gap, E_L and E_H are sub-bandgaps. The CB QFL is the quasi Fermi niveau of conductance band and VB QFL is the quasi Fermi niveau of valence band. (1) and (2) stand for photon absorption under the band gap, (3) stands for photon absorption above the band gap. Source [4]. (B) IB and monogap efficiency diagram as a function of E_L sub-bandgap. (C) The layer diagram of IBQD solar cell.

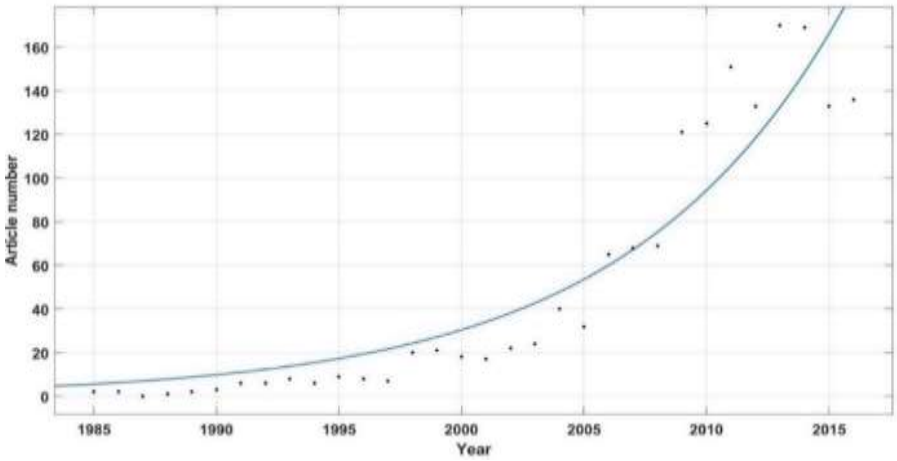


Figure 2

Number of publications on droplet epitaxy between 1985 and 2016

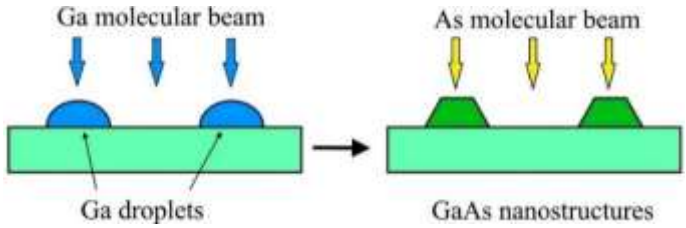


Figure 3
Formation of nanostructures during droplet epitaxy process (source: [8])

The characteristics of nano-structures during the formation process depend on several variables. In the case of low-temperature and high background arsenic pressure, QDs are formed. The size, the spatial density and the distribution of Ga droplets can be controlled by Ga flux, Ga coverage and the temperature of the substrate. The morphology of the nano-structure can also be controlled by substrate temperature and background pressure of arsenic as well [7].

During the growth process, it is essential to know the technological parameters (eg. arsenic pressure, substrate temperature and Ga flux), for the types of nano-structures formed and the details of this formation process. The clarification of these issues is facilitated by our developed software called "Quantum Structure Analyzer 1.0". The operation of this software is based on Kohonen's Self-Organizing Network [9].

2 Clustering Quantum Structures

The goal of this paper is to establish a grouping or clustering of quantum structures with regard of the above mentioned technological parameters. Several technological factors exert an influence on the formation of a single nano-structure. Moreover, the type determination of a nano-structure is not a simple task. The quick recognition, that is, the transition based on the physical appearance of the structure from one type to the other is continuous. This makes ordering a given nano-structure to one type or to another, on the basis of a simple rule is impossible. For that reason, different types are identified by neural networks. These networks are not programmed in the usual rule based system applications, but are trained to recognize clusters. The training can be supervised or unsupervised. A good example for unsupervised learning techniques is the self-organizing map (SOM). It is an artificial topographic mapping inspired by neuro-biological research [10].

The topological arrangement is created by multiple repetitions of the following process (Figure 4).

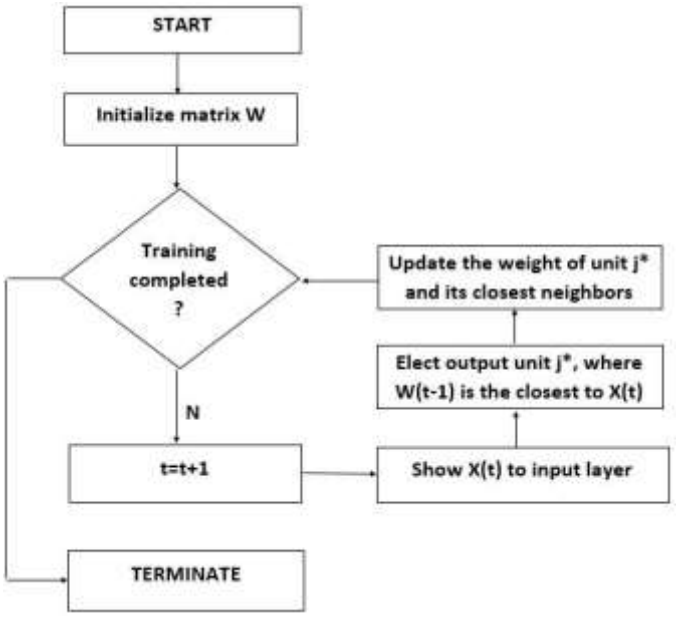


Figure 4
Flowchart of Kohonen SOM

3 Classifying Quantum Structures with Fuzzy Inference System

A possible way to classify quantum structures is to define a shape factor. Because of the continuous transition between different types of nano-structures, the shape factor can be calculated with the help of fuzzy logic.

Fuzzy logic is based on the fact, that a two valued (true/false) logic, is inappropriate to emulate human thinking and it is also inappropriate to describe certain phenomena. Thus, fuzzy logic uses intermediate values, between true and false. There are statements that where it is impossible to decide if they were true or false, but intermediate values indicate the 'degree of truth'. This was the reasoning behind the invention the fuzzy logic by L. Zadeh in the 1960s [12] [13]. A X fuzzy set can be defined by a so-called membership function. This Membership functions relate a value from the interval [0,1] to every value of the x base set. The value between 0 and 1 reflects the "extent" to which the given x value is a member of the X fuzzy set:

$$\gamma_x \in X \rightarrow [0,1] \tag{1}$$

where γ_x is the membership function of X fuzzy set, that unequivocally defines the set. There are different types of membership functions, in this paper we use the most frequent triangular and trapezoid shapes [12] [13].

One application of the theory described above is the fuzzy inference system. These systems are based on a rule base model. This model consists of fuzzy sets and "if-then" constructions. These systems are frequently used and they have various applications, for example [14] [15] [16] [17]. There are many types of fuzzy inference system, for example Mamdani, Sugeno, Tsukamoto etc. [18]. The inference steps of the algorithm are the following:

1. Fuzzification
2. Aggregation
3. Defuzzification

In the fuzzification, the crisp data is converted into fuzzy data. During the aggregation the fuzzy sets, that represent the outputs of each rule are combined into a single fuzzy set. In the defuzzification the fuzzified data is converted back into the crisp number. There are many ways for the defuzzification, for example in case of the centroid method:

$$x^* = \frac{\sum_{i=1}^{m} \mu_c(x_i) * x_i}{\sum_{i=1}^{m} \mu_c(x_i)}, \tag{2}$$

m is the quantization levels in the output, x_i is the ith data, $\mu_c(x_i)$ is the ith membership function.

In this model, Mamdani-type fuzzy inference system is applied. In this case, the general form of the rules are the following:

$$k: IF\ x\ is\ A_i^k\ AND\ y\ is\ B_j^k\ THEN\ z\ is\ C_l^k, \tag{3}$$

where k = 1, 2, . . . , R, i = 1, 2, . . . , N, j = 1, 2, . . . , M and l = 1, 2, . . . , L. N, M and L are the numbers of membership functions for input and output variables, R is the number of the rules [18]. In the model, the default settings (AND operation: MIN operator, OR operation: MAX operator, implication method: MIN operator, the aggregation: MAX operator, centroid defuzzification algorithm) are used. The details of the fuzzy model will be discussed in the next chapter.

4 Results and Discussion

The input of SOM (also referred to as „teaching data") can be seen in Table 1. A data vector is ordered to each sample. This vector contains the parameters of the nanostructure: the column I. is the number of the data vector; the column II. is the temperature of the substrate; the column III. is the flux of the component (Ga, In, Al, etc.); the column IV. is the surface coverage; the column V. is the arsenic pressure; the column VI. is the annealing time; the column VII. is the annealing temperature; the column VIII. is the base diameter of nano-structure; the column IX. is diameter of the ridge circle; the column X. is the distance of the substrate surface and the highest point of the nanostructure; the column XI. is the distance between the highest point of the nanostructure and the bottom point of the nanostructure; the column XII. is the spatial distribution of the nano-structures on the surface. The columns VIII. IX. X. XI. are the geometrical parameters of the nanostructure, and their interpretation will be introduced in the Figure 10.

I.	*II.*	*III.*	*IV.*	*V.*	*VI.*	*VII.*	*VIII.*	*IX.*	*X.*	*XI.*	*XII.*
1	200	0.19	2.75	1.00E-04	10	350	60	0	7	0	1.20E+10
2	250	0.025	2.75	1.00E-04	10	350	107.5	0	37	0	4.40E+08
3	200	0.75	3.75	5.00E-05	1	350	50	0	5	0	3.60E+10
4	300	0.75	3.75	4.00E-06	5	300	60	40	2	2	1.50E+09
5	300	0.05	1.75	1.00E-05	0.33	300	100	40	20	15	1.30E+08
6	260	0.025	3.75	2.00E-04	10	350	167	0	50	0	1.60E+08
7	260	0.025	3.75	2.00E-04	10	350	250	0	35	0	1.60E+08
8	620	0.4	3.2	7.00E-07	5	620	350	150	25	55	8.00E+06
9	620	0.4	3.2	9.00E-07	5	620	350	150	25	45	9.00E+06
10	640	0.8	2	1.00E-07	3	640	200	200	15	82.4	9.00E+06
11	640	0.8	2.4	1.00E-07	3	640	200	200	15	91.12	9.00E+06
12	650	0.8	3.2	1.00E-07	3	600	15.8	15.8	11	22	8.00E+06
13	650	0.8	2	1.00E-07	2	650	300	200	2	62	8.00E+06
14	200	0.19	3.75	6.40E-05	10	350	60	0	7.5	0	1.50E+10
15	200	0.19	3.75	5.00E-05	1	350	60	0	7	0	1.20E+10
16	250	0.025	3.75	5.00E-05	10	350	110	0	32	0	4.40E+08
17	300	0.75	3.75	4.00E-06	5	300	80	35	3.6	3.6	1.50E+09
18	200	0.19	3.75	6.40E-05	10	350	40	0	7	0	3.60E+10
19	200	0.19	3.75	4.00E-06	10	300	60	60	2	2	1.50E+09
20	507	0.08	3.2	1.00E-07	2	620	210	200	4	16.5	7.50E+07
21	507	0.08	3.2	1.00E-07	2	620	200	200	0.5	3	1.60E+08
22	300	0.75	10.5	1.00E-06	0.33	300	40	10	3.3	3.3	1.50E+09
23	300	0.75	10.5	1.00E-06	1	300	70	30	3.3	7.3	4.50E-07
24	200	0.19	6	1.00E-06	1	350	40	40	2.5	2.5	8.00E+09
25	500	0.04	4	3.00E-09	30	500	290	150	15	25	4.50E+07
26	520	0.8	2.4	1.00E-07	0.05	520	200	100	3	16	5.00E+06
27	520	0.8	2.4	1.00E-07	0.05	520	200	100	4	24	1.25E+07

28	620	0.47	2.82	9.00E-07	3	620	160	10	0	5	4.00E+08
29	160	0.79	3.75	1.00E-04	10	350	60	0	7.5	0	2.00E+11
30	200	0.79	3.75	1.00E-04	10	350	60	0	7.5	0	9.00E+10
31	250	0.79	3.75	1.00E-04	10	350	250	0	35	0	1.00E+10
32	260	0.79	3.75	1.00E-04	10	350	250	0	35	0	8.00E+10
33	500	1	3	5.00E-09	0	600	185	54	4	21.5	4.50E+07
34	500	1	3	5.00E-09	0	620	185	54	3	20.5	4.50E+07
35	500	1	3	5.00E-09	1	620	185	64	2	19.5	4.50E+07
36	540	0.8	3.2	1.00E-06	2	620	200	100	2.5	9.5	4.00E+07
37	540	0.8	3.2	1.00E-07	2	620	200	100	2.5	16.5	9.00E+07
38	520	0.8	8	3.00E-06	2	620	200	100	2	5	5.00E+06
39	520	0.8	8	1.00E-07	2	620	200	100	2	20	1.25E+07

Table 1

Input data of self organizing mapping. The (I.) is the serial number, the (II.) is the temperature of the substrate. The (III.) is the flux of the component (Ga, In, Al, etc.), the (IV.) is the surface coverage. The (V.) is the background pressure of arsenic, and (VI and VII respectively) the time period and temperature of annealing. The characteristic geometrial sizes of the nanostructure (diameter of base circle (VIII.), diameter of the ridge circle (IX.), the distance of the substrate and the highest point of the nanostructure (X.), the distance of the highest and lowest points of the nanostructure (XI.)) and the spatial distribution of the nanostructures on the surface (XII.).

The Kohonen SOM is an iterative algorithm. The theory of the Kohonen SOM algorithm is applied by our developed software code named Quantum Structure Analyzer 1.0. The results of this algorithm are shown as the function of iterative steps on Figure 5. The forming of Kohonen graph can be seen after step 5 (A), after step 10 (B), after step 100 (C), after step 500 (D), after step 1000 (E) and after step 2000 (F).

On Figure 6, It can be seen that the temperature of the substrate grows vertically upward and the background pressure of the arsenic decreases statistically horizontally to the right. On Figure 7, it can be seen that component flux grows vertically upward and the surface coverage grows statistically horizontally to the right.

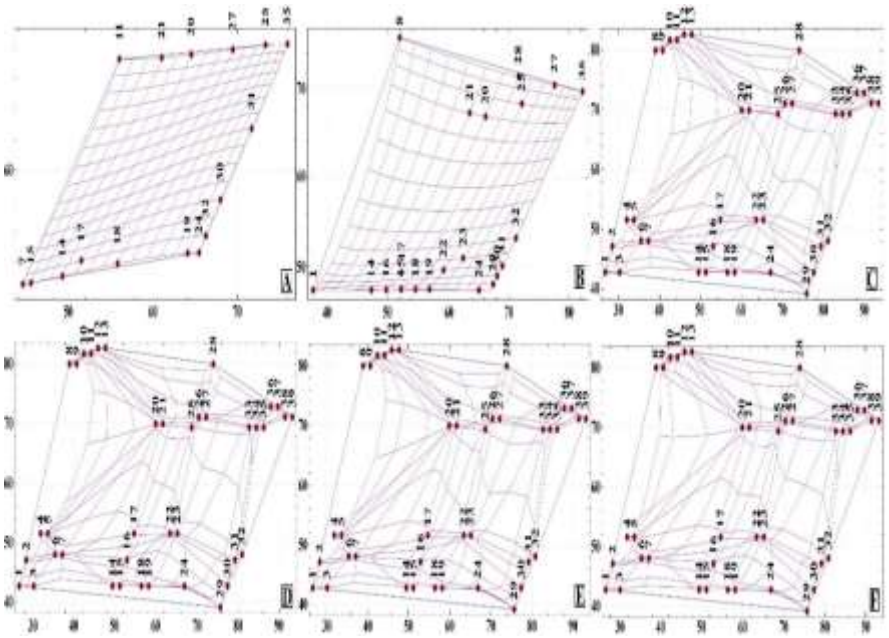


Figure 5

Snapshots of Kohonen Self-Organizing Maps after 5 steps (A), after 10 steps (B), after 100 steps (C), after 500 steps (D), after 1000 steps (E) and after 2000 steps (F)

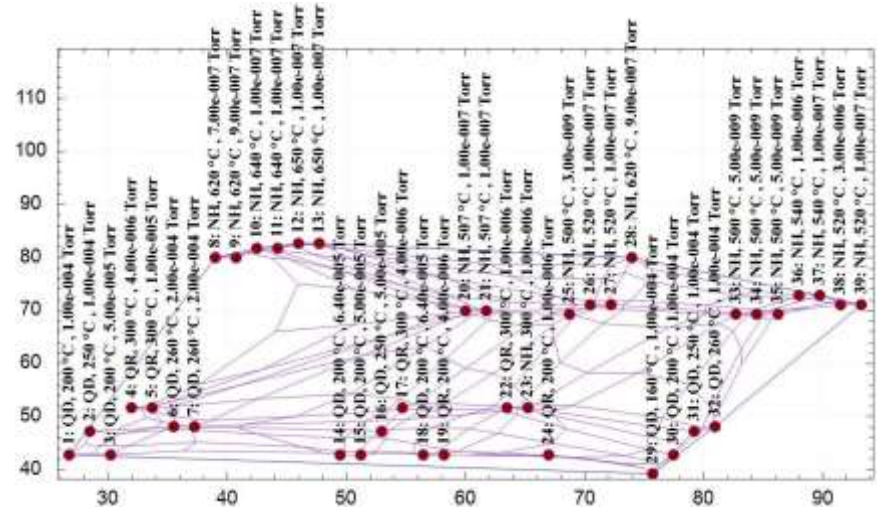


Figure 6

Substrate temperature – Arsenic background pressure diagram. The temperature of the substrate grows vertically upward and the background pressure of the arsenic decreases statistically horizontally to the right.

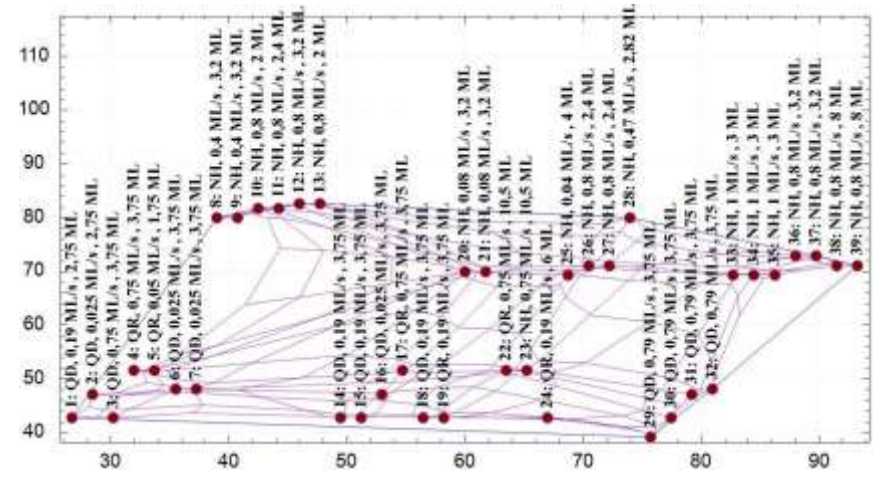


Figure 7

Component flux – surface coverage diagram. Component flux grows vertically upward and the surface coverage grows statistically horizontally to the right.

On Figure 8, it can be seen that temperature of annealing grows upward and the time of annealing statistically decreases to the right.

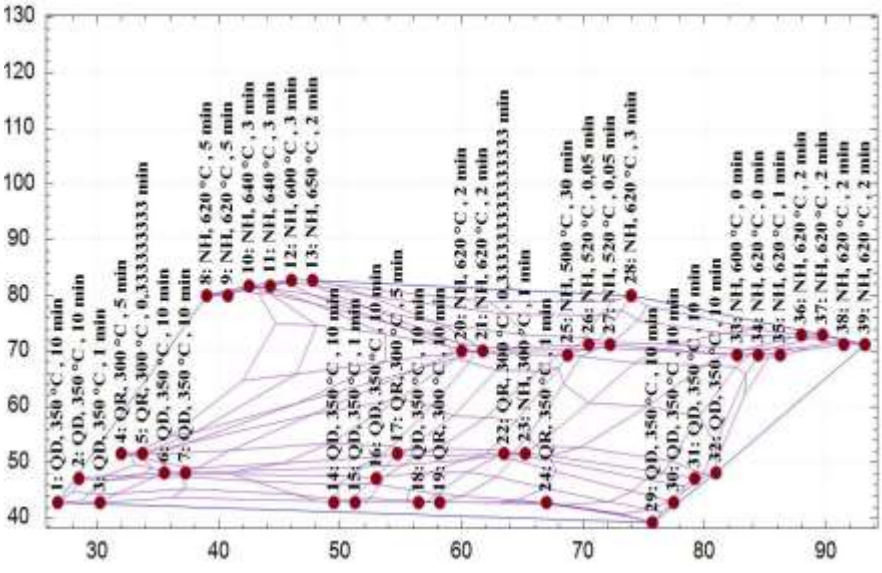


Figure 8

Temperature of annealing – Time of annealing diagram. The temperature of annealing grows upward and the time of annealing statistically decreases to the right.

The classification of nano-structures can be carried out several ways. Figure 9 displays the original classification based on the literature. The original data was obtained from many articles, for example Nemcsics et al. [19] [20] [21], Heyn et al. [22] [23], Kuroda et al. [24].

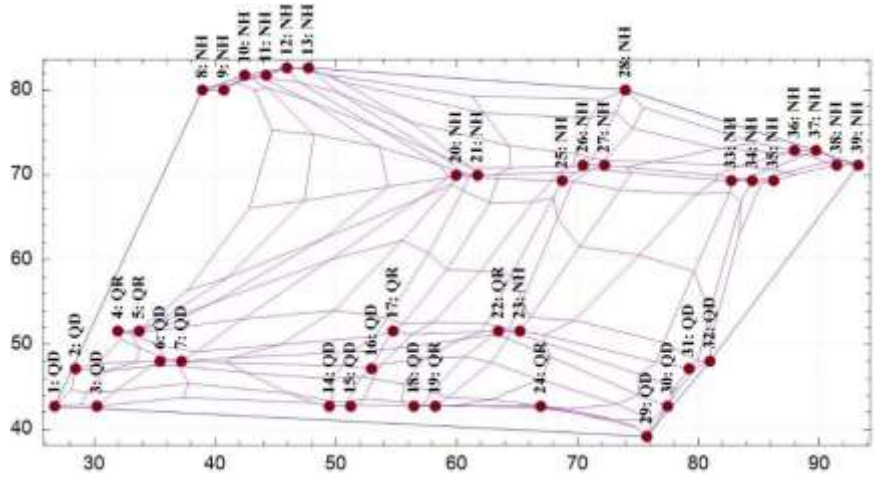


Figure 9

Original clustering of nanostructures (based on the literature)

The nano-structures were also classified according to geometrical attributes. The theory of fuzzy inference systems is applied by fuzzy inference system based geometrical quantum structure classification, which is developed in the Matlab environment, with the using of the Matlab Fuzzy Toolbox. Refer to Figure 10 for interpretation of the dimensions. An important thing is, that the interpretation of the geometrical parameters are the same in the Figure 10 part a, Figure 10 part b and Figure 10 part c as well. A is the diameter of the base circle of the nanostructure:

$$A = A(d) = \begin{cases} d, if\ C \geq 0{,}1\ nm \\ 0, otherwise \end{cases} \tag{4}$$

The value B is the diameter of the ridge circle of the nanostructure. C is the distance of the top of the nanostructure from the substrate. D is the distance of the top of the nanostructure from the global or local minimum of the nanostructure. In Table 1, the size A corresponds to the column VIII., the size B corresponds to the column IX., the size C corresponds to the column X. and the size D corresponds to the column XI.

The shape factor can be calculated according to B, C or D. If the shape factor is defined according parameter B then QDs can be separated from QRs and from

NHs. But in this case neither the latter two cannot be separated nor the type of "hybrid" (transitionary) nano-structures cannot be determined.

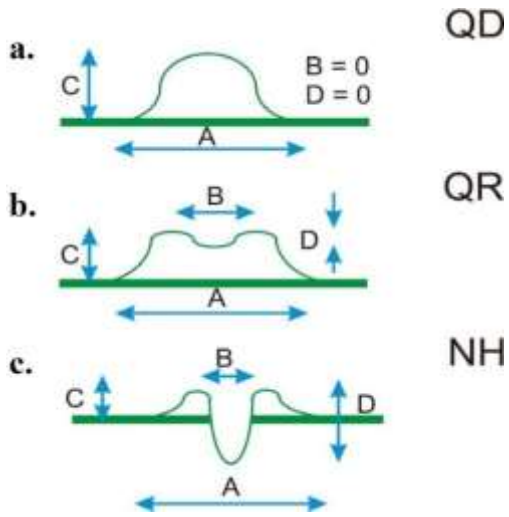


Figure 10
Geometrical dimensions of different nanostructures

The operation of the fuzzy model is based on Figure 10. In this case, the size C and the size D geometrical parameters are considered. The size C is the distance of the top of the nanostructure from the substrate. The size D is the distance of the top of the nanostructure from the global or local minimum of the nanostructure. If size D is smaller than or equal to 2 nanometer, then the structure is considered as a quantum dot. If size D is larger than 1 nanometer, but smaller or equal than C+1 nanometer, then it is a quantum ring. If size D is larger than C nanometer, then it is a nano-hole. Between 1 and 2 nanometers QD - QR hybrid, between C and C+1 nanometer QR - NH hybrid is detected. Contrary to this, if the shape factor is defined by fuzzy sets based on value of D and parameterized with value of C, then QRs and NHs can be clearly separated and type of hybrid nano structures can be identified as well (eg. QD - QR hybrid or QR - NH hybrid) (Figure 11A). The membership function of output can be seen on Figure 11B. Since it is possible that one or more rules of a rule base are true (meaning that one or more neuron fires), the result is defuzzyfied with centroid method and a real number is obtained. Depending on the range this number fits into, the type may be QD, QR, NH, QD-QR hybrid or QR-NH hybrid. The mapping is described by if-then fuzzy rules that can be seen in Table 2. If only one of the rules is true (or one rule 'fires' in fuzzy terminology) then the object is the nano-structure that belongs to the premise (QD, QR or NH). If two rules are true a hybrid form is obtained.

Input and condition	Output
D < 2 nm	quantum dot
D >= 1 nm and D =< C+1 nm	quantum ring
D > C nm	nano-hole

Table 2
The calculation of the shape factor is described with fuzzy rule base

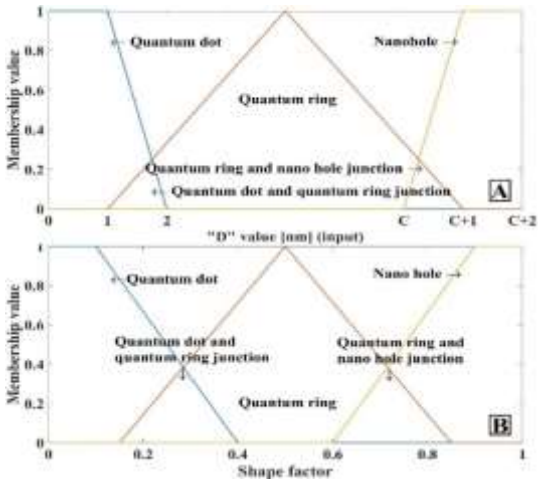


Figure 11

(A) Membership function if geometrical dimensions C and D were used to calculate form factor. Areas with red border represent hybrid nanostructures (quantum dot-quantum ring hybrids or quantum ring-nano-hole hybrids) (B) Membership functions of output. Hybrid structures are labeled on figure.

On Figure 12, the outline of classification with fuzzy inference system is shown. The classification is based on the geometrical features of the nano-structures. It is also shown that to what extent belongs the given sample to a nanostructure class.

The 100% implies that there is a clear type of nano-structure. If the structure under consideration is a hybrid type (eg. a QD - QR hybrid) the result explains the degree of "QDness" and "QRness". This ratio is calculated by relative error of the defuzzified output value to the limit of the relevant interval. The calculation formula is shown below.

$$QD(\%) = \frac{(Df - QR_{min})}{(QD_{max} - QR_{min})} * 100 \tag{5}$$

where D_f is the crisp output of the fuzzy model (defuzzified value), QR_{min} is the lower limit of the "QR interval" and QD_{max} is the upper limit of the same interval. QD(%) is the degree of "QDness". The formula for "QRness" is shown below.

$$QR_1(\%) = 100 - QD(\%) = \frac{(QD_{max} - Df)}{(QD_{max} - QR_{min})} * 100 \quad (6)$$

where $QR_1(\%)$ is the proportion of the quantum ring in percent and the meaning of the further variables is the analogous to those of 5.

The ratio for quantum ring-nano-hole hybrids are calculated in the same way:

$$QR_2(\%) = \frac{(Df - NH_{min})}{(QR_{max} - NH_{min})} * 100 \quad (7)$$

where D_f is the crisp (defuzzified) output of fuzzy model, NH_{min} is the lower limit of nano-hole interval and QR_{max} is the upper limit of quantum ring interval QR(%) is the degree of quantum "ringness". The ratio of nanohole is shown below:

$$NH(\%) = 100 - QR_2(\%) = \frac{(QR_{max} - Df)}{(QR_{max} - NH_{min})} * 100 \quad (8)$$

where NH(%) is the degree of "NHness" and the other parameters are the same as in equation 7. This classification has two advantages. It is simple and accurate solution.

There is an additional possibility in the classification of nanostructures. If the classification is based on controllable technological parameters then it can be interpreted as an engineering tool. These parameters are the substrate temperature, the component flux, the background pressure of arsenic, the time interval and temperature of annealing. On Figure 13, the results of fuzzy inference classification to the nano-structures can be seen. The meaning of the 100% proportion is the same as in the case of geometrical classification and the proportion calculation of the nanostructure junctions are also the same.

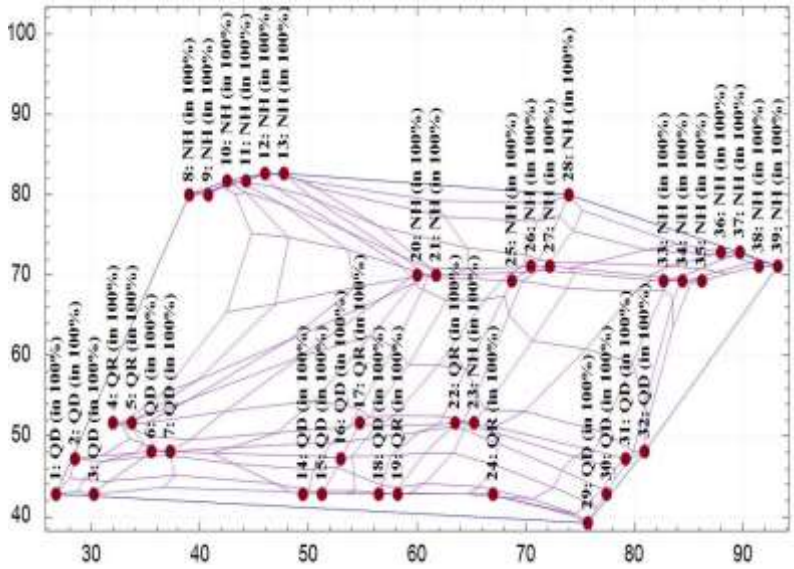


Figure 12

Result of clustering of nanostructures by fuzzy inference system based on the geometrical dimensions

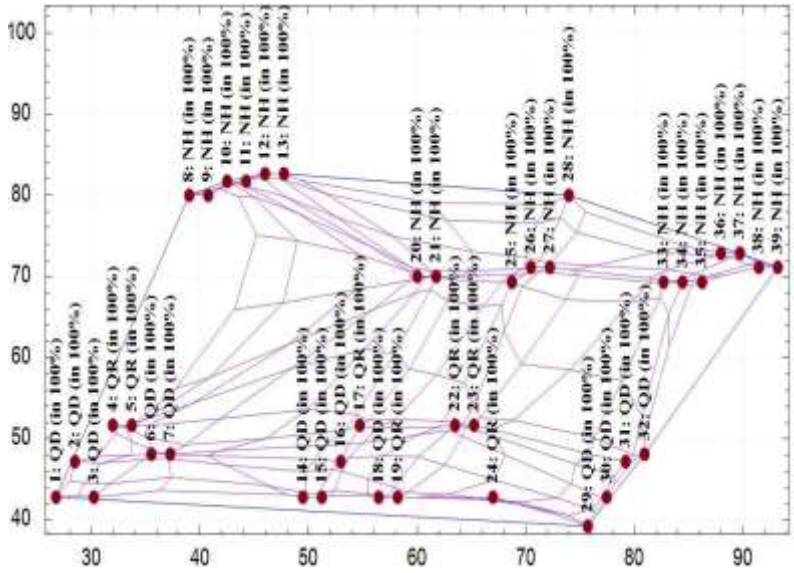


Figure 13

Result of the clustering of nanostructures based on technological parameters substrate temperature and arsenic background pressure

There is an anomaly in line 23, because the algorithm mistakenly identifies NH as QR. This sample is unusual as NHs are formed generally above 500 ℃ and under

arsenic pressure of 10^{-7} Torr. In sample 23, the NH was formed at low substrate temperature $T_{substrate}$=300 °C and at medium arsenic pressure P_{as}= 10^{-6} Torr, after 1 minute of annealing at T=300 °C [25].

Conclusion

In this paper, we examined the clustering of nano-structures forming in an self-organizing process. The shape, size and spatial distribution of these structures determine the characteristics of devices that built of them. For this reason, it is important to know that how the technical parameters of manufacturing process effect each type of nanostructure. In this paper, a self-organized clustering algorithm based on shape factor was examined. In addition to this a few possible clustering methods were outlined. First, a clustering algorithm based on Kohonen SOM, then we combined this method with fuzzy inference algorithm. In the other case, two approaches were studied. In the first approach, the shape factor is determined by geometrical dimensions of the nano-structures. According to the second approach, the shape factor is determined by directly adjustable technological parameters (substrate temperature, arsenic background pressure, etc.).

The above described process is verified with an example of a QD based solar cell. A possible realization of the QD contained solar cell, mentioned in the introduction is in the work of Kerestes et. al. [3]. In this example, the diameter of the base circle is between 10 and 18 nm and its height is between 2 and 5 nm. The average cell surface density is $9.7*10^{10}$ 1/cm^2. Related to our approximations, at this sample the substrate temperature was between 252 and 254 ^{o}C, the gallium flux was 0.025 ML/s. The ambient pressure of the arsenic component was between $5.01*10^{-5}$ and $1*10^{-4}$ Torr and the annealing time was 10 minute with a 350 ^{o}C annealing temperature applied.

References

[1] Z. Zheng, H. Ji, P. Yu, Z. Wang, "Recent Progress Towards Quantum Dot Solar Cells with Enhanced Optical Absorption," *Nanoscale Res Lett*, vol. 11, pp. 266-234, May. 2016.

[2] W. Jiang; C. Siming; S. Alwyn; L. Huiyun, "Quantum dot optoelectronic devices: lasers, photodetectors and solar cells," *Journal of Physics D Applied Physics*, vol. 48, p. 363001, 2015.

[3] C. Kerestes, S, Polly, D. Forbes, C. Bailey, A. Podell, J. Spann, P. Patel, B. Richards, P. Sharps, S. Hubbard, "Fabrication and analysis of multijunction

solar cells with a quantum dot (In)GaAs junction," *Progress in Photovoltaics Research and Applications*, vol. 22, no. 11, pp. 1172–1179, 2013.

[4] A. Luque, A. Martí, C. Stanley, "Understanding intermediate-band solar cells," *Nature Photonics*, vol. 6, pp. 146–152, February 2012.

[5] T. NODA, T. MANO, M. ELBORG, K. MITSUISHI, K. SAKODA, "Fabrication of a GaAs/AlGaAs Lattice-Matched Quantum Dot Solar Cell," *J. Nonlinear Optic. Phys. Mat.*, vol. 19, no. 4, pp. 681-686, 2010.

[6] A. Benahmed, A. Aissat, A. Benkouider, J. P. Vilcot, "Modeling and simulation of InAs/GaAs quantum dots for solar cell applications," *Optik - International Journal for Light and Electron Optics*, vol. 127, no. 7, pp. 3531–3534, April 2016.

[7] N. Koguchi, S. Takahashi, T. Chikyow, "New MBE growth method for InSb quantum well boxes," *Journal of Crystal Growth*, pp. 688-692, 1991.

[8] S. Sanguinetti, N. Koguchi, "Droplet Epitaxy of Nanostructures," in *Molecular Beam Epitaxy: From Research to Mass Production, 1st Edition*. Waltham, MA,USA: Elsevier Science, 2013, pp. 95-111.

[9] M. Farkas, L. T. Kóczy, Á. Nemcsics, "A hybrid approach on dimension reduction and fuzzy clustering of droplet epitaxial grown quantum structure experiments," , 3rd Symposium on Computational Intelligence, Győr, 2009.

[10] R. Rojas, "Kohonen Networks," in *Neural Networks, A Systematic Introduction*. Berlin, Germany: Springer-Verlag, 1996, pp. 391-412.

[11] T. Kohonen, *Self-Organizing Maps*, 3rd ed.: Springer, 2001.

[12] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[13] L. T. Kóczy, D. Tikk, *Fuzzy systems*. Budapest, Hungary, 2007.

[14] S. Wang, F.L. Chung, S. HongBin, H. Dewen, "Cascaded centralized TSK fuzzy system: universal approximator and high interpretation," *Applied Soft Computing*, vol. 5, no. 2, pp. 131-145, January 2005.

[15] S. Preitl R-E. Precup, "Stability and Sensitivity Analysis of Fuzzy Control Systems. Mechatronics Applications ," *Acta Polytechnica Hungarica*, vol. 3, no. 1, pp. 61-76, 2006.

[16] D. Ichalal, B. Marx, J. Ragot, D. Maquin, "Fault Detection, Isolation and Estimation for Takagi–Sugeno Nonlinear Systems," *Journal of the Franklin Institute*, vol. 351, no. 7, pp. 3651–3676, July 2014.

[17] H-Y. Li, R-G. Yeh, Yu-Che Lin, Lo-Yi Lin, Jing Zhao, Chih-Min Lin, Imre J. Rudas, "Medical Sample Classifier Design Using Fuzzy Cerebellar Model Neural Networks," *Acta Polytechnica Hungarica*, vol. 13, no. 6, pp. 7-24, 2016.

[18] N. Siddique, H. Adeli, *Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing*.: John Wiley & Sons, 2013.

[19] Á. Nemcsics, A. Stemmann, J. Takács, "To the understanding of the formation of the III–V based droplet epitxial nanorings," *Microelectronics Reliability*, vol. 52, pp. 430-433, 2012.

[20] Á. Nemcsics, B. Pődör, L. Tóth, J. Balázs, L. Dobos, J. Makai, M. Csutorás, A. Ürmös, "Investigation of MBE grown inverted GaAs quantum dots," *Microelectronics Reliability*, vol. 59, pp. 60-63, 2016.

[21] Á. Nemcsics, Ch. Heyn, A. Stemmann, A. Schramm, H.Welsch, W. Hansen, "The RHEED tracking of the droplet epitaxial grown quantum dot and ring structures," *Materials Science and Engineering B*, vol. 165, pp. 118-121, 2009.

[22] Ch. Heyn, A. Stemmann, A. Schramm, H. Welsch, W. Hansen, Á. Nemcsics, "Faceting during GaAs quantum dot self-assembly by droplet epitaxy," *Applied Physics Letters*, vol. 90, no. 20, p. 203105, 2007.

[23] Ch Heyn, T. Bartsch, S. Sanguinetti, D. Jesson, W. Hansen, "Dynamics of mass transport during nanohole drilling by local droplet etching," *Nanoscale Research Letters*, pp. 10-67, Dec. 2015.

[24] T. Kuroda,T. Mano,T. Ochiai,S. Sanguinetti, K. Sakoda,G. Kido, N. Koguchi, "Optical transitions in quantum ring complexes," *PHYSICAL REVIEW B*, vol. 72, p. 205301, 2005.

[25] K.H.P. Tung, H.W. Gao, N. Xiang, "Time evolution of self-assembled GaAs quantum rings grown by droplet epitaxy," *Journal of Crystal Growth*, vol. 371, no. 15, pp. 117-121, 2013.

Revealing Influencing Factors of Check-in Time in Air Transportation

Enikő Nagy, Csaba Csiszár
Budapest University of Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Department of Transport Technolgy and Economics, Stoczek utca 2, 1111 Budapest, Hungary
eniko.nagy@mail.bme.hu, csiszar.csaba@mail.bme.hu

Abstract: Air passengers are particularly faced with uncertainty during their travel. Information regarding the expected check-in time is not sufficient enough. In many cases, there is no infrastructure to measure the time of check-in process and to inform passengers. The aim of our research was to elaborate a method based on historical data in order to reveal the influencing factors and their effects on time elements (queuing and service time). We have considered various air-carrier operational types, periods of the year and destinations. The cases of each type and their combinations have been fully investigated. The most important influencing factors are: passenger numbers, baggage to passenger ratio, ratio of wheelchair passengers and the number of open check-in counters. The results serve as input data for prediction of check-in time in a personalized passenger information service.

Keywords: airport check-in; historical data; passenger queues; regression analysis

1 Introduction

Air travel requires more fatiguing preparation and causes more frustration than other modes of transportation due to stochastic process elements. Airport check-in queuing time is the sixth factor that causes discomfort for passengers according to a study [1]. Additionally, queues have negative impact on customers' satisfaction [2] and if the queues are too long, some passengers may even miss their flight [3]. Stress and uncertainty during travel are often caused by lack of information (e.g. about check-in time). These negative effects are more significant in case of incidents as delays, cancellations, long queues at check-in, etc. Passengers expect a smooth door-to-door travel experience during the entire journey [4].

In most of the airports there is no infrastructure to measure the check-in time elements, which is the basis for information provision. However, it would be useful to know, in advance, whether it is necessary to arrive earlier to the airport

even if it cannot be measured directly. In accordance with check-in, some studies [5] [6] [7] dealt with the modeling and the optimization of check-in and cumulative diagrams were used in order to model the operation of check-in counters. In [8] different queuing models were analyzed and compared. In [9] it has been found that the service quality is mainly driven by the number of available check-in counters, the dynamic arrival rate of passengers, and the distribution of the service time. According to the study [10] the number of needed check-in counters depends on the average speed of passenger flows through the check-in points and average check-in service time. In [11] the optimal number of check-in counters has been determined. These studies considered the check-in time depending on the available infrastructure but did not deal with the characteristics of flight or the composition of the passengers as influencing factors. Additionally, the study [12] stated that queuing theory is too restricted to predict and calculate queuing times. However, the model of integrated database is available to provide sufficient information for air passengers [13]. Based on the literature review it has been found that airport service time has strong effect on quality of service. Currently, only very few studies regarding check-in time analysis and the revealing of influencing factors are available.

In our research we have elaborated an analysis method (grouping and slicing method), searching for the influencing factors of check-in time. We also produced a correlation analysis for the same database. It is widely used as dependences not only between pairs of variables, but between larger groups of variables can be quantified in this manner [14]. We compared the results of both methods and highlighted those, check-in time, influencing factors that are determined by both methods. The purpose was to identify the basic and specified properties of the flight which influence the check-in time and to determine their effects. The following initial hypotheses were supposed:

Check-in time depends on the properties of the flight, which are the following:

- Basic (static) properties: type of airline, season, destination
- Specific (semi-dynamic) properties: number of passengers, baggage/passenger ratio, ratio of wheelchair passengers, number of used lanes.

According to the revealed correspondences the check-in service and queuing time are to be calculated without any immobile measuring infrastructure using only the flight characteristics. It is useful for both the airports 2-3 days prior to the scheduled flight to allocate the resources (e.g. check-in counters) and for the passengers during their preparation for the journey.

Section 2 summarizes the data collection method and the database structure of the analysis. Section 3 describes the elaborated analysis method. In Section 4 the results, in Section 5 a discussion is provided and the conclusions are drawn in the final section.

2 Data Collection

The airport check-in process has been disaggregated and analyzed. Time of the entire, completed check-in process (t_c) is calculated as a sum (1) of the following time values:

- Queuing time (t_q)
- Service time (t_s)

$$t_c = t_q + t_s \tag{1}$$

At the airport, that provided the sample data (Budapest Airport - BUD), the process of passenger check-in and baggage drop-off are handled together at the same counter. Passengers for a certain flight are standing in a single queue and proceed to the counters immediately before the check-in process. We have focused only on check-in aided by personnel.

2.1 Data Recording Steps

The sample database contained the following items:
Flight data: from Airport Operational Database (AODB), in reference to the check-in process (e.g. number of passengers, number of pieces of baggage, etc.).
Queuing data: recorded check-in time data (e.g. queuing and service time) between January 2013 and August 2016. The data recording was carried out by the employees of BUD through an Android mobile application that registered the arrival and leave to/from the check-in counter. The analyzed database contains 13,400 check-in time records belonging to 424 flights.

One employee registered either one flight or one flight group (in case of common check-in) trough continuous monitoring the dedicated check-in counters. Premium service counters were not included in data collection at Budapest Airport as passengers of first/business classes or frequent flyers are handled in a separate queue. In this way, the method ensured that all the passengers of one flight were standing in a single queue. The person had to monitor only the end of the queue for arriving passengers and all the dedicated check-in counters for leaving passengers. The flow chart of the data recording operations and the display of the application are summarized on Figure 1.

Flight details are uploaded from AODB /1/. The flight resource group /3/ depends on uploaded data, current date and time. Only the flights concerned in the near time window are available for selection. The employee monitors whether any passenger arrives to the queue or leaves from the counter /6/. In case of an arriving passenger, with the ‘Add passenger’ button the timestamp is registered and the queue length is increasing /7/. In the case of a leaving passenger, the button of the number of the check-in counter - from which the passenger is leaving - is pressed and the timestamp is registered /8/. A new counter can be opened with the long press of button ’X’ /10/. With the same method, it can be closed (deleted) /11/. If

all the passengers have left and all the counters are closed, data are exported to the created database /12/.

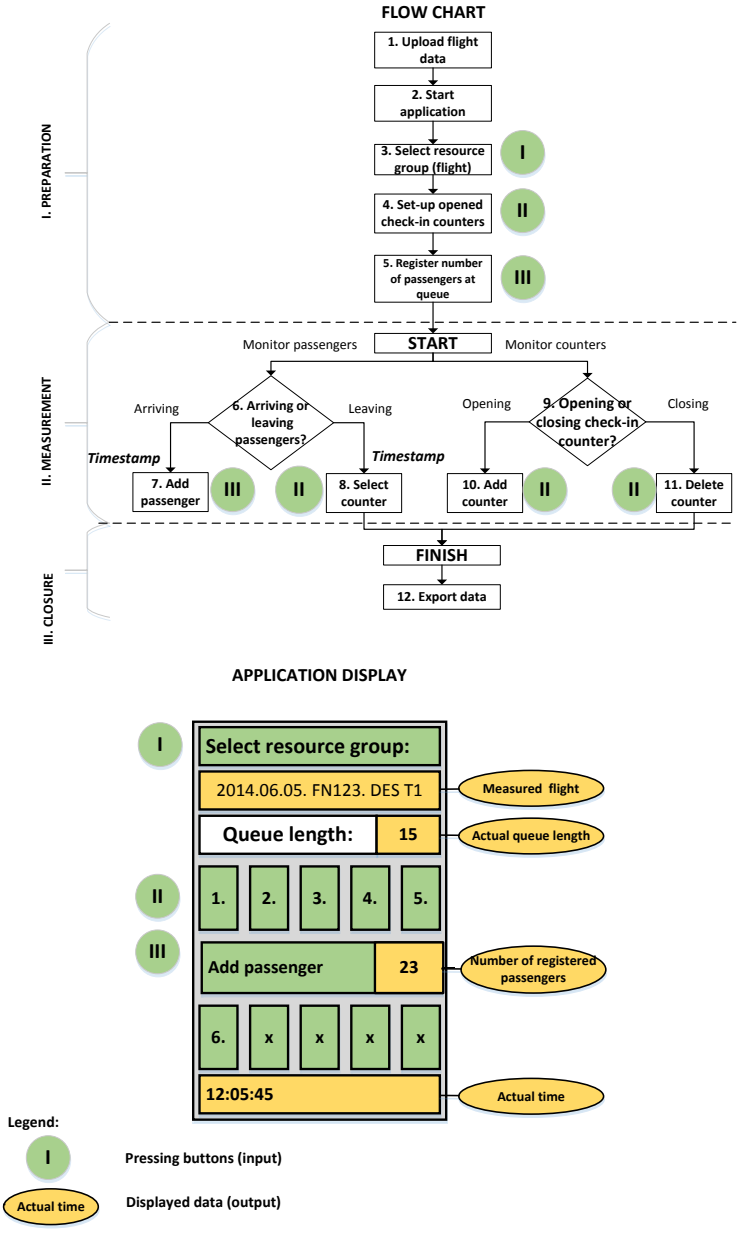


Figure 1
Flow chart of data recording and the application display considering the practice of BUD

2.2 Database Structure

Figure 2 illustrates the simplified database structure.

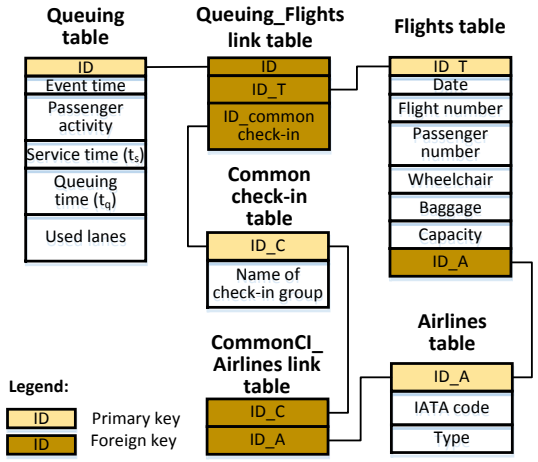


Figure 2
Simplified structure of the database considering the practice of BUD

The model contains only the attributes that are necessary to understand the analysis method. Table 1, 2 and 3 contains the explanation and data type of attributes with an example for better understanding.

Table 1
Structure of Queuing table

Name of attribute	Explanation	Data type	Example
ID	ID of recording	counter	136
Event time	Date and time of event	date	2013.02.12. 18:38
Passenger activity	Arriving (A) to queue or departing (D) from check-in counter	text	D
Service time (t_s)	(in seconds)	numeric	152
Queuing time (t_q)	(in seconds)	numeric	205
Used lanes	Number of opened counters	numeric	3

Table 2
Structure of Flights table

Name of attribute	Explanation	Data type	Example
ID_T	ID of traditional check-in	counter	1
Date	Scheduled date	date	2013.01.11.
Flight number	Master flight number	text	FN123
Passenger number	Number of passengers	numeric	147
Wheelchair	Number of passengers with reduced mobility	numeric	2
Baggage	Number of pieces of baggage	numeric	80
Capacity	Seat capacity	numeric	150

Table 3
Structure of Common check-in table

Name of attribute	Explanation	Data type	Example
ID_C	ID of common check-in	counter	1
Name of check-in group	Fictitious name for airline groups using common check-in	text	GROUP1

The 'Queuing' table contains the measured data, whereas 'Flights' Table contains the specific characteristics of flights. The 'Airlines' table contains the IATA codes and the type of airline (low-cost or traditional). Grouping of airlines was based on their business model (as airline type).

3 Analysis Method

We have elaborated an analysis method in order to reveal the influencing factors of check-in time.

1. We introduced a grouping and slicing method according to the static and semi-dynamic properties of a flight.
2. The introduced method was adapted for BUD. We calculated the statistical properties (minimum/maximum values, medians, quartiles, mean values) of each sub-group in order to identify the influencing factors.
3. We carried out correlation and regression analysis regarding all the basic and semi-dynamic properties of flights. The results of correlation and regression analysis as well as the results of slicing method have been compared.

3.1 Grouping and Slicing Method

The marking system of statistical values (e.g. mean value) has been introduced as in Figure 3.

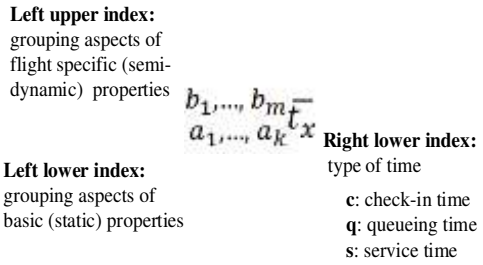


Figure 3
Marking system of statistical values

The basic (static) and specific (semi-dynamic) properties of the flights have been classified in Table 4. It summarizes the notation as well. Both the number of grouping aspects (k, m) and the number of categories in case of aspects (l_k, n_m) can be extended. In this way a flexible method has been created. The presented grouping aspects and the categories are based on the available types in a middle-sized airport. Grouping aspects are independent.

In cells of the grouping aspect the corresponsive database table and column names are also indicated by italic. The grouping is to be performed by these columns either directly or by logical implications / calculated expressions.

Baggage/passenger ratio reflects the amount of pieces of baggage per person. The number and intervals of categories have been determined by our practical experience.

Table 4
Variables and their values used during grouping

	Vari-able	Grouping aspect/table-column	Values	Meaning of values (groups)
Basic (static) properties	**a_1**	type of airline/ *Airlines-Type*	**0**	all airlines
			1	traditional airline
			2	low-cost airline
			l_1	...
	a_2	type of destination /*Flights-Flight number*	**0**	all destination
			1	Europe
			2	not-Europe
			l_2	...
	a_3	season / *Queuing-*	**0**	all seasons

		Event time	**1**	winter
			2	summer
			l_3	**...**
	a_k	k. grouping aspect	**l_k**	**…**
Specific (semi-dynamic) properties	**b_1**	passenger number / *Flights-Passenger number*	**0**	all data
			1	0-50 passenger
			...	...
			4	150-200 passengers
			n_1	...
	b_2	baggage/ passenger ratio /*Flights-Baggage* and *Passenger number*	**0**	all data
			1	0-0.25
			...	...
			8	1.75-2
			n_2	...
	b_3	ratio of wheelchair passengers / *Flights-Wheelchair*	**0**	all data
			1	0-1%
			...	...
			7	10%
			n_3	...
	b_4	Used lanes (check-in counters) / *Queuing – Used lanes*	**0**	all data
			1	1
			…	…
			5	5
			n_4	…
	b_m	m.grouping aspect	**n_m**	…

3.2 Adaptation of Grouping and Slicing Method for BUD

In the case of BUD k=3 and m=4. Accordingly, number of 'dimensions' as grouping aspects are 3 and 4. The number of options (as different data elements) are l_k and n_m.

Several statistical values regarding check-in service and queuing time have been calculated for the sub-groups according to basic (static) properties. Mean values have been calculated for the sub-groups according to specific (semi-dynamic) properties.

3.3 Correlation and Regression Analysis

The same variables (a_k, b_m) have been used for the correlation and regression analysis as in grouping and slicing method. Qualitative variables (a_1, a_2, a_3) are indicated as dummy variables based on Table 5. Final results are presented in this

paper. The results of the regression analysis can be applied for prediction method regarding check-in time.

Table 5
Dummy variables

a_1	type of airline	traditional	1	Dummy 1
		low-cost	0	
a_2	type of destination	Europe	1	Dummy 2
		not-Europe	0	
a_3	season	winter	1	Dummy 3
		summer	0	

4 Results

4.1 Results of Grouping and Slicing Method

Influence of Basic (Static) Properties

Statistical values (minimum, maximum, median, first quartile, third quartile) have been calculated for all the combinations of basic properties and the results (in seconds) are summarized in Table 6, Figure 4 and 5. Results are provided only for the data and its sub-groups that are available in the recorded sample database. The calculated statistical values were examined and compared in group-pairs in order to determine dependency. Based on this comparison t_s depends on a_1 ,a_2, b_2, b_3 variables, whereas t_q depends on a_2, a_3, b_1, b_2, b_3 variables. The minimum and maximum values show which airline type, destination type and season performs as the best and the worst.

Table 6
Mean values according to basic (static) properties

Type of time	**All data**	**Min**		**Max**	
$\mathbf{t_q}$	678.53	$\mathbf{t_q}^{000}_{111}$	395.58	$\mathbf{t_q}^{000}_{122}$	946.32
$\mathbf{t_s}$	88.3	$\mathbf{t_s}^{000}_{212}$	62.21	$\mathbf{t_s}^{000}_{121}$	120.76
$\mathbf{t_c}$	766.83	$\mathbf{t_c}^{000}_{111}$	484.96	$\mathbf{t_c}^{000}_{122}$	1057.12

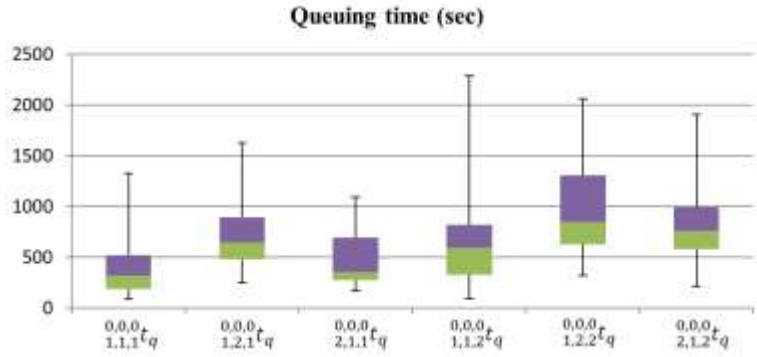


Figure 4
Statistical values for queuing time

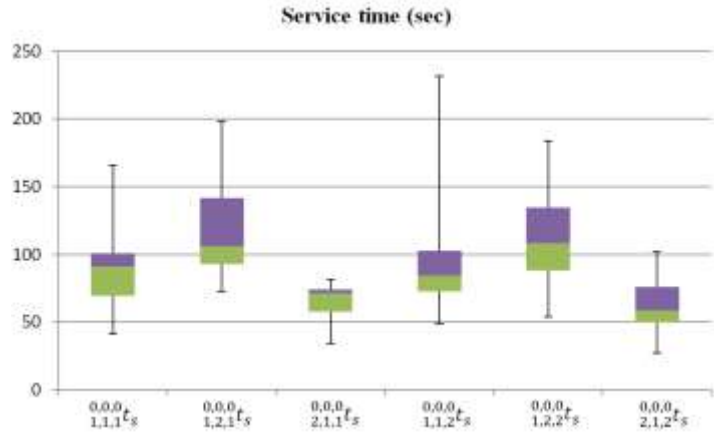


Figure 5
Statistical values for service time

Influence of Flight-specific (Semi-dynamic) Properties

Average values have been calculated for the semi-dynamic properties of the flights and the results (in seconds) are summarized in Figure 6 and 7. The values belonging to certain grouping aspects are indicated as independent variables on the horizontal axis of the diagrams. The functions of polynomial trend lines are also presented on the figures.

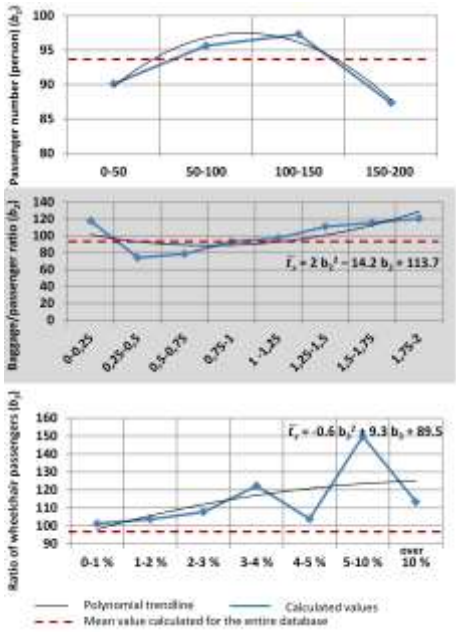


Figure 6
Mean values of service times

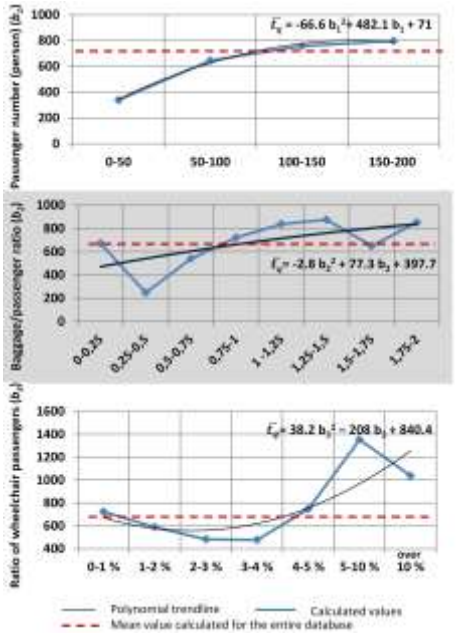


Figure 7
Mean values of queuing times

4.2 Results of Correlation and Regression Analysis

The result of correlation analysis in case of service and queuing time are summarized in Table 7.

Table 7

Correlation results

	t_s	t_q	a_1	a_2	a_3	b_1	b_2	b_3	b_4
t_s	1								
a_1	0.286	-0.200	1						
a_2	-0.234	-0.402	0.169	1					
a_3	0.069	-0.194	0.114	0.053	1				
b_1	-0.155	0.434	-0.450	-0.197	-0.113	1			
b_2	0.378	0.341	0.108	-0.384	-0.071	0.079	1		
b_3	0.150	0.083	0.073	0.013	0.005	0.180	0.067	1	
b_4	0.210	0.191	0.015	-0.267	-0.077	0.688	0.235	0.191	1

highlighted background: medium correlation (absolute value >0.3)

Table 8

Regression statistics

	Service time	Queuing time
r	0.58	0.66
r^2	0.34	0.43
Adjusted r^2	0.32	0.42
Standard error	21.56	301.59

According to the value of adjusted r^2, the reliability of the results is medium-strength both in case of service and queuing time calculation.

After removing those variables that have no or very weak influence, we recalculated the results. Final results are summarized in Table 9 and 10, where the elements of the equation are highlighted with cultured background**.**

Table 9

Results of statistical analysis (service time)

		Coefficients	Standard error	t value	p value
Intercept	**Name of variable**	74.33	6.93	10.73	$1.83*10^{-22}$
a_2	type of destination	-7.42	3.47	-2.14	0.034
b_1	passenger number	-0.31	0.043	-7.68	$3.18*10^{-13}$

$\mathbf{b_2}$	baggage/ passenger ratio	21.17	4.6435	4.56	$7.83*10^{-6}$
$\mathbf{b_3}$	ratio of wheelchair passengers	563.96	204.99	2.75	0.006
$\mathbf{b_4}$	used lanes	11.44	1.8	6.36	$8.86*10^{-10}$

Based on the calculations the service time can be predicted according to the equation (2):

$$t_s=74.33 - 7.42*a_2 - 0.31*b_1 + 21.17*b_2 + 563.96*b_3 + 11.44*b_4 \quad (2)$$

Table 10

Results of statistical analysis (queuing time)

		Coefficients	Standard error	t value	p value
Intercept	**Name of variable**	287.51	122.11	2.35	0.019
a1	type of airline	303.41	110.48	2.75	0.006
a2	type of destination	-310.23	54.46	-5.7	$3.3*10^{-8}$
a3	season	-150.56	52.84	-2.85	0.005
b1	passenger number	6.51	0.72	8.98	$5.7*10^{-17}$
b2	baggage/ passenger ratio	300.96	65.6	4.59	$7*10^{-6}$
b4	used lanes	-181.53	29.14	-6.23	$1.9*10^{-9}$

Queuing time can be predicted according to the equation (3):

$$t_q=287.52 +303.41*a_1–310.3*a_2–150.56*a_3+6.51*b_1+300.96*b_2– 81.53*b_4 \quad (3)$$

4.3 Comparison of the Results

We compared the results of the two methods (grouping and slicing method vs. correlation analysis) in Table 11. In the case of the grouping and slicing method the calculated statistical values were examined and compared in group-pairs. In case of connection between the queuing/service time and the variables, ✓mark is displayed. The strength of the connections has been revealed by correlation analysis. In the table, strong or medium connection has been marked with ✓and weak or very weak connection has been marked with X. A green background shows that both methods resulted in the same dependency, while light green shows if the variable has no influence and dark green if it has. Accordingly, these influencing factors have to be taken into consideration in for a prediction method.

Table 11
Dependencies of times on variables

	Name of variable	Service time (t_s)		Queuing time (t_q)	
		Grouping and slicing	Correlation analysis	Grouping and slicing	Correlation analysis
a_1	type of airline	✓	X	X	X
a_2	type of destination	✓	X	✓	✓
a_3	season	X	X	✓	X
b_1	passenger number	X	X	✓	✓
b_2	baggage/passenger ratio	✓	✓	✓	✓
b_3	ratio of wheelchair passengers	✓	X	✓	X
b_4	used lanes	N.A.	X	N.A.	X

✓: dependency
X: non-dependency
green background: same dependency in both methods
light green: variable has no influence
dark green: variable has influence

5 Discussion

Based on the results of the applied two methods the following have been found:

- Service time (t_s) of low cost airlines is 30-40% lower than in case of the traditional airlines (due to the lower number of check-in baggage and the higher number of prepared boarding passes).
- Queuing time (t_q) does not depend on the type of airline (a_1), but does on the type of destination (a_2).
- Service time (t_s) does not, but the queuing time (t_q) depends on the number of passengers on flight (b_1): the higher is the number of passengers the higher is the queuing time (4).

$$\overline{t_q} = -66.6*b_1^2 + 482{,}1*b_1 + 71 \quad (4)$$

- The higher is the baggage/passenger ratio (b_2) the higher is the service time (t_s) and the queuing time (t_q) (5), (6).

$$\overline{t_s} = 2*b_2^2 - 14.2*b_2 + 113.7 \quad (5)$$

$$\overline{t_q} = -2.8*b_2^2 + 77.3*b_2 + 397.7 \quad (6)$$

According to the results obtained from capacity analysis model [7], 76 passengers pass with 118 pieces of luggage through one check-in counter per hour. In this case $\overline{t_s}$ =47.4 sec/passenger. However, we found that in the case of BUD, according to (5), $\overline{t_s}$ = 96.5 sec/passenger belongs to the same baggage/passenger ratio.

Deviation between the two values could be that in case of [7], the results are calculated by fuzzy logic model, while in our case it is a real measured data, where service time could depend on the passengers' behavior. Additionally, BUD handles significant through – check-in requirements as well.

- According to correlation analysis it has been found that there is very weak correlation between check-in time (t_c) and ratio of wheelchair passengers (b_3) on the flight, however the grouping method showed an increased tendency. The false result is the consequence of the small sample (only 10% of the analyzed flights had wheelchair passengers on board).
- The smallest service-time (t_s) belongs to the low-cost airline, summer period, European destination combination. It is because low-cost airlines try to minimize the cost by forcing passengers to travel with less baggage and with a prepared boarding pass. It reduces the time spent at the check-in counter, in this way airlines have to pay less for the handling companies. Conversely, traditional airlines provide boarding pass printing and baggage check-in at the counters and that takes longer.
- The smallest check-in time (t_c) belongs to the traditional airline, winter period, European destination combination. The result is the consequence of lower queuing time (t_q) of traditional airlines. As the order of magnitude of queuing time is higher than the service time, the huge service time (t_s) does not affect the time of the overall process.
- Non-European destinations require more service and queuing time. An explanation is the baggage/passenger ratio on flight (b_2).

The method is mainly developed for airport operators in order to have a general overview about expected check-in time values. In the case of having information only about the static properties of the flight (e.g. 2 weeks before departure), the average check-in time values calculated by grouping and slicing method can be used for informing passengers. In case of having information about the semi-dynamic properties of flight, using the equations of the regression analysis gives more precise data. With the combination of these two methods, the expected check-in time can be predicted notably. These influencing factors are available

from flight schedule and seat reservation system. The calculated results could also serve as input data for passenger information applications (e.g. airport application, airline application or integrated application) on mobile devices to display the expected check-in time for passengers. It would be provided as public information; similarly, like an actual flight schedule. Those who have downloaded the application could check the information from anywhere. Data reliability could be improved by the further analysis of the method and the usage of more historical data. This is our goal in further research.

Conclusion

Our main findings show that the influencing factors of check-in time and their effects are the following:

- Check-in time slightly depends on the basic (static) properties of the flight and mainly the queuing times depend on the type of destination (a_2)
- Check-in time is influenced by specific (semi-dynamic) properties of the flight, mainly by the number of passengers (b_1) and the baggage/passenger ratio (b_2).

Accordingly, these influencing factors have to be taken into consideration for the case of a prediction model.

Based on the result, the proposed measures, in order to decrease time elements are as follows:

- Reducing queuing time is more effective if the entire check-in time is to be reduced
- Proper information provision (about baggage, gate info etc.) for passengers before check-in, in order to avoid long service time at the counters
- Decrease of service time by the initiation of boarding pass pre-printing and the promotion of travelling without baggage
- Reduction of service time by baggage drop-off in the city (e.g. at airport shuttle bus/ train stations or launching baggage delivery service);
- Opening of more check-in counters (or separated queues), for the case of when more than 5% of the total number of passengers are wheelchair passengers
- Better integration of seat reservation system of airlines and airport operational database in order to calculate the necessary numbers of check-in counters more precisely

- Data management with shorter data transmission cycle time (e.g. between seat reservation systems of airlines and airport databases regarding passenger number, baggage number, etc.).

During the research we learned that more precise determination of category intervals (considering several additional factors) resulted in more accurate results. Furthermore, the grouping aspects may vary depending on the characteristics of the dataset (e.g. analysis of check-in process in case of special items on flight: dogs, ski equipment, bikes etc.).

Further research directions are the amendment of calculations using additional grouping aspects and the elaboration of an advanced prediction method of check-in time, based on these analysis method/results and the historical data being available from different sources but mapping the same physical process. Automation in other sectors of transportation alters the conventional processes and researches are increasingly focusing on its impacts [15] [16]. Therefore, the aviation sector should be prepared for similar challenges. The autonomous airports, in the future, need more precise prediction information concerning check-in time elements, in order to plan their capacity more perfectly. The integration of passenger handling functions (check-in, baggage drop-off, security check, passport control) needs further analysis of the time elements. As a research goal, we are going to build these time elements into an elaborated method. As a future opportunity, the model will be further developed to measure and optimally minimize the security lines at the airport.

Acknowledgement

Authors of the paper would like to thank you for the employees of Operations Department of Budapest Airport Ltd. for the provided data used for our calculation.

SUPPORTED BY THE ÚNKP-17-3-III NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES

References

[1] Greghi, M. F.; Rossi, T. N.; Souza, J. B. G.; Menegon N. L. (2013) Brazilian Passengers' perceptions of Air Travel: Evidences from a Survey, Journal of Air Transport Management 31: 27-31 http://dx.doi.org/10.1016/j.jairtraman.2012.11.008

[2] Katz, L.; Larson, B.; Larson, R. (1991) Prescription for the Waiting-in-Line Blues: Entertain, Enlighten, and Engage, Sloan Management Review 4: 44-53

[3] Lange, R.; Samoilovich, I.; Rhee, B. (2013) Virtual Queuing at Airport Security Lanes, European Journal of Operational Research 225 (1): 153-165 http://dx.doi.org/10.1016/j.ejor.2012.09.025

[4] Esztergár-Kiss, D.; Csiszár, Cs. (2015) Evaluation of Multimodal Journey Planners and Definition of Service Levels Int. J. ITS Res. (2015) 13: 154 http://dx.doi.org/10.1007/s13177-014-0093-0

[5] Tošić, V. (1992) A Review of Airport Passenger Terminal Operations Analysis and Modeling, Transportation Research Part A: Policy and Practice 26 (1): 3-26 http://dx.doi.org/10.1016/0965-8564(92)90041-5

[6] Janic, M. (2003) Modelling Operational, Economic and Environmental Performance of an Air Transport Network, Transportation Research Part D: Transport and Environment 8 (6): 415-432 http://dx.doi.org/10.1016/S1361-9209(03)00041-5

[7] Chang, H.; Yang, C. (2007) Do Airline Self-Service Check-in Kiosk Meet the Needs of Passengers?, Tourism Management http://dx.doi.org/10.1016. (2007)

[8] Kleinrock, L. (1975) Queueing Systems Volume I: Theory. New York: John Wiley & Sons, Inc.

[9] Stolletz, R. (2011) Analysis of Passenger Queues at Airport Terminals, Research in Transportation Business & Management 1(1): 144-149 http://dx.doi.org/10.1016/j.rtbm.2011.06.012

[10] Koray Kıyıldı, R.; Karasahin, M. (2008) The Capacity Analysis of the Check-in Unit of Antalya Airport using the Fuzzy Logic Method, Transportation Research Part A: Policy and Practice 42 (4): 610-619 http://dx.doi.org/10.1016/j.tra.2008.01.004

[11] Bruno, G.; Genovese, A. (2010) A Mathematical Model for the Optimization of the Airport Check-In Service Problem, Electronic Notes in Discrete Mathematics 36: 703-710 http://dx.doi.org/10.1016/j.endm.2010.05.089

[12] Joustra, P. E.; Van Dijk, N. M. (2001) Simulation of Check-in at Airports, Proceedings of the 2001 Winter Simulation Conference

[13] Karádi, D; Nagy, E.; Csiszár, Cs. (2015) Integrated Information Application on Mobile Devices for Air Passengers, 4th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). 536 p., Budapest, Hungary

[14] Ferenci, T; Kovács, L (2014) Using Total Correlation to Discover Related Clusters of Clinical Chemistry Parameters, SISY 2014: IEEE 12th International Symposium on Intelligent Systems and Informatics, Subotica, Serbia, 2014.09.11-2014.09.13 (IEEE) Subotica: IEEE Hungary Section, 2014. pp. 49-54

[15] Tettamanti, T; Varga, I.; Szalay, Zs. (2016) Impacts of Autonomous Cars from a Traffic Engineering Perspective, Period. Polytech. Transp. Eng., Vol. 44, No. 4 (2016), pp. 244-250
http://dx.doi.org/10.3311/PPtr.9464

[16] Földes D., Csiszár, Cs. (2016) Conception of Future Integrated Smart Mobility. Smart Cities Symposium, 26-27 May 2016, Prague, Czech Republic, pp. 29-35
http://dx.doi.org/10.1109/SCSP.2016.7501022