Urban Traffic Congestion Alleviation Relying on the Vehicles' On-board Traffic Congestion Detection Capabilities*

Zoltán Fazekas[1], Mohammed Obaid[2], Lamia Karim[3] and Péter Gáspár[1,4]

[1] HUN-REN Institute for Computer Science and Control (HUN-REN SZTAKI), Kende u. 13-17, H-1111 Budapest, Hungary, {zoltan.fazekas, peter.gaspar}@sztaki.hun-ren.hu

[2] Department of Automotive Technology, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Stoczek u. 6, H-1111 Budapest, Hungary; obaid.mohammed@mail.bme.hu

[3] National School for Applied Sciences (ENSA), Hassan First University of Settat, Avenue de l'université, B.P. 218, Berrechid, Morocco; lamia.karim@uhp.ac.ma

[4] Department of Control for Transportation and Vehicle Systems, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Stoczek u. 2, H-1111 Budapest, Hungary; gaspar.peter@kjk.bme.hu

*This work is dedicated to the memory of Prof. Azedine Boulmakoul, a dear friend and an excellent colleague from Morocco.

Abstract: Traffic simulation experiments were carried out for an urban road network to explore the effect of road vehicles' individual traffic congestion avoidance efforts, in which on-board visual line-of-sight (LoS) exteroceptive sensors (ECSs) and related on-board traffic congestion detection (OTCD) capabilities are put to use on the network level traffic situation. OTCD requires a visual LoS constellation between the subject vehicle and some vehicles in the vehicle queue ahead. The experiments concern themselves with the comparison of undisturbed, disturbed and mitigated traffic. PTV Vissim traffic simulator was used in the experiments. The process of congestion detection, avoidance and mitigation was tentatively modelled via proxy parameters. Two series of experiments are reported herein. A new approach to route planning has been identified and earmarked for future research.

Keywords: traffic simulation; traffic congestion; driver assistance systems; exteroceptive sensors; route planning

1 Introduction

Intelligent road vehicles (IRVs) with their advanced driver assistance systems (ADAS), on-board smart sensors, navigational and info-communication devices have already become part of everyday life in many countries worldwide. Autonomous road vehicles (AVs) have yet to achieve such prevalence [1].

Many of the automotive sensors that appear in, or on-board IRVs and AVs are visual line-of-sight (LoS) sensors. Sensors of this kind can detect obstacles, vehicles, road objects, humans, animals, etc. that they can actually 'see', or in other words, that are in the sensors' visual LoS regions. Mono and stereo cameras, infrared cameras, LIDARs, and radars are exemplars of such sensors. Some of the automotive LoS sensors look within the subject vehicle, in this sense these sensors are proprioceptive sensors, while others, i.e., the ones looking out of the subject vehicle, are exteroceptive sensors (ECSs). In-depth reviews of LoS sensors and their typical applications in vehicular measurements are given in [2-4].

1.1 Mottos and Conjectures

Three relevant quotes are included here as mottos, from [5-7]:

The first motto:

> *'Road accidents and traffic congestion are two critical problems for global transport systems. Connected vehicles and automated vehicles are among the most heavily researched and promising automotive technologies to reduce road accidents and improve road efficiency. However, both automated vehicle and connected vehicle technologies have inherent shortcomings, for example, line-of-sight sensing limitation of automated vehicles' sensors and the dependency of high penetration rate for connected vehicles.'*

The second motto:

> *'Autonomous systems should be designed with the ability to produce several alternative ways to complete their tasks ...*
> *... the system should not only be able to find alternative routes when it encounters traffic jams and blocked roads. If the system realizes that it cannot complete its mission, it should be programmed to transfer the passenger to another mode of transportation – such as taking them to a train station ...'*

The third motto:

> *'For geometric design [of roads], the most useful form of classification is functional classification, as it defines the spectrum of road usage from pure mobility to pure accessibility. This, in turn, supports the selection of the design speed and the design vehicle. These two parameters, in combination with current and anticipated traffic volumes, define geometric standards of horizontal and vertical alignment, and intersections or interchanges and definition of the cross-section.'*

As pointed out in the first motto, road accidents and traffic congestion are critical problems for transport systems. Statistical data on serious and fatal road accidents, see [8], keep frustrating transport stakeholders and common road users worldwide alike. Beside accidents, urban traffic congestion has many other detrimental effects, and these manifests themselves at different scales of the society and economy [9]. Three conjunctures/hypotheses for the present study are included below; the first one logically links road accidents and urban road traffic congestion.

> **Conjecture 1:** A high proportion of serious and fatal accidents occurs in urban areas at times of traffic congestion. Any decrease in accident numbers associated with traffic congestion is likely to lessen also the total figures of serious and fatal accidents.

Motivated also by the above conjecture, ways are sought worldwide to better understand and reduce urban road traffic congestion. To this end, road traffic simulation experiments were carried out – in the frame of the study presented herein – for an urban subnetwork to explore the joint effect of road vehicles' individual rational traffic congestion avoidance efforts. The simulation experiments were de-signed to model traffic congestion that forms due to some unexpected, non-recur-ring traffic incident. As in such cases, the incident itself and the forming congestion remain unreported for some time, typically for some minutes, the deteriorating traffic situation necessitates the application of local congestion detection, avoidance and alleviation approaches. One such approach, proposed in [10], relies on dedicated sensors and enhanced vehicle detection capabilities that are available on-board the road vehicles moving in the traffic, and perceive the build-up of vehicle queues in the vicinity. This approach is revisited herein and the effect of certain parameters are looked at in some detail.

> **Conjecture 2:** Visual LoS exteroceptive sensors – on-board vehicles – and their supporting on-board traffic congestion detection (OTCD) systems could and should play an important role in urban road traffic congestion alleviation, at least locally, and especially before:
>
> a) Network-level traffic congestion alleviation interventions are initiated
>
> b) The effect of these interventions is felt in the road network, and also locally.

Such sensors and such OTCD capabilities – either artificial, or human – on-board vehicles participating in the urban road traffic were assumed and considered in the simulation experiments mentioned above. The experiments concern themselves with the comparison of undisturbed, disturbed and mitigated road traffic.

The local traffic conjunction mitigation approach used herein led to considerable traffic improvements in a number of cases, however, such an improvement cannot always be guaranteed. This is phrased as a conjecture below.

> **Conjecture 3:** Not all urban traffic situations/incidents can be handled, not all transport assignments/missions can be supported by the application of local traffic congestion detection, avoidance and alleviation approaches.

To prove the truth of Conjecture 3, a fairly common, but sensitive transport assignment/task – along an important urban route – is considered. The traffic along the route is strongly hindered by a possibly minor traffic incident. It turns out that the adverse traffic situation formed cannot be promptly resolved, and so the commenced transport task cannot be accomplished via the application of local traffic congestion avoidance and alleviation approaches.

In order to properly tackle similar sensitive transport assignments/tasks, a new objective for route planning and optimization is proposed herein. The route planning and optimization that considers also this new objective should be studied and further researched as it could attract interest from the military, police, and security fields, and find applications in such missions and routing tasks, as well as in fairly common, ordinary routing tasks.

1.2 Modelling Traffic That Involves Vehicles with Vehicle-Queue Detection Capabilities

The disturbance brought about to the urban road subnetwork under study is a sudden, unexpected and non-recurring, possibly minor traffic incident that causes vehicle queues and traffic congestion in the area. The drivers/AVs – relying on their own LoS ECSs and OTCD capabilities – see/detect the traffic queues in the congested area, and try to avoid them. This detection requires an LoS constellation between the subject vehicle/driver and at least some of the vehicles stuck in the queues ahead. The intention to avoid the congested road sections and the resulting re-routing of the vehicles affect the traffic flow in the area and to some extent mitigate the traffic congestion.

Traffic simulation experiments were carried out to model the above outlined complex spatio-temporal process. It, however, was modelled – for reasons given in Section 2.2 – in a simplified way. Two series of traffic simulation experiments, referred to as case studies, concerning the subnetwork were carried out and their simulation results are reported herein. These experiments continue and extend those described in book chapter [10]. Indeed, the urban road subnetwork analyzed herein is the same as one of the three road subnetworks analyzed in the cited book chapter, however, the traffic conditions, namely the obstacle locations, and/or some parameters used in the simulations do differ.

Furthermore, in the cited book chapter, strong emphasis was put on the alternatives of OTCD based traffic mitigation approach, in particular, on the various communication means, channels, services and options that help drivers/AVs to keep away from congested urban areas. These alternatives were found superior – on a network-level – to OTCD based traffic congestion mitigation, particularly for highly developed urban areas, and after an initial delay. However, application niches and target user groups for OTCD systems were identified in book chapter.

Accepting the conclusions drawn in [10], the focus of the present study is to answer the following questions.

- What happens to the traffic if car drivers/AVs still use the aforementioned individualistic traffic congestion avoidance approach?
- Do the OTCD based individual congestion avoidance efforts add up and alleviate traffic congestion in the area?
- What happens to the traffic elsewhere in the road subnetwork?

Clearly, sharing information on detected vehicle queues via available communication channels/infrastructure is an excellent way forward in network-level traffic congestion mitigation, but again this is applicable only in highly developed urban areas, see [11]. Herein, however, these communication possibilities are set aside, but the general approach of decentralized traffic congestion control – advocated in the above cited article – is followed.

1.3 Further Related Literature

One way of dealing with urban road traffic congestion is building new roads and improve the transport infrastructure and related communication facilities, and transport services in general. For various reasons, e.g., for economic, architectural, geographic, topographical reasons, this way of improving the traffic situation might not be possible. In such cases, alternative traffic congestion mitigation measures are required.

The spread of AVs is usually deemed beneficial also for easing traffic congestion, see e.g., [12]; nevertheless, according to [13], the urban transportation infrastructure needs to be ready for the sustainable deployment of AVs. The author opines that a key aspect of this readiness is to introduce certain modifications in road design and adjust traffic control accordingly. The author investigated the adjustments needed for interrupted traffic flow, and modelled the altered circumstances and conditions using the PTV Vissim traffic simulator [14] [15]. Results published in [13] indicate that even if the recommended modifications are fully adhered to and properly implemented, AVs alone, i.e., without connectivity features, do not improve the capacity of the road network considerably. Simulation results presented in [12], on the other hand, show that the capacity of an urban road network increases quasi-linearly with AV-penetration; maximum traffic flow, for instance, increased by 25% as the mentioned parameter was changed from 0% to 100%.

The authors of [16] opine that the study of OTCD capabilities have been neglected in the literature. One of the reasons for this disinterest could be that visual detection of vehicle-queues in the vicinity is easy for experienced and/or local drivers, but it is quite challenging to devise, formulate, implement and validate reliable real-time OTCD algorithms that can be deployed in ADAS/AD subsystems. In contrast to this disinterest, a considerable number of traffic congestion detection methods and systems that rely on data from static traffic sensors were reported during the last decade, e.g., [17] [18]. Several of the methods and systems proposed in this period make use of artificial intelligence. It is quite likely that some of these methods and systems are adaptable for the tasks of OTCD based avoidance and mitigation. Another reason for the mentioned disinterest is that in the era of advanced vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications [19] [20], communication-based solutions are serious rivals to OTCD systems.

In case of AVs, the OTCD capability should technically build upon some more common LoS detection capabilities provided by certain ADAS/AD subsystems and functions onboard. The research on such – in the present context subordinate – detection capabilities is reviewed in [21].

Two frameworks presented therein are relevant for the present study. The first one concerns traffic conflict-based vehicle intelligence, while the second one is a cooperative control framework. Both frameworks refer to tasks, subsystems and capabilities, e.g., detection of traffic participants and nearby obstacles, managing travel information, making use of the vision, radar and multi-sensor fusion subsystems, predicting traffic conflicts and escaping/avoiding the more serious ones, which are related to the present topic. Escaping unexpected incidents, or even disasters, by autonomous systems, such as AVs, was the topic of [6] from which our second motto has been borrowed. The authors of the cited paper argue that despite the unpredictability of such events, handling of the arising situations is often possible.

According to [22], numerous traffic simulators were developed in the last decades. In conjunction with study described in [13], the Vissim traffic simulator has been already mentioned. For the traffic simulation experiments presented herein, also this traffic simulator was chosen[1] and used. Vissim microscopic traffic simulator [14] [15] runs different simulation factors and parameters at a complex microscopic level, and runs the simulation with high number of iterations until achieving near real-life conditions. It can handle both real and user defined maps to study a number of different conditions. It is often the simulator of choice for real road networks [26] as it can run complex equilibrium assignments.

[1] When taking this decision, also the planned application of Vissim-MATLAB cooperation via the COM interface, see [23-25], was considered. See details in Subsection 2.2.

The network-level benefits of navigation and route guidance systems (NRGSs) used by drivers/vehicles were assessed in [27]. In the traffic simulations presented therein, such systems were used and relied on by growing proportions of drivers/ vehicles. The authors compared total travel times, as well as total delays within the road network for eleven different parametrized sub-scenarios (PSSs.)

In the simulations presented in [28], the authors used Vissim's Dynamic Traffic Assignment (DTA) simulation option [15]. The PSSs drawn on in their assessment were parametrized by the percentage of the traffic demand equipped with NRGSs. According to [28], there are several benefits of employing DTA in simulations. However, they opine that it makes the simulation task more complex, furthermore, it necessitates additional parametrization, calibration and validation of the model.

Due to the lack of built in LoS filtering support in the commonly used versions of the Vissim, the complex process of vision-based traffic congestion detection and avoidance presented herein was tentatively modelled via proxy parameters. A similar proxy-based modelling approach was taken in [23] [24] regarding the utility assessment of a driver assistance system and the effectivity of a traffic control, respectively.

1.4 Scope and Structure of The Manuscript

The joint effect of the individual OTCD capabilities is assessed in the simulation-based experiments presented herein. The question of how such artificial capabilities can be reached and implemented has not been addressed herein, but clearly such capabilities are within the reach of the present-day, leading-edge automotive technology [3]. The simulation experiments concern themselves with the undisturbed, disturbed and mitigated traffic within an urban road subnetwork. Even though the traffic is modelled in a simplified way, experimentation with the proxy parameters helps to understand their roles, and relates some of them to the urban texture, see [29].

The rest of the manuscript is organized as follows. Section 2 presents the simulation experiments concerning the OTCD based urban traffic congestion mitigation. In Subsection 2.1, the simulation environment is briefly described. Then, in Subsection 2.2, the simulation approach is presented and justified. In Subsection 2.3, the target road subnetwork is introduced. In Section 3, two case studies concerning unexpected traffic congestions are included. The first case study is presented in Subsection 3.1, while the second in Subsection 3.2. The results of the simulation experiments are discussed in Subsection 3.3. In Section 4, the sensitive transport assignment referred to in Subsection 1.1 is presented and discussed. In Section 5, conclusions are drawn and further research is suggested.

2 Traffic Simulation Based Experiments

2.1 The Simulation Environment

As it was mentioned earlier, Vissim simulator was chosen and used for the simulation experiments reported in the present study. The traffic models created in Vissim normally rely on its own generic traffic flow models. These characterize and describe vehicle movements both in longitudinal and lateral directions, moreover the model may include roads with different lane structures. The simulator has a number of traffic conflict resolution models, which can simulate and manage cross-roads, junctions and pedestrian crossings, as well as road locations with road layout changes. All these are road locations where traffic conflicts are likely to arise. Relying on Vissim's built in capabilities, one can build realistic road network models, and can obtain realistic simulation results. Still, it has its inherent limitations; and these need to be addressed when complex traffic schemes, unusual traffic aspects, out of the ordinary/vehicle functions, or new communication techniques/services are to be modelled and implemented.

2.2 The Simulation Approach

Currently, the generic versions of the Vissim simulator framework do not provide built in LoS filtering support for their users; or in other words, the visibility between vehicles/road users cannot be checked within a customary simulation, and so the visibility relation between vehicles/road users can not be used as a condition for an action, or decision. Such relations, however, could be approximated through and utilized in a Vissim-MATLAB cooperation. Albeit such an implementation would result in a more faithful modelling and simulation of the traffic that involves also vehicles with OTCD capabilities, it is left as a target for future work. This decision was taken with a view on the more involved and longer effort needed for Vissim-COM-based developments reported by both [23] [24]:

- A possible introductory move forward could involve semi-automatic generation and evaluation of LoS regions of the queues within a road subnetwork. The move could start with sampling the subnetwork either manually, or algorithmically, so that only a few hundred discrete road locations remain that still characterize the subnetwork properly.

The difference between the physical and the sampled road location-based visibility is illustrated via a simple example shown in Figure 1. Two vehicles wait in a queue that has formed because of a car crash near an urban road crossing. The figure shows that the last car in the queue is not in the visual LoS region of the purple car that approaches to the road crossing as a building blocks its view.

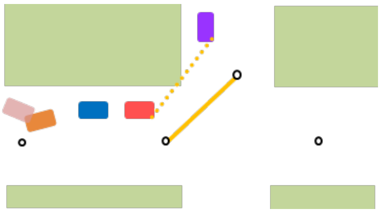


Figure 1

An example for which the sampled road location-based visibility and the vehicle based physical visibility do not coincide. The lack of the LoS connection is indicated by a dotted line, while the solid line indicates that the two sampled road locations are in LoS connection.

Though the physical visibility – in real road traffic – and the sampled road location-based visibility – in a restricted microscopic traffic simulation – provides information on whether a vehicle is present, or not, at the other, i.e., viewed, road location, it is not possible to deduce from this fact if there is a queue there, or not.

It was mentioned in Subsection 1.4 that the queue detection method implemented by the OTCD system is not addressed in the present manuscript, but its traffic simulation aspects should be touched upon:

- In a dedicated lane-based microscopic traffic modelling/simulation, some characteristic traffic descriptor (e.g., recent average vehicle speed in the lane) should be collected and calculated for the close vicinity of each discrete road location that has been sampled and chosen for assessing the visibility.
- In Vissim-based simulations, queue counters – associated with each of the aforementioned discrete road locations – should be introduced and used for the purpose of detecting queues from a distance.

In the tentative approach used herein, however, the intricate spatio-temporal process of traffic congestion detection and avoidance relying on individual OTCD capabilities is modelled and simulated in a much-simplified manner by using three proxy parameters. The first of these serve as a defining parameter of sub-scenarios, while the second and third represent specific time-delays.

The first proxy parameter – denoted by γ – falls in range of 0.0 to 1.0; and for convenience, is given in percentages. It is used for the purpose of quantifying the mitigation of the traffic hindrance that has been caused by some disturbance; more precisely, it specifies the percentage of the road vehicles equipped with and relying on their own OTCD systems among the vehicles. In the simulation experiments, these percentages represent the proportions of vehicles diverted from the blocked paths to other compatible paths. Proxy parameter γ also serves as a defining parameter of the sub-scenarios; γ was assigned different discrete values for these, namely 0%, 5%, …, 25%, and 30% (i.e., step of 5%)[2].

The second proxy parameter represents the time required for driving around the obstacle in the traffic, it is denoted by T_{stop} as it was implemented through vehicles' stopping at the obstacle; while the third proxy parameter represents the time required for taking a rerouting decision, and it is denoted by $T_{decision}$. Proxy parameter T_{stop} can be linked to the obstacle size relative to the road-width/lane-width, and also to the seriousness of the traffic incident, as well as to the traffic intensity; while $T_{decision}$ represents the time required for a driver/AV who/that has reached the vicinity of the obstacle to find that:

a) A vehicle queue has built up there, due to some unusual reason

b) This queue will not dissolve quickly

c) To establish – through his/her local knowledge, or by use of a navigation device – that one or more alternative route is available

d) The time required for the planning of the modified route, should also be included.

Choosing these time delays properly is a precondition of achieving a realistic estimation of the LoS based traffic congestion avoidance and mitigation process. If fixed time-delays are used in the simulation, then these delays can be estimated, e.g., by falling back onto some very simple traffic model. Another possibility would be to make use of real-life time measurements taken in different traffic conditions in the area close to the intended obstacle location. Alternatively, the intersections near the intended obstacle location could be looked at via focused microscopic traffic simulations to come up with realistic estimates for the delays. Plausible ad hoc choice is a further possibility. T_{stop} and $T_{decision}$ are used in the first case study as fixed time-delays; they, on the other hand, are varied in the second.

2.3 The Road Subnetwork Used in the Simulation Experiments

The subnetwork chosen for the simulation experiments is a subnetwork of roads in Budapest, Hungary. It is situated on the west bank of the River Danube. The Móricz Zsigmond Circus (MZSC) is in the center of the target subnetwork, as it is shown in Figure 2. The travel demands used in the simulations are artificial, but realistic. Each demand is associated with an ordered pair of numbered locations, i.e., with the origin (O) and destination (D) locations, within the

[2] Increasing γ beyond a limit, in this case beyond 30%, did not improve the traffic situation any further, and for this reason, the corresponding sub-scenarios are not includeed herein.

subnetwork. These O-D pairs are referred to as routes in the text, and are denoted as in Route 2→1.

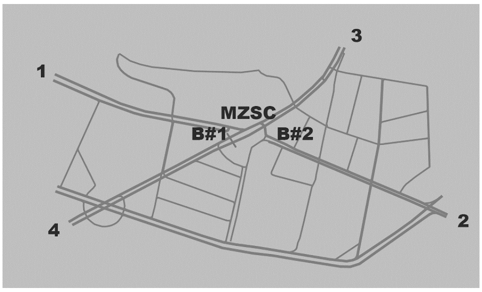


Figure 2
The modelled subnetwork of roads

In the map, four numbered locations are indicated; each of these generates and resorbs traffic to/from the other numbered locations, respectively. The lengths of shortest paths for Routes 2→1, 2→4, 3→4 are 2.4 km, 2.5 km and 1.5 km, respectively.

3 Traffic Simulation Case Studies

3.1 The First Case Study

The First Scenario (FS) of the First Case Study is looked at in Sub-subsection 3.1.1, the Second (Base) Scenario (SBS) in Sub-subsection 3.1.2, the PSSs of Second Scenario in Sub-subsection 3.1.3, and the Enroute Dynamic Rerouting (EDR) for Second Scenario in Subsubsection 3.1.4.

3.1.1 First Scenario: The Undisturbed Traffic

The FS specifies an undisturbed traffic – within the subnetwork shown in Figure 2 – without any hindrance:

- The subnetwork moves a total of 3600 vehicles between its four numbered road locations in an hour.
- A simulation run models a one-hour interval. Within this interval, vehicles are 'generated' only in the first 45 minutes, in the remaining 15 minutes the subnetwork is left on its own to discharge.
- Each of the four numbered location 'generates' equal number of vehicles, i.e., 300 vehicles, heading for each of the other three numbered road

locations within the simulated period, i.e., total travel demand D_{total} is 3600 vehicles/hour.

- Major junctions within the subnetwork are modelled as signalized junctions.

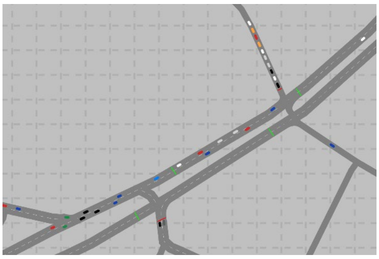


Figure 3

A snapshot of the undisturbed traffic according to the FS within a rectangular area at and near MZSC. The road vehicles appear as miniscule elongated color blobs.

The traffic that forms under the above conditions is not very intense, and is undisturbed, as can be perceived from the sub-image of a simulation snapshot shown in Figure 3. In Table 1, the average travel times (ATTs) on certain routes and corresponding numbers of arrived vehicles (NAVs) are shown. No recent traffic information is communicated to and used by the drivers/AVs modelled in this simulation run. The routes included in Table 1 – and also in Table 2 – are the ones that are to be hit hardest by the obstacle popping up according to SBS.

Table 1

The average travel times (ATTs) and the numbers of arrived vehicles (NAVs) per hour for the FS on two specific routes

Route	**ATT [s]**	**NAV**
2 → 4	427	300
3 → 4	160	300

Table 2

The ATTs and the NAVs per hour for the FS on two specific routes; however, in this case the drivers/AVs could make use of recent traffic information communicated to them

Route	**ATT [s]**	**NAV**
2 → 4	545	267
3 → 4	150	300

In Table 2, the ATTs and NAVs per hour are shown for the aforementioned routes. These traffic descriptors are also for the FS, however, in this case the drivers/AVs – modelled in the traffic simulation – had access to recent traffic information, which had been received by their NRGSs. Comparing the corresponding traffic descriptors presented in Tables 1 and 2, one can see that the utility of the recent traffic information – in case of undisturbed and calm traffic – is not distinct.

3.1.2 Second Base Scenario: An Obstacle Greatly Hinders the Traffic

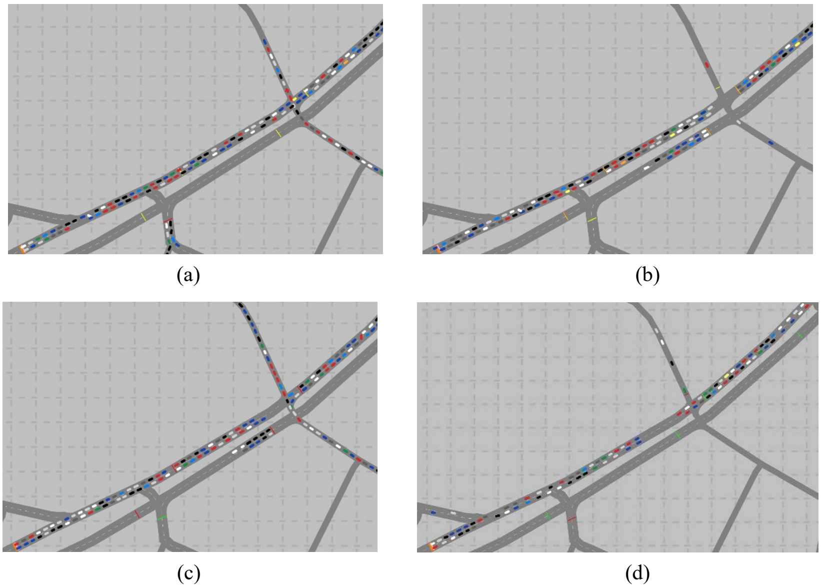
In the SBS, an obstacle pops up (e.g., bus breaks down) close to MZSC on the road that leads from MZSC to Location 4:

- The obstacle is modelled to stop each vehicle – passing it in the mentioned direction – for 60 s. In other words, T_{stop} is fixed, and is set to 60 s for the simulations implementing the SBS and the PSSs.

The full-blown effect of the obstacle's appearance in the road subnetwork on the traffic situation can be perceived by viewing Figure 4a. A continuous vehicle queue has formed on the road that leads from Location 3 to Location 4 – in this direction – in the road section shown in the figure.

3.1.3 Parametrized Sub-scenarios of the Second Scenario: Easing the Traffic Congestion via Vehicle Diversions to Compatible Alternative Paths

The second scenario has been split up into a number of PSSs based on proxy parameter γ.



(a) (b) (c) (d)

Figure 4

The full-blown effect of the obstacle's appearance on the traffic (a). The traffic jam is then eased via diversions of 10% (b), 20% (c) and 30% (d) of the vehicles concerned.

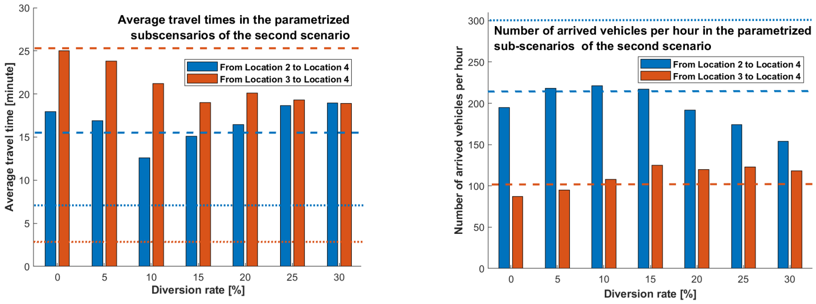
- The congestion is mitigated through diversions of growing proportions to alternative paths of the respective routes; seven discrete values of γ were used in separate simulation runs. In these, 0%, 5%, 10%, … , 25% and 30% (i.e., step of 5%) of vehicles were diverted with a sixty-second delay from the blocked path to alternative compatible paths, i.e., $T_{decision}$ is fixed, and set to 60s.

Simulation snapshots corresponding to four out of the seven aforementioned simulation runs have been included herein. The snapshots for PSSs corresponding to 0%, 10%, 20% and 30% (i.e., step of 10%) vehicle diversions[3] are shown in the subfigures of Figure 4.

Diagrams presented in Figure 5 show the ATTs and NAVs per hour, respectively, for vehicles travelling along Routes 2→4 and 3→4 according to the PSSs. The diagrams indicate that certain diversions, namely the 10% and the 15% ones, have perceptively eased the traffic situation on these "hard hit" routes, respectively.

3.1.4 Easing the Traffic Congestion of the Second Base Scenario via EDR

In the simulation experiments reported in this subsubsection, the effect of traffic congestion mitigation via EDR is looked at. Unlike the more detailed study presented in [27], herein the percentage of the vehicles equipped with NRGSs has not been varied in fine grades, rather either all the drivers/AVs made use of NRGSs, or none of them did.



(a) (b)

Figure 5

The effect of EDR – making use of recent road traffic information – and of the PPSs on ATTs (a) and on NAVs per hour (b) for Routes 2→4 and 3→4 for the second scenario. Dashed lines indicate ATTs and NAVs, respectively, for the second scenario with EDR, while the dotted lines indicate these traffic descriptors for the undisturbed traffic, i.e., for FS.

[3] Note that the case of γ = 0%, is actually the SBS.

The benefit of receiving recent traffic information about the road network – via NRGSs – is assessed herein based on two specific traffic descriptors, namely a) ATTs, and b) NAVs per hour, calculated for two selected routes, namely for Route 2→4 (blue[4]) and Route 3→4 (red), of the subnetwork:

- The aforementioned traffic descriptors computed for the undisturbed traffic – without and with EDR, see Tables 1 and 2, respectively – have already been looked at, and compared. Based on this comparison, one can see that EDR – relying on recent traffic information received via NRGS – is not particularly useful in case of undisturbed traffic, as the traffic flows smoothly anyway.
- On the other hand, in the congested traffic that builds up according to the SBS, EDR– making use of recent traffic information – may considerably improve the traffic situation[5]. This improvement reveals itself also in the traffic descriptors mentioned above; these are shown diagrammatically in the respective subfigures of Figure 5.

3.2 The Second Case Study

In the present subsection, the simulation experiments presented in [10] – in respect of using individual OTCD capabilities for traffic congestion detection and alleviation – are reiterated. More concretely, all the parameters defined earlier, i.e., γ, T_{stop}, and $T_{decision}$, will be varied. Moreover, in some of the experiments, also parameter D_{total}, which specifies the total travel demand in the target subnetwork, will be varied. T_{stop}, $T_{decision}$ and D_{total} are varied around their respective fixed values considered in the first case study. By varying these, the aim was to better understand their roles in the model, to delimit the spatio-temporal process of congestion detection and avoidance based on OTCD capabilities.

The present subsection is structured as follows. Subsubsection 3.2.1 presents the FS, in which the undisturbed road traffic within the target subnetwork is examined; then, in Subsubsection 3.2.2, the SBS and PSSs are examined.

3.2.1 First Scenario with Different Total Travel Demands

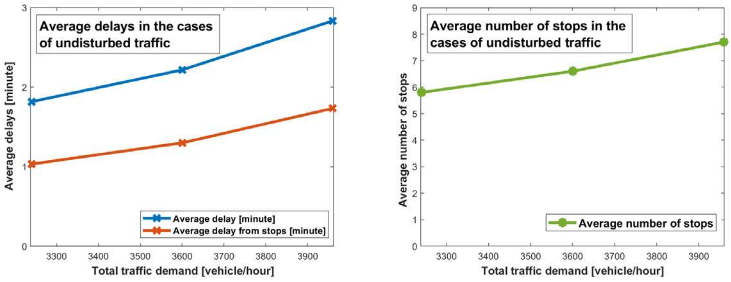
In this subsubsection, the undisturbed road traffic in the subnetwork shown in Figure 2 is looked at under three different total travel demands. More concretely:

[4] The traffic descriptor values for Route 2→4 are shown as blue bars in the subfigures of Figure 5, whereas the descriptor values for Route 3→4 are shown as red bars there.

[5] The improvement achieved through EDR is less though than that achieved by the best of the parametrized sub-scenarios (see Figure 5). Please recall that these sub-scenarios serve as proxies for the modelled OTCD based traffic congestion alleviation approach.

- The traffic that takes shape under 90%, 100% and 110% of D_{total} used in the first case study are looked at. There, D_{total} was chosen to be 3600 vehicles/hour. Each of these total travel demands is divided equally between the twelve considered routes, and the distribution of the travel demands in time – within each simulation run – is the same as it was in the first case study.

Figures 6 and 7 present diagrams of four specific traffic descriptors. In Figure 6a, the average delays (blue) and average delays due to stops (red) experienced by drivers/AVs travelling in the subnetwork are shown. The tendency of these delays matches the intuition: with higher traffic demands these delays grow.



(a) (b)

Figure 6
The average delays and the average delays from stops (a) and the average number of stops (b) – within the undisturbed traffic – for various traffic demands

Figure 6b presents the average number of stops drivers/AVs had to make during their travel through the subnetwork because of the other vehicles and of the traffic signals. The tendency of this traffic descriptor again matches the intuition: with higher traffic demands the descriptor grows.

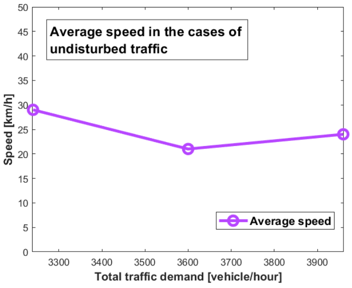


Figure 7
Average vehicle speeds for three different traffic demands served by undisturbed traffic

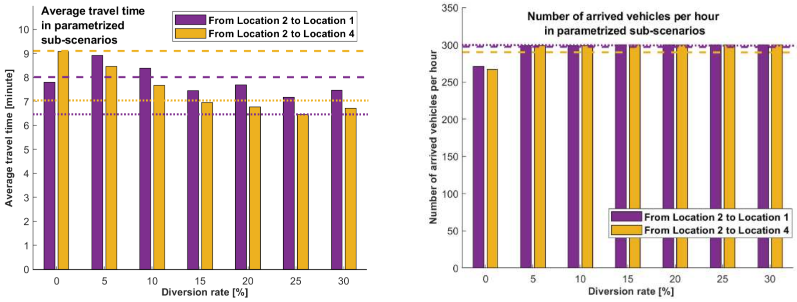
In Figure 7, the average vehicle speed within the target road subnetwork is shown. The diagram does not seem to follow a clear pattern, and does not match the intuition; with higher traffic demands, the average vehicle speed first decreases, then slightly increases.

3.2.2 The Second Base Scenario and its Parametrized Sub-Scenarios

The smooth road traffic is hindered by the unexpected appearance of an obstacle at the road location labelled with B#2 in Figure 2:

- The dotted lines in the subfigures of Figure 8 indicate ATTs and NAVs per hour, respectively, for the FS, i.e., for the undisturbed, smooth traffic.
- The traffic congestion formed because of the obstacle is then mitigated – in consecutive simulation experiments – by increasing γ from 0% to 30% by steps of 5%. The traffic situations that take shape in the SBS and the PSSs are characterized also by the aforementioned diagrams. These show how the ATTs and the NAVs per hour change – as proxy parameter γ increases – for routes hit hard by the obstacle, namely for Routes 2 → 1 and 2 → 4.
- The dashed lines indicate ATTs and NAVs per hour, respectively, for the second scenario with EDR based on recent traffic information.

The diagrams in the subfigures of Figure 8 are to be compared to the diagrams prepared for the first case study, i.e., to Figures 5a and 5b. The diagrams in Figures 5 and 8 have been prepared for the nominal D_{total} only, and shows the ATTs and NAVs per hours of particularly hard-hit routes.



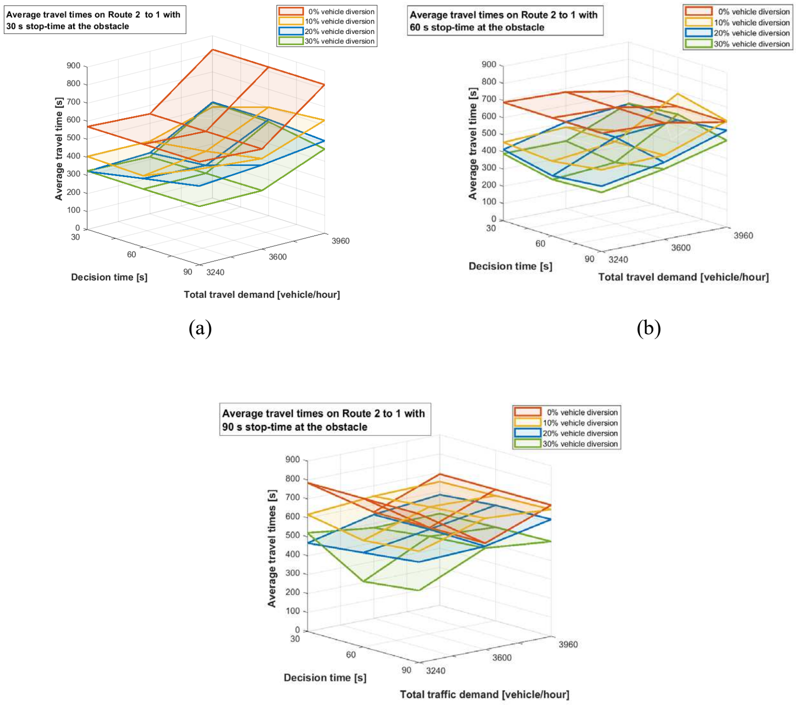
(a) (b)

Figure 8

The traffic situations that take shape in the SBS and the PSSs of the second case study. Also, the effect of EDR – based on recent traffic information – is shown. The ATTs (a) and the NAVs per hour (b) for Routes 2→1 and 2→4 for the nominal total traffic demand.

To illustrate how – according to the PSSs of the second scenario within the second case study – ATTs along a particularly hard-hit route, namely along Route 2 → 1, change as γ, T_{stop}, $T_{decision}$, and D_{total} are varied in consecutive simulation experiments, three separate diagrams are included here as subfigures of Figure 9.

In the mentioned diagrams, proxy parameter T_{stop} was set to 30s, 60s, and 90s, respectively. The interpolated ATT-surfaces corresponding to the following γ values have been included therein: 0%, 10%, 20% and 30% (i.e., step of 10%).



(a) (b)

(c)

Figure 9

The ATTs on Route 2→1 – for various decision times, traffic demands, diversion ratios – in the traffic disturbed by the obstacle. The stop-times at the obstacle are 30s (a), 60s (b) and 90s (c), respectively.

3.3 Discussion of the Traffic Simulation Results

According to the microscopic traffic simulations carried out in conjunction with the first case study, the following traffic conditions and changes can be observed for the FS, the SBS, and the PSSs based on Figures 3, 4 and 5:

- Figure 3 shows calm, fairly light, undisturbed traffic that has formed according to the FS.
- In Figure 4a, on the other hand, the full-blown effect of the obstacle, which appeared in the target road network at a busy road location B#1, on the traffic can be perceived: a continuous vehicle queue has built up on a hard-hit road, and the traffic situation has worsened considerably. This situation can be gauged by considering that the ATT on the hardest hit route, i.e., Route 3→4 has grown about 9 times of its FS value, while the NAV per hour has reduced to about 25% of its FS value.

- In case of the PSSs corresponding to γ values of 5% and 10%, the ATTs and NAVs per hour, see Figure 5, have improved compared to the respective values computed for the SBS in respect to the Routes 2→4 and 3→4. A screenshot of the traffic according to the latter PSS is shown in Figure 4b. These improvements are due to the fact that some the drivers/ AVs started using low-capacity local roads to reach their destinations. These diversions have resulted in somewhat reduced road traffic near the obstacle location B#1, and have decreased the length of the queue that has originally formed there.
- For the PSSs corresponding to γ values of 15% and 20%, the ATTs on both considered routes started to increase, i.e., worsen, while the NAVs per hour continued to increase, i.e., improve, according to Figure 5. This is due to the higher number of vehicles using now the local roads; however, using local roads results in longer travel times, thereby increasing also the considered ATTs. On the other hand, the queue has become considerably shorter near the obstacle location B#1, as it can be verified in Figure 4c for the latter PSS.
- For the PSSs of 25% and 30% diversions, the ATTs, as well as the NAVs per hour considered have worsened according to Figure 5. The explanation for this is that even though the queue near the obstacle location has become considerably shorter than for the preceding PSSs, the local roads are now getting congested causing more delays and causing vehicles to be stuck there. A screenshot of traffic according to the latter PSS is shown in Figure 4d.
- In Figure 5, the ATTs and NAVs per hour computed for the different PSSs can be compared to those computed for the FS, and to those computed for the hindered case when recent traffic information was used by the drivers/AVs for EDR.

According to the microscopic traffic simulations carried out in conjunction with the second case study, the following traffic conditions and changes can be observed for the FS, the SBS, and the PSSs based on Figures 6-9:

- In Subsubsection 3.2.1, the FSs[6] of the second case study with different total travel demands were looked at. In this context, Figures 6 and 7 present diagrams of four specific traffic indicators – computed for the undisturbed traffic – versus D_{total}. The descriptors presented in this manner were the average delay, average delay due to stops, average number of stops and average vehicle speed. The first three of these four behaved as expected, the fourth, however, did not show a clear pattern.

[6] Note that the FSs are shared by the two case studies, as only the hindered cases are different for the two case studies.

- In Subsubsection 3.2.2, the traffic situations that take shape in the SBS and the PSSs of the second case study were looked at. In this context, Figures 8 present diagrams of ATTs and NAVs for two hard-hit routes at the nominal total traffic demand.
- Also, in Subsubsection 3.2.2, the SBS and PSSs of the second case study were looked while several parameters were varied as:
  - In the subfigures of Figure 9, the ATTs on a particularly hard-hit route have been presented for various values of T_{stop}. The ATTs behave in regards to this proxy parameter as expected, for longer stop-times the ATTs tend to grow.
  - In each subfigure of Figure 9, the interpolated ATT-surfaces –drawn in red, yellow, blue and green – tend to lower as γ increases from 0% to 30% by steps of 10%. The ATTs behave in regards to this proxy parameter more or less as expected.
  - In regards to D_{total}, the ATTs tend to grow, though this behavior of theirs is not that clear cut.
  - In regards to proxy parameter $T_{decision}$, the ATTs do not follow a clear pattern. $T_{decision}$ could be related to urban texture[7] near and around the obstacle and the queues.

The hindrances caused by the obstacles associated with the two case studies presented herein can be gauged by comparing Figures 5 and 8. Based on this comparison, one finds that the SBS according to the first case study caused considerably greater hindrance to the traffic than the SBS of the second case study. This comparison, as well as the analysis of further traffic descriptors that not have not been included herein, signifies a higher utility of the OTCD capability in the first case study than in the second.

4 Traffic Situations Not Relieved by Local Traffic Congestion Mitigation Approaches

In the previous section, two nonrecurrent traffic congestion cases were looked at in two separate case studies. The cases considered therein can be mitigated by local traffic congestion mitigation approaches, including also the OTCD based approach described in Subsection 1.2. On the other hand, according to Conjecture

[7] In a lightly built-in area, the LoS detection of the traffic congestion could be quicker and more reliable than in a densely built area. Therefore, when modelling the visibility from the nearby crossroads, the lightly built-in area could be associated with a shorter $T_{decision}$, while the densely built area with a longer $T_{decision}$,..

3 put forward in Subsection 1.1, there are traffic situations and transport tasks/assignments that do not lend themselves to local traffic mitigation approaches. An example[8] from Budapest, Hungary for such a situation is given below:

- The transport task is to get/drive to the Budapest Airport (BUD) from the city center in a short time by car in order to reach either an outgoing flight departing from BUD, or to meet a passenger arriving to the airport onboard an incoming flight.

Of course, one can choose the fast, but accident-prone, narrow expressway, along which one can – in undisturbed traffic – drive relatively fast, but in the peak hours, and in case of some even minor traffic incident on the road, the traffic comes to a stand-still for a while; or one can choose some slower route that has more alternatives, moreover, the route itself and all its alternative routes taken into consideration bypass the expressway toward the airport.

Based on the simulation experiments presented in Section 3, and on the above example from Budapest, one can realize the importance of available alternative routes both, in traffic congestion mitigation and in routing. Then, when analyzing and tackling these and similar adverse traffic situations, one could think in terms of groups of routes rather than in terms of individual routes[9]. For instance, sticking to Budapest, if someone wants to drive from a location on the Pest side, i.e., from the eastern half of the city, to a location on the Buda side, i.e., to the western half, they can choose between different Danube bridges to reach the intended destination. In this example, the bridges could well be used to identify the route groups.

In certain road networks, depending on the urban texture of and the traffic control and management arrangements implemented, a route can be slightly modified via taking a short local detour within its own route group when the driver/AV encounters a minor disturbance along their planned route. In this sense, there are route groups that are robust against minor traffic disturbances, and there are groups, which are not.

Based on the route group concept, a new, or at least uncommon, route optimization criterion can be specified. Furthermore, this approach can and should be combined with the LoS problem/task/approach used herein. Referring back to the second motto, with the above outlined routing approach, which considers the existence, and number of local route alternatives, the queues arising because of

[8] Note that this example underlines also the relevance of the third motto. A further point to the mentioned motto: the handbook cited therein was published before the proliferation of intelligent, and/or connected vehicles, and AVs; since then, and due to the spread of such vehicles on roads, the number of essential road design parameters should be increased from the stated two to three. See further details in [30].

[9] The notion of route is used now in its original sense, not as O-D pairs.

some unexpected minor incident at some neuralgic road location could be totally bypassed, and the transport assignment – or mission as referred to in the motto – need not be aborted at all. The outlined route optimization approach could find applications in military, police and security related route planning tasks, but could also in fairly common, everyday tasks.

It is not to say that the above outlined routing approach is always applicable in real-world situations. The mission mentioned in the second motto needs to be aborted during its execution (because of some traffic jams and blocked roads encountered). Relying on the above outlined routing approach, the aforementioned mission might not even start: it may become a mission impossible right at the beginning (due to some real-world time constraint, e.g., driving along the safer route featuring many local alternatives routes might take too long to reach the intended flight). The above outlined approach is simply too caution.

When – after further research, modelling, implementation and validation efforts – the approach outlined above reaches the maturity to deal with real-world applications, it will need to consider:

a) Actual traffic intensity
b) Multilane sections of the roads
c) Different vehicle priorities
d) Dealing with short critical sections of the route
e) Looking at the most critical traffic situations could also facilitate its application in real-world traffic situations and routing tasks.
f) The seriousness of the traffic incident
g) The traffic intensity dependent accident rates
h) Visibility degradation due to adverse weather conditions

Conclusions

In the frame of the study presented herein, traffic simulation experiments were carried out for a concrete urban road subnetwork, to explore the joint effect of road vehicles' individual traffic congestion avoidance efforts, in which, onboard visual LoS, ECSs and related OTCD capabilities were put to use. Two realistic case studies were carried out to investigate the effect of non-recurrent, unexpected traffic incidents. The presented Vissim-based simulation experiments concerned themselves with the comparison of undisturbed, disturbed and mitigated traffic.

In addition, the traffic effects of the hindrances described in the case studies were compared. The process of congestion detection, avoidance and mitigation was tentatively modelled via proxy parameters in the manuscript. In certain simulation experiments, several proxy parameters were varied and the corresponding results assessed.

In Section 4, a new objective was proposed for route planning and optimization in sensitive traffic situations and in case of sensitive transport assignments. Further research aspects were mentioned in the text.

Acknowledgements

This research was supported by the European Union within the framework of the National Laboratory for Autonomous Systems (grant: RRF-2.3.1-21-2022-00002). It was also supported by the NRDI Office, Budapest, Hungary (grant: 2018-2.1.10-TÉTMC-2018-00009).

References

[1] Payalan, Y. F.; Guvensan, M. A. Towards next-generation vehicles featuring vehicle intelligence. IEEE Trans. Intell. Transp. Syst. 2019, 21, 30-47

[2] Ortiz, F. M.; Sammarco, M; Costa L. M. M.; Detyniecki, M. Appli-cations and services using vehicular exteroceptive sensors: a survey. IEEE Trans. Intell. Veh., 2022, 1-20

[3] Marti, E.; de Miguel, M. A.; Garcia F.; Perez, J. A review of sensor technologies for perception in automated driving, IEEE Intell. Transp. Syst. Mag. 2019, 11(4), 94-108

[4] Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Gläser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst. 2021, 22(3), 1341-1360

[5] He, J.; Tang, Z.; Fu, X.; Leng, S.; Wu, F.; Huang, K.; Huang, J.; Zhang, J.; Zhang, Y.; Radford, A.; Li, L. Cooperative Connected Autonomous Vehicles (CAVs): Research, Applications and Challenges. In Proceedings of IEEE 27th International Conference on Network Protocols, Chicago, IL, USA, 2019

[6] Marron, A.; Limonad, L.; Pollack, S.; Harel, D. Expecting the Unexpected: Developing Autonomous System Design Principles for Reacting to Unpredicted Events and Conditions. In: Proceedings of IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, pp. 167-173, 2020, Seoul, South Korea

[7] Wolhuter, K. Geometric Design of Roads Handbook. CRC Press, Boca Raton, FL, USA, 2015

[8] World Health Organization (WHO): European regional status report on road safety 2019, Available online: https://apps.who.int/iris/handle/10665/336584 (accessed on 18 Oct, 2022)

[9] Jayasooriya, S. A. C. S.; Bandara, Y. M. M. S. Measuring the Economic Costs of Traffic Congestion. In Proceedings of 2017 Moratuwa Engineering Research Conference, pp. 141-146, Moratuwa, Sri Lanka, 2017

[10] Fazekas, Z.; Obaid, M.; Boulmakoul, A.; Karim, L.; Gaspár, P. Utility assessment of line-of-sight traffic jam and queue detection in urban environments for intelligent road vehicles, in: Boulmakoul, A., Karim, L., Bhushan, B. (Eds.), Intelligent Distributed Computing for Trajectories. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, 2022, pp. 207-228

[11] Thandavarayan, G.; Sepulcre, M.; Gozalvez, J. Cooperative perception for connected and automated vehicles: evaluation and impact of congestion control. IEEE Access 2020, 8, 197665-683

[12] Lu, Q.; Tettamanti, T.; Hörcher, D.; Varga, I. The impact of autonomous vehicles on urban traffic network capacity: an experimental analysis by microscopic traffic simulation. Transp. Lett. 2020, 12, 540-549

[13] Bohra, S. Design Manual Adjustments and Infrastructure Needs for Automated Vehicles: City of Toronto context. MSc thesis, Ryerson University, Toronto, Ontario, Canada, 2019

[14] PTV VISSIM. Available online: https://www.ptvgroup.com/en/solutions/products/ptv-vissim/ (accessed on 22 Dec, 2022)

[15] PTV Planung Transport Verkehr AG., VISSIM 4.10 User Manual, 2005, Karlsruhe, Germany

[16] Parisot, C.; Meessen, J.; Carincotte, C.; Desurmont, X. Real-time Road Traffic Classification Using On-board Bus Video Camera. In: Proceedings of 11th International IEEE Conference on Intelligent Transportation Systems, pp. 189-196, Beijing, China, 2008

[17] Chakraborty, P.; Adu-Gyamfi, Y. O.; Poddar, S.; Ahsani, V.; Sharma, A.; Sarkar, S. Traffic congestion detection from camera images using deep convolution neural networks. Transp. Res. Rec. 2018, 2672 (45), 222-231

[18] Akhtar, M.; Moridpour, S. A review of traffic congestion prediction using artificial intelligence. J. Adv. Transp. 2021, article-id. 8878011

[19] Zanella, A.; Bazzi, A.; Pasolini, G.; Masini, B.M. On the impact of routing strategies on the interference of ad hoc wireless networks. IEEE Trans Commun 2013, 61(10), 4322-4333

[20] Bazzi, A.; Berthet, A. O.; Campolo, C.; Masini, B. M.; Molinaro, A.; Zanella A. On the design of sidelink for cellular V2X: a literature review and outlook for future. IEEE Access 2021, 8, 97953-97980

[21] Hu, L.; Ou, J.; Huang, J.; Chen, Y.; Cao, D. A review of research on traffic conflicts based on intelligent vehicles. IEEE Access, 2020, 8, 24471-483

[22] Ullah, M. R.; Khattak, K. S.; Khan, Z. H.; Khan, M. A.; Minallah, N.; Khan, A. N. Vehicular traffic simulation software: A systematic comparative analysis. Pakistan J. Eng. Appl. Sci. 2021, 4(1), 66-78

[23] Ramadhan, S. A.; Joelianto, E.; Sutarto, H. Y. Simulation of traffic control using Vissim-COM interface. Internetworking Indones. J. 2019, 11 (1), 55-61

[24] Bansal, P. Matlab-Vissim interface for online optimization of green time splits. arXiv preprint: arXiv:2007.15208, 2020

[25] Valentine, D. T.; Hahn, B. Essential MATLAB for Engineers and Scientists. Academic Press, London, United Kingdom, 2022

[26] Chen, M. Modeling and Simulation Analysis of Road Network Based on VISSIM. In: 2019 International Conference on Intelligent Transportation, Big Data and Smart City, Changsha, China, 2019, pp. 32-35. IEEE

[27] Farhan, M.; Martin, P. T. Evaluation of the benefits of route guidance system using combined traffic assignment and control framework. In: Proceedings of 17th ITS World Congress, ITS Japan, ITS America, Busan, South Korea, 2010

[28] Savrasovs, M.; Pticina, I.; Zemlyanikin, V. Wide-Scale Transport Network Microscopic Simulation Using Dynamic Assignment Approach. In: Proceedings of the International Conference on Reliability and Statistics in Transportation and Communication, Riga, Latvia, 2017, pp. 241-251

[29] Permana, A. Y.; Susanti, I.; Dewi, N. I. K.; Wijaya, K. Morphology of urban space: model of configuration using logic of space theory in densely populated of Bandung City. Journal of Architectural Research and Education 2019, 1(1), 18-35

[30] Fazekas, Z.; Balázs, G.; Gyulai, C.; Potyondi, P.; Gáspár, P. Road-type detection based on traffic sign and lane data. J. Adv. Transp. 2022, 6766455

Avoiding Mistakes in Bivariate Linear Regression and Correlation Analysis, in Rigorous Research

László Barna Iantovics

George Emil Palade University of Medicine, Pharmacy, Science and Technology of Targu Mures, Gheorghe Marinescu, 38, 540142, Targu Mures, Romania, barna.iantovics@umfst.ro; ORCID iD: https://orcid.org/0000-0001-6254-9291

Abstract: Data science and artificial intelligence are emergently, very fast-evolving fields, being applied to a large diversity of real-life problem-solving. In this context, some methods are applied without verifying assumptions that must be met, for the correct applicability and the necessary model fit. Such mistakes could lead to misinterpretations of the results. One of the application domains, that is very affected in this sense, is healthcare, where misinterpretations could have dangerous effects on human health. Based on an in-depth study of the scientific literature, it was identified that bivariate linear regression (BLR) even is considered simple, is one of the methods that sometimes leads to confusion in application. With this in mind, this paper proposes in an algorithmic form of a methodology that consists of assumptions, that must be passed by the BLR, so that the applicability is correct and should pass the required threshold model fit. Also, presented in algorithmic form is the decision for the correct calculus of the bivariate linear correlation coefficient (BCC). There are other considerations, like the necessary sample sizes for the two variables in the case of BCC and BLR. The proposed methodology, herein, will be useful for researchers, since BLR is frequently applied in research in diverse domains, like industry and healthcare, individually or combined with methods of data science and artificial intelligence.

Keywords: data science; linear regression; model fit; prediction; artificial intelligence; mathematical modeling; goodness-of-fit; mistakes encountered in clinical research; correlation coefficient; data misinterpretation

1 Introduction

Methods of Data Science and Artificial Intelligence including Intelligent Systems are successfully applied for a wide diversity of real-life problem-solving. An optimization method of robotic mobile agent navigation uses a neural network [1]. In [2] a comprehensive review of recent trends in measuring machine intelligence is presented. Even though measuring machine intelligence is of high interest, there are very few methods focused on measuring machine intelligence [3] [4]. Methods based on statistics combined with methods of artificial intelligence and data science are frequently applied together to combine the advantages of the constituting methods. A novel statistical methodology is applied for the detection of cooperative multiagent systems with extreme intelligence (those that are statistically significantly more intelligent than a set of considered intelligent systems) [5].

Research in all domains, industry, and healthcare, comprises problems or subproblems that involve bivariate correlation (BC) and/or regression analysis (RA) which includes the bivariate linear regression (BLR) analysis [6]. For instance, soil water content prediction using regression models could offer support in the decision-making processes [7]. The research [8] focused on the study of the bivariate correlation between health literacy and cell phone addiction among Iranian healthcare students. A frequent mistake in research that involves BC and BLR is that these methods are applied without verifying the necessary assumptions, which should be passed for model fit and correctness of their applicability even in papers published in the best journals. This could lead to erroneous interpretation of the research results. One of the most affected domains in this sense is healthcare, where misinterpretations of BC and BLR could lead even to loss of lives [9]. Based on this motivation, this paper proposes a mathematically grounded modeling of the assumptions that must be passed by BLR to be applicable and to pass the requested threshold model fit. At the same time, it presents the algorithmic decision for the correct calculus of BCC.

The upcoming structure of the paper is organized as follows: Section 2 presents a brief survey of the state-of-the-art research that is based on BC and BLR; Section 3 presents the proposed methodology; and Section 4 presents the experimental evaluation of the proposal. Finally, in Section 5, the conclusions of the paper are presented.

2 State-of-the-art Applications of BC and BLR

BC and BLR have applications in many real-life problems solving. Prediction could be helpful in human decision-making. Frequently prediction methods are based on bivariate linear regression. Prediction based on BLR is approached in various research, such as rice sheath blight field resistance prediction [10], reduced energy consumption prediction [11], and many others.

Anti-social behavior identification in online discussions frequently is important to be identified. For this problem solving a classification method that involves diverse regression methods is proposed [12].

Concerns of decision-makers toward the profit obtained by using cloud computing technology are studied in [13]. The proposed method includes linear regression.

A study on the suitability of predictive control on an advanced mechatronic system consisting of a laboratory helicopter is presented in [14]. The developed model includes a linear regression method.

Sometimes linear regression is combined with diverse methods of artificial intelligence or data science for solving prediction problems like traffic with climate condition prediction [15], city-wide demand-side prediction [16], and many others.

Various regression methods are applied for diverse real-life problem-solving. Another widely applied regression method is logistic regression, applied for problem-solving like identifying candidate disease genes [17], sampling on-demand [18], semantic web service matchmaking [19], etc.

Obtaining real data in industry and healthcare for research frequently is difficult. A novel method for data quality assessment of synthetic data obtained by simulation is based on a statistical approach [20]. It is treated the bivariate logistic regression that is frequently applied mostly in healthcare-related research. There are proposed mathematically grounded assumptions that must be met for the application of bivariate logistic regression to be correct and pass the requested threshold model fit.

The presented bibliographic study shows that even though BC and BLR are traditional methods, they still have even recent times many applications.

3 Assumptions Testing in BC and BLR

3.1 Mistakes in Bivariate Correlation and Regression Analyses

Even simple descriptive statistics and data normality analyses presented in research have some common mistakes [21]. A decision rule for data central tendency indicator establishment in [22] is presented. In [23] is treated the subject of avoiding mistakes in quantitative statistical analyses in political science. A valuable step-by-step guide for the correct application of different statistical tests is presented in [24].

There are some common statistical errors that appear in papers published in radiology journals [25]. The study involved 157 selected papers from 20 radiology journals that were published between 2016 and 2017. The selected articles were assessed regarding different kinds of statistical errors like mistakes in statistical tests applied, wrong interpretation of p values, and some others. The mistakes were treated considering the journals based on their impact factors.

According to [26], the most common mistakes in some medical research consist of wrong sample size determination; mistakes in bias related to sampling; mistakes in making adjustments in multiple comparisons; wrong interpretation of the p-value by considering clinical relevance, choosing wrong statistical tests, etc.

A guide for avoiding some mistakes in scientific investigations related to epidemiology, and public health in [27] is presented. There are suggested questions that must be responded to before the effective beginning of a research.

In many papers, there are reported mistakes that could appear in the statistical regression analyses [28-31]. Statistical analyses performed on animal science present some common mistakes that are presented in [32]. Studies on confounding bias for heritability include different mistakes that are treated in [33]. A study on the biases in summary statistics of slopes and intercepts in linear regression that includes errors in both independent and dependent variables is presented in [34].

As a general conclusion based on the performed bibliographic study can be formulated that the usual mistake in the case of statistical methods based on linear regression is that they are applied without being based on necessary assumptions that should be met for their applicability to be correct, and the model fit to be passed the required threshold. This laziness frequently leads to misinterpretations. Linear correlation and regression analysis are among these methods.

3.2 The Proposed Validation and Analysis Methodology

In the following, are presented the assumptions that must be passed by BLR and the model fit analyses. Also, the correct calculus of BCC is treated. The presented mathematically grounded assumptions are applicable in any type of research that involves BC and BLR.

The BLR problem includes two variables measured at continuous levels (interval or ratio variables) denoted in the following *VrX* and *VrY*. $|VrX|$ and $|VrY|$ denote the sample sizes of *VrX* and *VrY*. $|VrX|=|VrY|=n$, $Df=n$-2, where *Df* represents the degrees of freedom. The *BivRegMet* algorithm presents the methodology for verification of the assumptions that should be met for BLR could be applicable and passing the necessary model fit thresholds. Previously to *BivRegMet* could be applied the *BivCorr* algorithm that describes the correct calculus of the correlation coefficient of *VrX* and *VrY*. PCc denotes the Pearson correlation coefficient. SCc denotes the Spearman correlation coefficient. In case *BivCorr* is applied, then the *BivRegMet* algorithm application can be decided on the response to the fact that it is a valuable parametric or nonparametric statistics, being applicable only in parametric case (both *VrX* and *VrY* passed the normality assumption). However, *BivCorr* if applied is a prefiltering assumption and model fit verification for the BLR application.

It must be noticed that in many types of research, the human evaluator (*HE*) should have a central role in the interpretation of the experimental evaluation results and the establishment of the values of different parameters. According to the two algorithms, *HE* is responsible for the establishment of the parameter's values (*CL*, α_{anova}, etc.), thresholds for model fit (necessary for the interpretation; the threshold for the strength of the correlation for instance), and supervises the evaluation process, by deciding in different decision points when necessary. Must be motivated that *HE* must make some visual examinations of graphical results. *HE* in his/her contribution will consider the specificity of the research, application area, and his/her background knowledge.

BivCorr: Bivariate Correlation Coefficient Calculus algorithm

```
IN: VrX; VrY; OUT: r; //correlation coefficient.
SignIndic; CorrStr;// significant correlation is detected?; correlation strength
Direc; // correlation direction: positive or negative?;
Step A1. Calculus of the correlation coefficient.
CL:=95%; //the implicit confidence level
Step A1.1. Verification of the normality assumptions.
@Verification of VrX and VrY normality using numeric goodness-of-fit tests;
@VrX, VrY normality results validation by visual examination of the Q-Q plots;
Step A1.2. Calculus of the correlation coefficient.
If (VrX and VrY are normally distributed) Then @Calculates PCc r. //parametric
    Else @Calculates SCc r. //nonparametric. $\rho$ is more usual notation.
EndIf
If ($r$ = 0) Then @ no correlation, the analysis is stopped. EndIf
Step A2. Assessment of correlation significance and correlation direction
@Establish the Research Hypotheses:
    $H_{cr}$: "r is statistically equal with 0" //null hypothesis, no correlation
    $H_{cra}$: "r is statistically significantly different from 0" //alternative hypothesis
@Calculates the CI, [Lr, Ur] of r at the CL confidence level.
If  (($r > 0$) and ($Lr > 0$))  Then //significant positive correlation.
  $H_{cr}$ rejected, $H_{cra}$ accepted; SignIndic := "Yes"; Direc := "+";
  ElseIf  (($r < 0$) and ($Ur < 0$)) Then //significant negative correlation.
      $H_{cr}$ rejected, $H_{cra}$ accepted; SignIndic := "Yes"; Direc :="-";
        Else $H_{cr}$ accepted. @there is no correlation, the analysis is stopped.
EndIf
Step A3. Establishment of the correlation strength.
```

```
If (SignIndic="Yes") Then @The strength of the correlation CorrStr is
    established based on the |r| value considering the classification from Table 1.
EndIf
EndBivCorr
```

In the case of both algorithms, each of the parameters is set to an implicit value that usually is considered the most appropriate. *HE* could change these values if considered, based on his/her consideration and taking into account other evidence if available (for instance in similar research, is obtained, a certain value of the correlation coefficient and in the actual study, is considered to have a comparable value).

α denotes the Type I error rate. It is recommended to approach the two-tailed α. β denotes the Type II error rate. The power is calculated as 1-β. r_{estim} represents an estimation of the correlation coefficient (the expected correlation coefficient). r_{estim} value can be based on some background knowledge (previous study for instance). Z_α denotes the standard normal deviate for α. Z_β denotes the standard normal deviate for β. It is given, β and an estimate of expectable r_{estim} size then can be calculated the necessary sample size *n* (and degree of freedom *Df*) (1, 2) [35]. For example, α=0.05 and power=0.8 (β=0.2), when correlation coefficients are increasing 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9, the sample sizes are decreasing as follows 193, 84, 46, 29, 19, 13, 9 and 6, respectively.

$$C = 0.5 \times \ln[(1+ r_{estim})/(1- r_{estim})] \quad (1)$$

$$n = [(Z_\alpha+Z_\beta)/C]^2 + 3 \quad (2)$$

Step A1.1, in the case of numerical evaluation of the normality assumption, for small sample sizes (n≤30), as a decision rule, recommended the application of the Shapiro-Wilk (SW) goodness-of-fit test [36], which is proven [37] as having the highest statistical power compared with the frequently used: Kolmogorov-Smirnov, Lilliefors (Lill), and Anderson-Darling tests. A limitation of the SW test consists of the sensitivity in the case of large samples. In case of larger sample sizes, we recommend the application of the Lill test [38-40] which represents an adaptation of the Kolmogorov-Smirnov test [41].

The Quantile-Quantile (Q-Q) plot [42] [43] is a scatter plot appropriate for the normality visual validation. A Q-Q plot is a drawn reference line. The visual study of the data normally involves the verification if the data points fall almost along this reference line. The larger the difference from the reference line, the larger the evidence is for the interpretation that the data fails to pass the normality assumption. Additionally, to the mentioned numerical verification of normality, it is recommended to make a visual validation based on the drawn Q-Q plot.

r denotes the correlation coefficient of *VrX* and *VrY*. According to Step A1.2 of the algorithm, if both *VrX* and *VrY* are normally distributed, the PCc *r* [44] [45] is computed, elsewhere the SCc *r* (more precisely notation is ρ) [44] [45] is computed. This decision, regarding the calculus of the correlation coefficient, is based on the fact that SCc is more robust than PCc (less sensitive to influential points), and considering this circumstance, it is more appropriate in the nonparametric case. As additional validation, *HE* can make a visual evaluation of the Scatter Plot created based on *VrX* and *VrY*. This simple approach is useful to visually present the relationship between the two studied continuous variables; indicate if there are influential points; and estimate if the relationship is linear.

When r=0, this indicates no correlation. Step A2 verifies if the correlation is statistically significant (if r is statistically significantly different from 0) based on the r value and the *CI* of r at the *CL*% level. If the difference is statistically significant, and if r>0 then the correlation is positive, if r<0 the correlation is negative.

Step A3 establishes *CorrStr* as the strength of the correlation according to Table 1 in case a statistically significant correlation is detected in the previous step. *CorrStr* value can be considered a model fit measure. *HE* could require a certain level of correlation strength. For instance, could establish that just a very strong correlation is acceptable.

Table 1

Range of correlations strength

Correlation coefficient value	Interpretation (Level of correlation strength)
$\|r\|\in [0.8, 1]$	Very strong
$\|r\|\in [0.6, 0.8)$	Strong
$\|r\|\in [0.4, 0.6)$	Moderate
$\|r\|\in [0.2, 0.4)$	Week
$\|r\|\in [0, 2)$	Neutral

Additionally, in the case of parametric correlation is recommended the calculus of the r^2, $r^2\in[0,1]$ is called the coefficient of determination [46]. r^2 is an indicator of the effect size. r^2 is the proportion of the variation in the *VrY* variable that is explainable/predictable by the *VrX* variable. For instance, r^2=0.82 indicates that 82% of the variance of the *VrY* variable is explained by the variance of the *VrX* variable. r^2 is a measure of the goodness-of-fit of the model, higher value means better model fit. The minimal required threshold value of *CD* that should pass r^2 must be established by *HE* based on the specificity of the research and considering how good the model fit should be. Frequently *CD*=0.7 can be considered as an implicit parameter value. In case if $r^2 \geq CD$ then the threshold passed, else the threshold does not pass. Table 2 presents the usual interpretation of r^2 values. *HE* will consider also the strength of the correlation jointly with the value of r^2 (in the case of parametric correlation is interpreted r^2).

Table 2
r^2 interpretation in parametric case

r^2	Interpretation
≥0.85	Very good
[0.75, 0.85)	Good
[0.6, 0.75)	Satisfactory
<0.6	Weak

Spearman's *r* (more precisely ρ) is an indicator of monotonicity. It reflects the extent to which an increase/decrease in *VrX* is associated with an increase/decrease in *VrY*, but the expanse of increase/decrease does not have to be constant over the whole range of values, as in the case of linear correlation.

BivRegMet: Assumptions for application of BLR

```
IN: VrX; VrY; Principal OUT: y(x);//regression equation
The visual plotted regression line and its confidence interval (CI);
Secondary OUT: DW// the Durbin-Watson statistics result.
SignSlope;//slope of the regression line is significantly different from 0;
RD;//root mean square deviation (RMSD), called standard deviation of residuals;
Step B1. Preliminary assumptions checking
@Obtain the DW applying the Durbin-Watson test.
@VIS1:The residual plot is created, plotted standardized predicted scores (X axes) against standardized residuals (Y axes). HE visually verifies the homoscedasticity.
@VIS2:Elaborated probability–probability (P-P) plot. HE makes a visual examination of the normality of residuals.
@HE establishes $R_{MSD}$ value. Calculate RD and the value of RMSD.//model fit
Step B2. Model Fit Overall Model test
@Establish the Research Hypotheses:
  $H_r$: The slope (denoted $a$) of the regression equation is statistically equal to 0.
  $H_{ra}$: The slope of the regression equation is significantly different from 0.
@Apply the ANOVA test. Let $Pval_{an}$ be the obtained p-value of the ANOVA test.
If ($Pval_{an} > \alpha_{an}$) Then $H_r$ is proved. SignSlope:=“No”; //NO significant difference.
  Else $H_r$ proving failed. $H_{ra}$ proved. SignSlope:=“Yes”;//significant difference.
EndIf
Step B3. BLR modeling if assumptions passed and model fit
If (($DW \sim 2$) and (VIS1 passed) and (VIS2 passed) and ($RD < R_{MSD}$) and (SignSlope=“Yes”)) Then
    @Constuct the linear regression equation $y(x) = a \times x + b$.
    @Plot the regression line including the CI at the CL level.
    @HE makes a visual validation of the regression line considering also its CI.
    @HE treats influential points if exist.//stepwise methodology described in
    Section 4 in applicative form.
        Else @"weak model-fit".//BLR is not approachable
EndIF
EndBivRegMet
```

Step B1 consists of some preliminary analyses. For the verification of the independence of values in the two variables (autocorrelation in the residuals) is applied the Durbin-Watson statistic obtaining a *DW* value. $DW\in[0,4]$, where $DW=2$ indicates there is no autocorrelation, $DW<2$ indicates positive autocorrelation, and $DW>2$ indicates negative autocorrelation. The residuals must have a constant variance with no dependence on the level of the dependent variable, a property called homoscedasticity (VIS1 verification). After the Durbin-Watson test, *HE* performs the visual examination of homoscedasticity on the residual plot created verifying that the error terms variance is constant crossways the dependent variable values. *HE* makes a visual examination of residuals normality on the P-P plot (VIS2). RMSD is a measure of the goodness-of-fit of the regression line. *RD* represents the calculated RMSD value. *RD* must be interpreted according to Table 3, comparatively with an established threshold value R_{MSD}.

Step B2 responds to the Null Hypothesis *Hr* and Alternative Hypothesis *Hra*, for whose verification is applied the ANOVA test. $Pval_{an}$ denotes the obtained p-value of the ANOVA test applied at the α_{an} significance level. α_{an} could have different values like 0.01, 0.001, etc. In most cases is recommended the α_{an} value 0.05. In Step B2 performed the regression analysis, where *HE* will decide on the application of BLR, based on the requirements DW~2; VIS1 visual examination passing; VIS2 visual examination passing; $RD<R_{MSD}$; and *SignSlope*="Yes".

Table 3

Root Mean Square Deviation Interpretation

RMSD value	Interpretation
≤ 0.75	Very good
(0.75, 1]	Good
(1, 2]	Satisfactory
>2	Not satisfactory

Finally, the regression equation is constructed and the regression line is plotted with the CI at the CL level being appropriate for visual validation of model fit. The CI helps in the visual and numerical appreciation of points that fall outside the CI. *HE* will use this information in making the final decision on the effective application of bivariate linear regression for the considered problem-solving. Influential points are important to be detected in regression analysis since they could have a large impact on the linear regression equation [47] [48]. Section 4 additionally presents applications of the step-by-step treating of influential points that could largely affect the regression equation.

4 Experimental Evaluation

In this section, for testing and evaluation of the proposed methodology it was performed an experimental evaluation case study focused on a prediction problem based on BLR. In this case, *VrX* is called independent (predictor) variable and *VrY* is called dependent (predicted) variable. Also, we present some additional elements that could be used by *HE* in the interpretation of the results and formulation of conclusions based on data analysis results. Table 4 presents the generated synthetic data used in the evaluation. The subscripts labeled “inf1”, “inf2” and “inf3” indicate influential points that will be identified and treated later.

Table 4

The data used for experimental evaluation

VrX	***VrY***	***VrX***	***VrY***	***VrX***	***VrY***
93	3.78	64	2.88	71	2.89
$61^{\mathbf{inf1}}$	$3.8^{\mathbf{inf1}}$	$62^{\mathbf{inf2}}$	$1.6^{\mathbf{inf2}}$	$73^{\mathbf{inf3}}$	$2.42^{\mathbf{inf3}}$
74	3.1	85	3.19	87	3.44
69	2.88	94	3.68	91	3.91
70	3.21	78	3.28	56	2
53	2.1	66	3.1		

Initially is applied *BivCorr* algorithm. Step A1.1, considering the low sample size of both variables $|VrX|=|VrY|=17$ (17<30), $Df=15$, the best option is the application of the SW test and making a visual validation of normality based on the Q-Q plot. For illustrative purposes of the interpretation of diverse results that could be obtained by different normality goodness-of-fit tests, the Lill test, along with the SW test was applied. For both tests, it was considered the $\alpha_{norm}=0.05$ significance level. Table 5 presents the results of the Lill and SW tests applied to both variables *VrX* and *VrY*. Both variables met the normality assumption for both tests of normality. *p* denotes the p-value obtained after application of a test, with $p>\alpha_{norm}$, which leads to the acceptance of the null hypothesis of the normality test, (H_0) that states the assumption of normality.

Table 5

The results of the applied normality tests

	Lilliefors			**Shapiro-Wilk**		
Variable	***Statistic***	***P***	$p > \alpha_{norm}$	***Statistic***	***P***	$p> \alpha_{norm}$
VrX	0.127	0.2	Yes	0.949	0.438	Yes
VrY	0.185	0.126	Yes	0.936	0.271	Yes

Figures 1 and 2 present the Q-Q plots for *VrX* and *VrY*. The visual interpretation of both figures leads to the same conclusions as the numerical SW and Lill tests for meeting the normality assumption.

In the following, a descriptive static (Table 6) was performed, even if it is not integrated into the algorithms. This could admit the formulation of some additional remarks. *SD* denotes the standard deviation. *Variance* represents the variance that is calculated as SD^2. For each variable, the mean was chosen as the central tendency indicator since both variables met the normality assumption. *CV* denotes the Coefficient of Variation, $CV=SD/mean\times 100$. The lower the *CV*, the lower the dispersion. *CV* admits the evaluation of data homogeneity-heterogeneity, by making a classification: $CV\in[0, CVa)$ indicates homogeneity; $CV\in[CVa, CVb)$ indicates relative homogeneity; $CV\in[CVb, CVc)$ indicates relative heterogeneity; $CV\geq CVc$ indicates heterogeneity. The most usual recommended values, which should be established by *HE*, are CVa=10, CVb=20, and CVc=30.

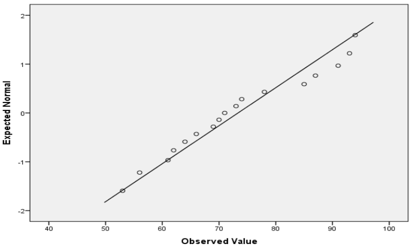


Figure 1

The Q-Q plot of *VrX*

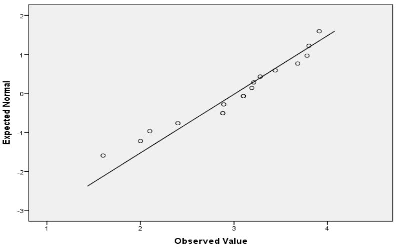


Figure 2

The Q-Q plot of *VrY*

Table 6

Descriptive statistics

Variable	*Mean*	*SD*	*Variance*	*CV*
VrX	73.353	12.85	165.123	17.52
VrY	3.014	0.665	0.442	22.06

Table 7 presents the results of descriptive statistics obtained by performing bootstrapping based on 1000 samples. L_{rd} denotes the lower bound of the 95% CI. L_{ur} denotes the upper bound of the 95% CI. *SE* denotes the Standard Error.

Table 7

Descriptive statistics results by bootstrapping

Variable		*Bias*	*SE*	L_{rd}	L_{ur}
VrX	Mean	-0.025	3.102	67.237	79.529
	SD	-0.628	1.622	8.832	15.095
VrY	Mean	0.005	0.155	2.695	3.311
	SD	-0.035	-0.035	0.408	0.825

According to Step A1.2, since both variables meet the normality assumption it was chosen the calculus of *r* as the PCc, with *r*=0.712 (Table 8). *r*>0 indicates the possibility of the existence of a positive linear correlation. In the column labeled

"*" the CI of *r* is presented at the *CL*=95%. *r*>0 and 0.351>0 indicate that there is a statistically significant positive correlation. In the column labeled "**" the 95% CI of *r* is calculated based on bootstrapping, with Bias=0.002 and *SE*=0.154, *r* proving to be statistically significant even at the 0.01 level. Bootstrapping was applied using 1000 samples. *r*>0 and 0.358>0 led to the formulation of the same conclusion, claiming the existence of a positive correlation.

Table 8

Results of correlation analysis

Pearson *r*	***p-value***	r^2	**95% CI of *r* ***	**95% CI of *r* ****
0.712	0.001	0.51	[0.351, 0.888]	[0.358, 0.935]

Step A3, According to the classification presented in Table 1, *r*=0.712 ($|r|\in[0.6, 0.8)$), indicate a strong linear correlation. The obtained p-value 0.001 indicates that the correlation is significant even at the 0.001 level. Since it is applied parametric statistics it is calculated the coefficient of determination r^2. *HE* set the *CD* value to 0.7. r^2, r^2=0.51, 0.51< *CD* indicates that the passing of the threshold *CD* failed, and even if the linear correlation is strong the model is not appropriate.

Anyway, *HE* decided on the continuation of the analysis also considering the existence of potential influential points. As a second step, based on the parametric statistics the BivRegMet algorithm was applied. It was obtained *DW*=1.73 very close to value 2, indicating slight negative autocorrelation. Figure 3 shows heteroscedasticity which is a violation of an assumption that should be passed (VIS1). Visual examination of Figure 4 shows the violation of the residuals normality assumption, the residuals of the regression do not follow a normal distribution (VIS2). The *RD* value 0.4824, according to Table 3 admits the formulation of very good goodness-of-fit of the regression line. These analysis results indicate the violation of some assumptions that should be passed, based on this fact *HE* could conclude that the model in the actual form is not appropriate, eventually if there are influential points their removal could have a remediating effect. In the following, there were formulated the hypotheses *Hr* and *Hra*. It was applied the ANOVA test, at the significance level α_{an}=0.05, with results in Table 9, with the predicted variable *VrY* and the predictor variable *VrX*. $Pval_{an}<\alpha_{an}$ admits rejection of *Hr* and acceptance of *Hra* according to that the slope of the regression equation is statistically different from 0.

Table 9

Results of the ANOVA test, indicator of the Overall Model Test

Model	**Sum of Squares**	**Df**	**Mean Square**	**F**	$Pval_{an}$	$p>\alpha_{an}$
Regression	3.588	1	3.588	15.422	0.00134	No
Residual	3.49	15	0.233			
Total	7.079	16				

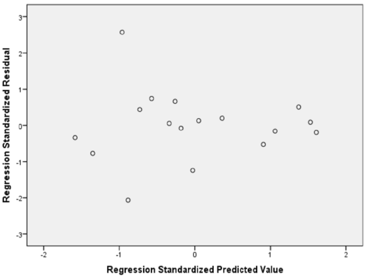


Figure 3
Scatterplot for visual evaluation of Homoscadacity: VaY

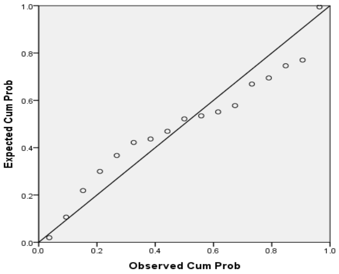


Figure 4
Normal P-P Plot of Regression Standardized residual: Dependent variable VaY

Table 10 presents the coefficients of the regression equation and for additional appreciation the SE, the lower bound (Lb), and the upper bound (Ub) of the 95% CI.

The obtained regression equation (3) is:

$$y(x) = 0.311 + 0.037 \times x \tag{3}$$

Table 10
The coefficients of the regression equation

Model	Best-fit value	SE	Lb	Ub
Slope	0.037	0.009	0.017	0.057
Y Intercept	0.311	0.698	-1.178	1.799
X Intercept	-8.431			

Figure 5 presents the plotted linear regression line, with the 95% CI. The visual interpretation of Figure 5 shows that 29.4% of points fall outside the 95% CI. This leads to the formulation of the remark of indication of weak prediction power.

Based on the previous analysis *HE* should formulate the conclusion that in this case, the model fit is above expectations of *HE* in performing prediction based on bivariate linear regression and another method should be chosen if all the actual data is available.

α is the probability of a Type I Error. β is the probability of a Type II Error. *power* denotes the power. Additionally, it was performed a post-hoc analysis by computing the achieved power, based on considered two-tails, r=0.712, the sample size=17, and established significance level α=0.05 obtaining the *power*=0.931, where β=0.069 (*power=1-β*).

HE based on the visual interpretation of the graphical representation of the regression line and its CI, identifies (X,Y)=(61,3.8) (marked in Table 4 with “inf1”) as a potential influential point that has a large influence on the regression line. HE decided on (61,3.8) removal and the analysis has been repeated. As result obtained: r=0.853 (increased with 0.141) indicates a very strong correlation; r^2=0.728 (increased with 0.218), with the threshold CD (CD=0.7), where r^2>CD, now this assumption is passed (initially with the “inf1” included does not passed); *RD*=0.3532 (decreased with 0.1292), which indicates a slight improvement; the new *DW* value decreased to 1.274 (that is worse than 1.723), indicating degradation (that is conflictual to the other indicators). Based on these facts *HE* decided on the removal of (X,Y)=(61,3.8) and continuation of the exploratory analysis of assumption passing, increase the model fit and the model prediction power.

HE based on the visual evaluation of the graphical regression line and its CI (Figure 5) identifies (X,Y)=(62,1.6) (marked in Table 4 with “inf2”) as a potential influential point. Removing this influential point and repeating the analysis resulted in, an increased r=0.887 (additional increase with 0.034) indicating a very strong correlation; r^2=0.787 (additional increase with 0.059), where $r^2>CD$ (CD=0.7); *RD* value decreased to 0.27 (additional decrease with 0.0832), which indicated a better model fit; the new *DW* value becomes 1.588 which is slightly better than 1.274 but is still worse than the initial 1.723 (with all the influential points included). *HE* decided on (62,1.6) removal and continuation of the exploratory analysis. The obtained regression equation with the two influential points, marked with “inf1” and “inf2” removed is presented in (4), plotted in Figure 6 with the 95% CI.

$$y(x)= 0.152 + 0.0388 \times x. \tag{4}$$

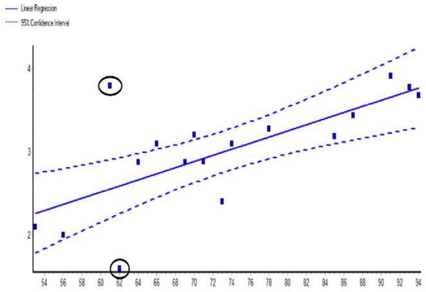


Figure 5

The regression line with the 95% CI

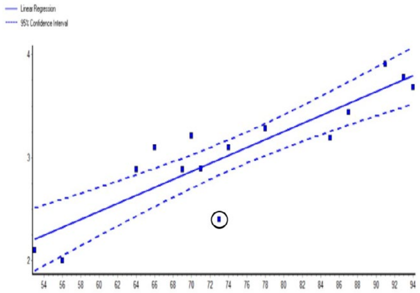


Figure 6

The regression line with the 95% CI with “inf1” and “inf2” influential points removed

HE based on the visual interpretation of the graphical representation regression line and its CI (Figure 6), identifies (X,Y)=(73,2.42) (marked in Table 4 with “inf3”) as a potential influential point removing it for performing an exploratory analysis. The obtained regression equation with the three influential points, “inf1”, “inf2” and “inf3”, removed is (5) plotted in Figure 7 with the 95% CI.

$$y(x)= 0.2324 + 0.03824 \times x \quad (5)$$

Applying BivCorr resulted in, increased r=0.924, p>0.001(additional increase with 0.037), indicating a very strong correlation (95%CI=[0.771,0.976]); increased r^2=0.853 (additional increase with 0.066), where r^2>CD, with the threshold *CD* passed. Applying BivRegMet, the *DW* value of 1.997~2 indicates no autocorrelation. Figure 8 indicates homoscadacity. Figure 9 shows that the assumption of residuals normality has been met. *RD* value decreased to 0.221 (additional decrease with 0.049).

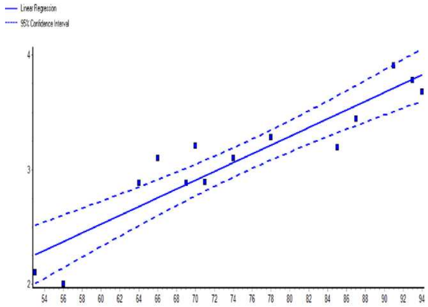


Figure 7

The regression line with the 95% CI with “infl”, “inf2” and “inf3” influential points removed

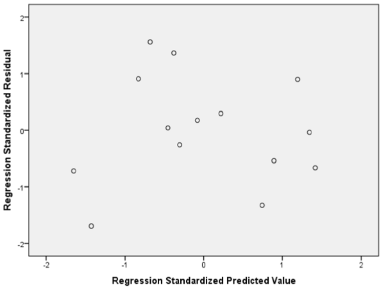


Figure 8

Scatterplot for evaluation of Homoscadacity: VaY with all the influential points removed

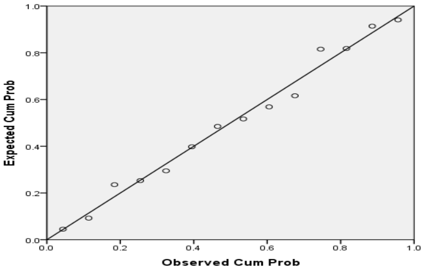


Figure 9

Normal P-P Plot of Regression Standardized residual: VaY with influential points removed

Studying Figure 7 visually, even though a few points fall outside the CI they do not fall too far. To do not make mistakes regarding overfitting can be considered that is not appropriate their removal. The application of the methodology proved that the BLR is applicable and the removal of all the influential points resulted in a better model fit, all the necessary assumptions were met, the final model having a better prediction power. It was proven also the important role of *HE* in evaluation even if some assumptions passed the required threshold the visual interpretation revealed influential points whose removal have improving effect.

Must be noticed that even a very strong correlation between two variables *VarX* and *VarY* does not prove causality. It could happen that *VarX* cases *VarY* or vice versa or there is a third factor *VarZ* that gives growth to the variation in both variables. This is also a situation that shows the important role of *HE*.

Conclusions

In the case of many real-life problems, methods based on statistics, which are sometimes combined with methods of artificial intelligence or data science, are frequently applied. In research that includes bivariate linear regression (BLR), assumptions that must be met for the applicability and the necessary model fit frequently are missed or wrongly applied. Because of this, there is no clear response to the question of whether the BLR is appropriate for a certain problem or subproblem solving. Among others, this is a usual situation in healthcare-research, where misinterpreted data analysis results could have dangerous effects.

With this in mind, this work presented, in the form of algorithmic methodology, the proposed assumptions that must be met for the correct application of BLR and the measurement of the strength of model fit, passing the model fit threshold. It presents an experimental testing and evaluation of the proposed methodology. The proposals from this paper, will be a useful source for researchers who would like to measure the strength of linear correlations, between two variables and/or apply BLR individually or combined with methods of artificial intelligence, data science, or other statistical methods for problem-solving to avoid making mistakes. The methodology can be applied to any type of research that involves BCC and BLR.

It must be said that in many types of research, the *HE* should have a central role in the interpretation of the experimental evaluation results, based on its human-specific background knowledge and contextual knowledge concerning the solved problem. For instance, this can be considered the case of prognosis based on time series, when the data collected in case of a phenomenon meets all the assumptions but is not characteristic of the phenomenon the linearity. For example, we consider the time series of a currency (dollar, euro) that could have a linear tendency over some time but the common sense of *HE* could indicate that linear regression-based prognosis is not a good approach. A proof regarding the importance of the *HE* role in our methodology consisted of the visual discovery of the influential point whose further elimination has as a result fulfilled all the necessary (numeric and visual) assumptions for applicability and model fit threshold. HE applied a trial-and-error methodology when testing the influence of the supposed influential points. For the autocorrelation tests the removal of the first two influential points had a negative effect but after the removal of the third influential point the situation was remediated *DW* indicated no autocorrelation. *HE* in the decisions that make should consider the specificity of the research, application area, personal experiences/knowledge and different background knowledge (other studies results).

Acknowledgment

This work is a part of the CHIST-ERA program supported by the Future and Emerging Technologies (FET) program of the European Union through the ERA-NET funding scheme under the grant agreements, title Social Network of Machines (SOON). This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI-UEFISCDI, project number 101/2019, COFUND-CHIST- ERA-SOON, within PNCDI III. We would like to thank you for the Research Center on Artificial Intelligence, Data Science, and Smart Engineering (ARTEMIS).

References

[1] Brassai, S. T., Iantovics, L. B., Enachescu, C., Optimization of Robotic Mobile Agent Navigation, Studies in Informatics and Control, 21(4), 2012, pp. 403-412

[2] Iantovics, L. B., Gligor, A., Niazi, M. A., Biro, A. I., Szilagyi, S. M., Tokody, D.: Review of Recent Trends in Measuring the Computing Systems Intelligence, BRAIN - Broad Research in Artificial Intelligence and Neuroscience, 9(2), 2018, pp. 77-94

[3] Iantovics, L. B., Rotar, C., Niazi, M. A. MetrIntPair-A Novel Accurate Metric for the Comparison of Two Cooperative Multiagent Systems Intelligence Based on Paired Intelligence Measurements, International Journal of Intelligent Systems, 33(3), 2018, pp. 463-486

[4] Iantovics, L. B. Black-Box-Based Mathematical Modelling of Machine Intelligence Measuring, Mathematics, 9(6), 2021, 681

[5] Arik, S., Iantovics, L. B., Szilagyi, S. M. OutIntSys - a Novel Method for the Detection of the Most Intelligent Cooperative Multiagent Systems, 24th Int. Conf. on Neural Information Processing (ICONIP 2017), 14-18 Nov. 2017, Guangzhou, China. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (Eds.), Neural Information Processing, Lecture Notes in Computer Science 10637, 2017, pp. 31-40

[6] Chen, S. H., Zhou, X. Q., Zhou, G., Fan, C. L., Ding, P. X., Chen, Q. L. An online physical-based multiple linear regression model for building's hourly cooling load prediction. Energy and buildings, 254, 2022, 111574

[7] Leij, F. J., Dane, J. H., Sciortino, A. Hierarchical prediction of soil water content time series. Catena, 209(2), 2022, 105841

[8] Soltani, E., Rezaei, M., Nasiri, M., Barasteh, S., Rahmati-Najarkolaei, F., Mazaheri, MA. The Bivariate Correlation of Health Literacy and Cell Phone Addiction amongst Iranian Healthcare Students, Journal of Clinical and Diagnostic Research, 13(6), 2019, IC1-IC5

[9] Indrayan, A. Statistical fallacies & errors can also jeopardize life & health of many, Indian J Med Res. 148(6), 2018, 677-679

[10] Zeng, Y. X., Dong, J. J., Ji, Z. J., Yang, C. D., Liang, Y. Linear Regression Model for the Prediction of Rice Sheath Blight Field Resistance. Plant Disease, 105(10), 2021, pp. 2964-2969

[11] Biswas, N. K., Banerjee, S., Biswas, U., Ghosh, U. An approach towards development of new linear regression prediction model for reduced energy consumption and SLA violation in the domain of green cloud computing. Sustainable Energy Technologies and Assessments, 45, 2021, 101087

[12] Machova, K., Mach, M., Hreskova, M., Classification of Special Web Reviewers Based on Various Regression Methods, Acta Polytechnica Hungarica, 17(3), 2020, pp. 229-248

[13] Atobishi, T., Bahna, M., Takacs-Gyorgy, K., Fogarassy, C., Factors Affecting the Decision of Adoption Cloud Computing Technology: The Case of Jordanian Business Organizations, Acta Polytechnica Hungarica, 18(5), 2021, pp. 131-154

[14] Jadlovska, A., Jajcisin, S., Predictive Control Algorithms Verification on the Laboratory Helicopter Model, Acta Polytechnica Hungarica, 9(4), 2012, pp. 221-245

[15] Artin, J., Valizadeh, A., Ahmadi, M., Kumar, S. A. P., Sharifi, A. Presentation of a Novel Method for Prediction of Traffic with Climate Condition Based on Ensemble Learning of Neural Architecture Search (NAS) and Linear Regression. Complexity, 2021, 8500572

[16] Kim, T., Sharda, S., Zhou, X. S., Pendyala, R. M., A stepwise interpretable machine learning framework using linear regression (LR) and long short-term memory (LSTM): City-wide demand-side prediction of yellow taxi and for-hire vehicle (FHV) service. Transportation Research Part C: Emerging Technologies, 120, 2020, 102786

[17] Lei, X. J., Zhang, W. X. Logistic regression algorithm to identify candidate disease genes based on reliable protein-protein interaction network. Science China Information Sciences, 64(7), 2021, 179101

[18] Liang, J. Y., Song, Y. S., Li, D. Y., Wang, Z. Q., Dang, C. Y., An accelerator for the logistic regression algorithm based on sampling on-demand. Science China Information Sciences, 63(6), 2020, 169102

[19] Wei, D. P., Wang, T., Wang, J. A logistic regression model for Semantic Web service matchmaking. Science China Information Sciences, 55(7), 2012, pp. 1715-1720

[20] Iantovics, L. B., Enăchescu, C., Method for Data Quality Assessment of Synthetic Industrial Data, Sensors 22(4), 2022, 1608

[21] Madadizadeh, F., Ezati Asar, M., Hosseini, M., Common Statistical Mistakes in Descriptive Statistics Reports of Normal and Non-Normal Variables in Biomedical Sciences Research, Iranian Journal of Public Health, 44 (11), 2015, pp. 1557-1558

[22] Iantovics, L. B., Dehmer, M., Emmert-Streib, F. MetrIntSimil-An Accurate and Robust Metric for Comparison of Similarity in Intelligence of Any Number of Cooperative Multiagent Systems, Symmetry, 10(2), 2018, 48

[23] King, G. How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science, American Journal of Political Science, 30(3), 1986, pp.666-687

[24] Marusteri, M., Bacarea, V. Comparing groups for statistical differences: How to choose the right statistical test?, Biochemia Medica, 20(1), 2010, 15-32

[25] Karadeniz, P. G., Uzabacı, E., Kuyuk, S. A., Kesin, F. K., Can, F. E., Seçil, M., Ercan, İ. Statistical errors in articles published in radiology journals. Diagn Interv Radiol 25(2), 2019, 102-108

[26] Raheem, Y. A. Statistics in medical research: Common mistakes, J Taibah Univ Med Sci. 18(6), 2023, 1197-1199

[27] Rovetta A. Common Statistical Errors in Scientific Investigations: A Simple Guide to Avoid Unfounded Decisions. Cureus. 15(1),2023, e33351

[28] Ignatiadis, N., Saha, S., Sun, D. L., Muralidharan, O., Empirical Bayes Mean Estimation With Nonparametric Errors Via Order Statistic Regression on Replicated Data, Journal of the American Statistical Association, 118(542), 2021, pp. 987-999

[29] Ranganai, E., Nadarajah, S. A predictive leverage statistic for quantile regression with measurement errors, Communications in Statistics - Simulation and Computation, 46(8), 2017, pp. 6385-6398

[30] Abuzaid, A. H., Hussin, A. G., Mohamed, I. B., Detection of outliers in simple circular regression models using the mean circular error statistic, Journal of Statistical Computation and Simulation, 83(2), 2013, pp. 269-277

[31] Chang, X. F., Yang, H. Performance of the preliminary test two-parameter estimators based on the conflicting test statistics in a regression model with Student's t error, Statistics, 46(3), 2012, pp. 291-303

[32] Serao, N. V., Tokach, M. D., Paton, N. Fundamentals, Common Mistakes, and Graduate Education in Statistics, Journal of Animal Science, 99, 2021, pp. 104-104

[33] Holmes, J. B., Speed, D., Balding, D. J., Summary statistic analyses can mistake confounding bias for heritability, Genetic epidemiology, 43(8), 2019, pp. 930-940

[34] Kalantar, A., Gelb, R. I., Alper, J. S. Biases in summary statistics of slopes and intercepts in linear regression with errors in both variables. Talanta, 42(4), 1995, pp. 597-603

[35] Hulley S. B., Cummings S. R., Browner W. S., Grady D., Newman T. B. Designing clinical research: an epidemiologic approach. 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2013. Appendix 6C

[36] Shapiro, S. S., Wilk, M. B. An analysis of variance test for normality (complete samples). Biometrika, 52, 1965, pp. 591-611

[37] Razali, N., Wah, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modelling and Analytics, 2, 2011, pp. 21-33

[38] Dallal, G. E., Wilkinson, L. An analytic approximation to the distribution of Lilliefors's test statistic for normality. American Statistician, 40, 1986, pp. 294-296

[39] Lilliefors, H. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 62, 1967, pp. 399-402

[40] Lilliefors, H. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. J. Am. Stat. Assoc. 64, 1969, pp. 387-389

[41] Chakravarti, I. M., Laha, R. G., Roy, J. Handbook of Methods of Applied Statistics; Wiley: New York, NY, USA, 1967, Vol. I, pp. 392-394

[42] Tsai, D. M., Yang, C. H., A quantile-quantile plot based pattern matching for defect detection, Pattern Recognition Letters, 26(13), 2005, pp.1948-1962

[43] Ben, M. G., Yohai, V. J., Quantile-quantile plot for deviance residuals in the generalized linear model, Pattern Recognition Letters, 13(1), 2004, pp. 36-47

[44] Bonett, D. G., Wright, T. A. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika, 65, 2000, pp. 23-28

[45] Stigler, S. M. Francis Galton's Account of the Invention of Correlation. Statistical Science, 4(2), 1989, pp. 73-79

[46] Chicco, D., Warrens, M. J., Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7(e623), 2021, e623

[47] Meloun M., Hill, M., Militký, J., Vrbíková, J, Stanická, S, Skrha, J. New methodology of influential point detection in regression model building for the prediction of metabolic clearance rate of glucose. Clin Chem Lab Med. 42(3), 2004, 311-322

[48] Sauerbrei, W., Buchholz, A., Boulesteix, A. L., Binder, H. On stability issues in deriving multivariable regression models. Biom J. 57(4), 2015, 531-555

Fusion of Finger Vein Images, at Score Level, for Personal Authentication

Bharathi Subramaniam[1], Sudha V Krishnan[1], Sudhakar Radhakrishnan[1] and Valentina E Balas[2]

[1]Department of Electronics and Communication Engineering, Dr.Mahalingam College of Engineering and Technology, Pollachi -642 003, Tamilnadu, India
E-mail: sbharathi@drmcet.ac.in, sudhashree@drmcet.ac.in, hod_ece@drmcet.ac.in

[2]Department of Automatics and Applied Software, Aurel Vlaicu University of Arad / Academy of Romanian Scientists, 77 B-dul Revolutiei, 310130 Arad, Romania; E-mail: valentina.balas@uav.ro

Abstract: A biometric system with a single biometric trait is less effective, owing to constraints, such as inter-class similarities, susceptibility to noisy pictures and spoofing. Integrating information from different biometric evidences aids in the resolution of difficulties in unimodal biometric systems. It is incredibly challenging in a biometric system to intrude into more than one trait at the same time. Researchers are becoming more interested in multimodal biometric systems due to benefits such as dependability, security, and robustness. A multimodal biometric system based on finger vein images is proposed in this paper, by combining information from the index, middle and ring fingers of the hand. The essential characteristics from the finger vein images are extracted using a Convolutional Neural Network with a ReLU activation function. The input test image features are then compared with the features stored in the database using the correlation-based matching technique, and the match scores are fused using the arithmetic mean-based score level fusion. The performance of the proposed work is analyzed using the finger vein images from STUMULA -HMT database. The results reveal that the suggested multimodal biometric system outperformed the existing techniques, with a maximum accuracy of 99.83%.

Keywords: Biometrics; finger vein authentication; score level fusion; Convolutional Neural Network

1 Introduction

Biometric authentication technology has majorly included the inherent characteristics of people like fingerprint, face, finger vein, iris, voice, gait and many more. In the beginning, fingerprints were paid more attention, and researchers have deployed fingerprint technologies in security applications and handheld devices. The possibility of changing the fingerprint through surgery [1], extracting latent print and spoofing attacks lead to distracting the researchers to another biometric trait. Similarly, face biometric also getting slows down by the possibility of sophisticated face masks [2] and the impact of accuracy because of the non-permanent facial makeup [3]. Nowadays, vein biometrics [4] gain more concentration among researchers because of its own security, reliability, liveness and hygiene. Even though finger vein biometrics provides security, still accuracy is deduced due to intra-class variation, noisy data, inter-class similarities and poor-quality images (Blurry, askew, dim and bright images [5]). Integrating several biometric modalities improves biometric data security and accuracy. The multimodal biometric system may be implemented in a variety of ways by combining biometric data at various levels. The two major possibilities for implementing fusion techniques in multimodal biometric system are at pre-matching and post-matching stages.

Usually, the biometric system learns user-specific identities in the enrollment phase and the identity is verified in the verification phase. Sensor level fusion is the process of amalgamating information obtained from multiple perspectives or different sensors. Feature level fusion integrates features from many sensors (several samples of the same trait, multiple characteristics) to create a concatenated resultant feature vector. Some feature reduction methods may be used to exemplify a larger facet of a fused feature vector. However, the fusion at this level also serves to improve system performance and is used for template creation [6]. The matching module in each biometric modality will deliver the matching score, after matching the input biometric template with the templates stored in the database. Score level fusion is frequently used in multimodal biometric systems [7] [8] because matching scores include enough information to discriminate between genuine and imposter situations while being fairly simple to acquire. The fusion of decisions derived from several modalities is referred to as decision level fusion. However, because the feature sets of the various modalities may be incompatible, fusion at this level is a tricky process to do in practice. Due to the limited availability of relevant data, fusion at the decision level might be seen as stringent.

As a result, the recommended technique utilizes score level fusion to improve the accuracy of the biometric system by merging the matching scores obtained from the index, middle, and ring finger vein images. The remainder of the paper is structured as follows. In Section II the related works identified with fusion of biometric traits for multimodal biometric recognition at various levels are examined. The proposed technique for Fusion of Finger Vein Images at Score Level is introduced in Section III. In Section IV the experimental results obtained by this proposed technique have been explained. The conclusion and future work are discussed in Section V.

2 Background of the Research Work

Many researchers presented their findings on the merging of biometric characteristics at several stages to enrich the biometric system's performance. Yan et al [9] used a feature level fusion approach to generate user-specific biometric templates by fusing data from numerous palm vein samples acquired from each and every individual. The palm vein is captured at a wavelength of 700 nm and they have obtained the EER is around 0.98%. Jinfeng yang et al. [10] presented feature level fusion-based personal identification by combining fingerprint and finger vein data with a unique supervised local-preserving canonical correlation analysis technique (SLPCCAM) and other feature fusion approaches. Among all their fusion approaches, the SLPCCA-based method has a low FAR of around 1.35%. Yong-Fang Yao et al [11] demonstrated a distance-based separable weighting method for fusing of face and palm print data at feature level with an average recognition rate of around 90.73%.

Jialiang Peng et al [12] developed a virtual multimodal biometric system by extracting GLBP, minutiae, Fourier descriptor, and phase congruency features from four biometric modalities: finger vein, fingerprint, finger shape, and finger knuckle print. Sugeno–Weber (SW) triangular norm is used to fuse the match scores. Maleika Heenaye et al. [13] demonstrated a hand vein based multimodal biometric system by utilizing dorsal and palmer vein biometrics of the hand. To normalize the matching scores into the range 0 to 1, min-max normalization techniques were employed. The sum rule, which creates the sum of the product of weights, is then used to fuse the matching scores together and normalized scores in each matching module. The fused biometric system has FRR of around 0.35%. Walia et al [14] presented fingerprint, finger vein, and iris-based multi-biometric systems, and matching was accomplished using Eigen distance, Euclidian distance, and hamming distance for each biometric system. Backtracking Search Optimization Algorithm (BSA) is used to optimize the classifier scores, and the match scores were transformed into belief masses using the Denoeux model. In the decision module, the belief mass is then compared to a threshold value, and they have obtained an EER of 1.57%. Sim et al. [15] suggested an iris and face biometrics-based multimodal system. Euclidian distance and hamming distance were utilized for matching the features of iris and face biometrics. Following that, the collected scores were fused together using weighted score level fusion. Dwivedi et al. [16] used the Rectangle Area Weighting (RAW) approach to show score level fusion of cancellable multi-biometric templates. Individual biometric system scores were computed using the mean-closure weighting (MCW) approach. The two-level cancelable fusion score approach outperforms unimodal cancelable systems. Khellat-Kihel et al. [17] presented finger vein, finger knuckle, and finger print biometrics based multimodal biometric system. Concatenating feature vectors from all three finger-based biometrics was used to achieve feature-level fusion. Then, at the decision level, unimodal judgments from a fingerprint, finger knuckle, and finger vein biometric system are merged to provide the final decision, with an accuracy of about 95.28%.

Kun Su et al [18] employed a multi-sample fusion approach to connect a finger vein and an ECG biometric system. The biometric system was built using feature-level and score-level fusion methods. The maximum EER of all methods was obtained by weighted sum rule-based score level fusion, which was about 1.27%. However, the serial and parallel feature fusion methods yielded an EER of around 7.5%. Ammour et al. [19] presented a multimodal biometric system by utilizing the face and iris. The fusion took place at the score and decision levels, with a recognition rate of around 86.66% for the min-max based normalization technique. Sengar et al [20] developed a palm print and fingerprint-based multi-biometric system based on Deep Neural Network that achieves an accuracy of around 91% with 1.10% FAR and 3% FRR.

3 Proposed Method

Figure 1 depicts the schematic of the proposed finger vein-based biometric system. Convolutional Neural Network [21-23] is utilized during the enrollment phase to extract deep feature sets from index, middle, and ring finger vein images from both the right and left hands. The acquired characteristics are saved in the database as the user's distinct identity. In the verification phase, matching takes place between the input image features and features in the database (stored enrolled templates) utilizing the correlation-based matching technique. Using arithmetic mean-based score fusion, the matching scores obtained from the matching modules are merged into a single number. If the final matching score is greater than the threshold value, the corresponding individual will be given access; otherwise, the person will be considered an imposter.

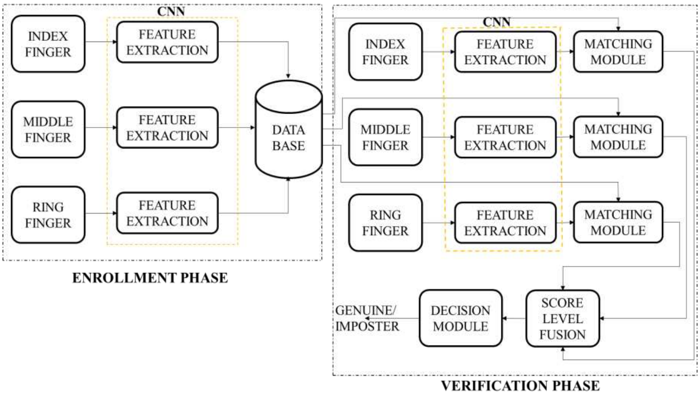


Figure 1
Block diagram of the proposed work

3.1 Database

The proposed research is carried out by utilizing the SDUMLA-HMT database [24], which comprises index, middle, and ring finger vein images taken from the hands of 106 people. Each finger was captured six times, and each image in the database is of size 320×240.

3.2 Feature Extraction

The convolutional neural network is a deep learning algorithm that extracts the feature sets from an image in the forward propagation and updates the weights and bias during backward propagation. After several iterations, CNN learns to extract the relevant features. The configuration of the proposed convolutional neural network is shown in Table 1. The input image is preprocessed and resized to 227×227. We have used six convolutional layers with different number of filters and filter sizes to extract relevant features. Filters convolve with the input image to produce feature maps. The decision of each neuron in the CNN is carried out by the activation functions used in the convolutional layers. In the proposed system, the CNN is trained and tested with ReLU activation function due to its own merits. In every convolutional layer, average pooling is done to reduce the size of the feature map. The architecture of the CNN is shown in Figure 2. Convolutional neural network was trained and tested with finger vein images from the database with various learning rates such as 0.03, 0.025, 0.01, 0.001, 0.0005, 0.0002 and 0.0001.

Table 1

Configuration of the proposed convolutional neural network

Layers	**No. of Filters**	**Filter size**	**Stride**	**Activation function**	**Size of feature map**
Input	-	-	-	-	227×227
Conv 1	32	3×3	1	ReLu	$227\times 227\times 32$
Avg. pooling 1		2×2	2		$114\times 114\times 32$
Conv 2	64	3×3	1	ReLu	$114\times 114\times 64$
Avg. pooling 2		2×2	2		$57\times 57\times 64$
Conv 3	64	3×3	1	ReLu	$57\times 57\times 64$
Avg. pooling 3		2×2	2		$29\times 29\times 64$
Conv 4	128	3×3	1	ReLu	$29\times 29\times 128$
Avg. pooling 4		2×2	2		$15\times 15\times 128$
Conv 5	256	3×3	1	ReLu	$15\times 15\times 256$
Avg. pooling 5		2×2	2		$8\times 8\times 256$
Conv 6	256	3×3	1	ReLu	$8\times 8\times 256$

Avg. pooling 6		2× 2	2		4× 4× 256
Flattening	-	-	-	-	4096× 1
Fully connected layer	-	-	-	Softmax	50× 1

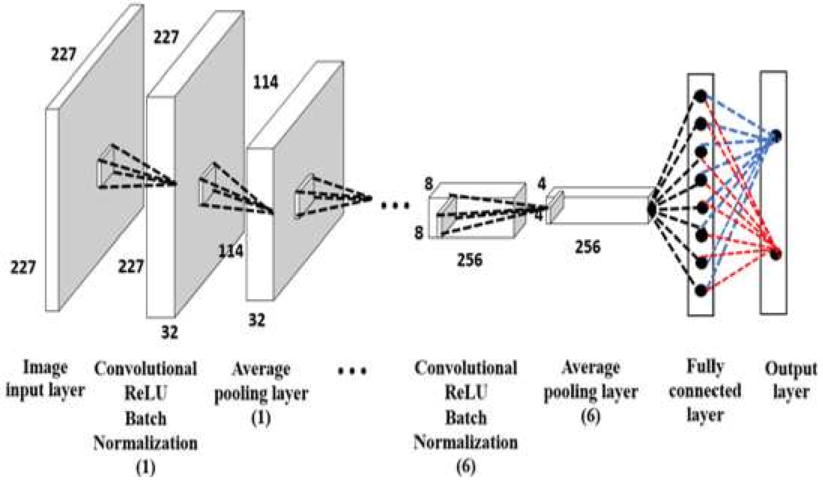

227
227
227
227
114
114
32
32
8
8
256
4
4
256
Image input layer
Convolutional ReLU Batch Normalization (1)
Average pooling layer (1)
• • •
Convolutional ReLU Batch Normalization (6)
Average pooling layer (6)
Fully connected layer
Output layer

Figure 2
Architecture of CNN used for feature extraction

3.3 Enrollment Phase

CNN produces the hierarchical representation of an image as a feature vector after the completion of multiple iterations. The deep layers produce the higher-level features which are constructed using the lower-level features produced in the earlier layers. Then the user-specific feature vectors are stored as their identity.

3.4 Verification Phase

The convolutional neural network is given with images of the user's index, middle, and ring finger veins as input. CNN delivers the relevant feature vector for each finger vein image based on the weights and bias applied during the training phase.

3.5 Matching

The correlation-based matching is performed here in order to compare the input biometric feature template with the enrolled feature templates. The sample values of the enrolled attributes and the test features of the person's index, middle, and ring finger images are given in Table 2. The matching score of enrolled and test features is obtained by finding the correlation between the enrolled and test features with the calculations shown in Tables 3,4 and 5 respectively for the index, middle, and ring finger.

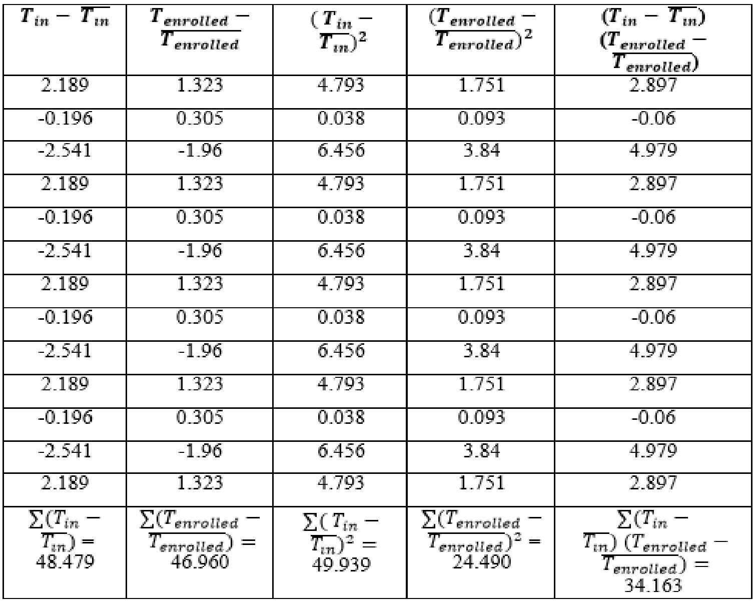
Table 2

Sample values of Train and Test features

Enrolled features			**Test features**		
Index finger	**Middle finger**	**Ring finger**	**Index finger**	**Middle finger**	**Ring finger**
50.668	49.193	50.096	48.283	48.283	49.879
48.283	48.283	48.389	47.265	48.283	48.389
45.938	46.089	46.610	45.000	46.035	46.610
50.668	49.193	50.096	48.283	48.283	49.784
48.283	48.283	48.389	47.265	48.283	48.358
45.938	46.089	46.610	45.000	46.035	46.650
50.668	49.193	50.096	48.283	48.283	50.376
48.283	48.283	48.389	47.265	48.283	48.389
45.938	46.089	46.610	45.000	46.035	46.610
50.668	49.193	50.096	48.283	48.283	50.389
48.283	48.283	48.389	47.265	48.283	48.389
45.938	46.089	46.610	45.000	46.035	46.610
50.668	49.193	50.096	48.283	48.283	49.389

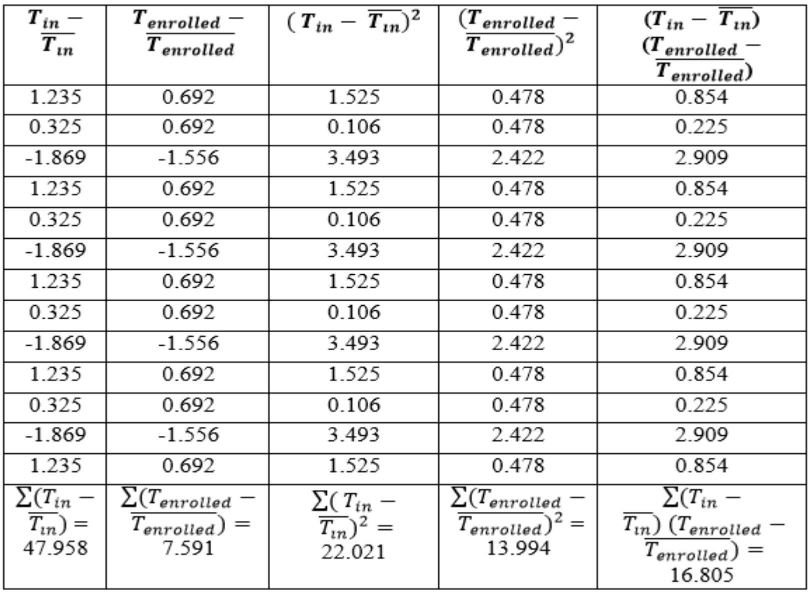
Table 3

Correlation calculation for index finger



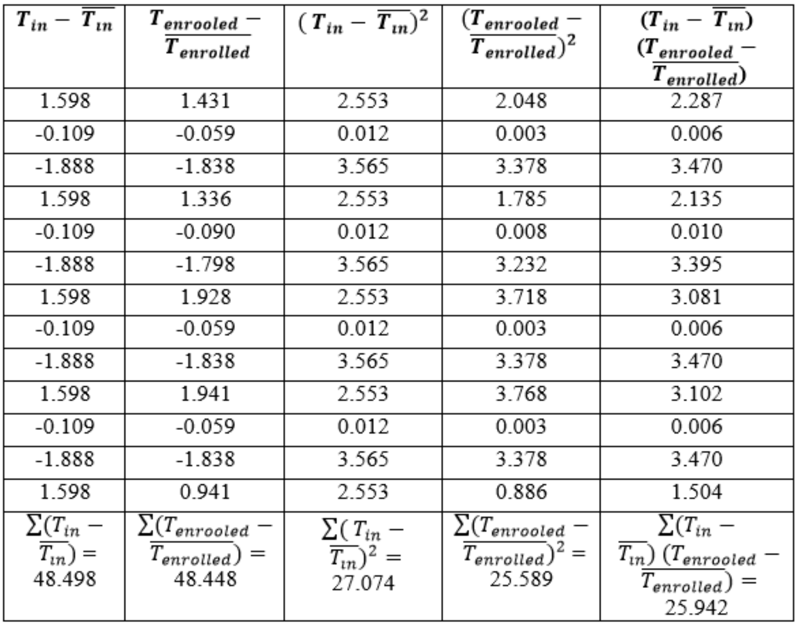
$T_{in} - \overline{T_{in}}$	$T_{enrolled} - \overline{T_{enrolled}}$	$(T_{in} - \overline{T_{in}})^2$	$(T_{enrolled} - \overline{T_{enrolled}})^2$	$(T_{in} - \overline{T_{in}})(T_{enrolled} - \overline{T_{enrolled}})$
2.189	1.323	4.793	1.751	2.897
-0.196	0.305	0.038	0.093	-0.06
-2.541	-1.96	6.456	3.84	4.979
2.189	1.323	4.793	1.751	2.897
-0.196	0.305	0.038	0.093	-0.06
-2.541	-1.96	6.456	3.84	4.979
2.189	1.323	4.793	1.751	2.897
-0.196	0.305	0.038	0.093	-0.06
-2.541	-1.96	6.456	3.84	4.979
2.189	1.323	4.793	1.751	2.897
-0.196	0.305	0.038	0.093	-0.06
-2.541	-1.96	6.456	3.84	4.979
2.189	1.323	4.793	1.751	2.897
$\sum(T_{in} - \overline{T_{in}}) =$ 48.479	$\sum(T_{enrolled} - \overline{T_{enrolled}}) =$ 46.960	$\sum(T_{in} - \overline{T_{in}})^2 =$ 49.939	$\sum(T_{enrolled} - \overline{T_{enrolled}})^2 =$ 24.490	$\sum(T_{in} - \overline{T_{in}})(T_{enrolled} - \overline{T_{enrolled}}) =$ 34.163

Table 4
Correlation calculation for middle finger



$T_{in} - \overline{T_{\imath n}}$	$T_{enrolled} - \overline{T_{enrolled}}$	$(T_{in} - \overline{T_{\imath n}})^2$	$(T_{enrolled} - \overline{T_{enrolled}})^2$	$(T_{in} - \overline{T_{\imath n}})$ $(T_{enrolled} - \overline{T_{enrolled}})$
1.235	0.692	1.525	0.478	0.854
0.325	0.692	0.106	0.478	0.225
-1.869	-1.556	3.493	2.422	2.909
1.235	0.692	1.525	0.478	0.854
0.325	0.692	0.106	0.478	0.225
-1.869	-1.556	3.493	2.422	2.909
1.235	0.692	1.525	0.478	0.854
0.325	0.692	0.106	0.478	0.225
-1.869	-1.556	3.493	2.422	2.909
1.235	0.692	1.525	0.478	0.854
0.325	0.692	0.106	0.478	0.225
-1.869	-1.556	3.493	2.422	2.909
1.235	0.692	1.525	0.478	0.854
$\sum(T_{in} - \overline{T_{\imath n}}) =$ 47.958	$\sum(T_{enrolled} - \overline{T_{enrolled}}) =$ 7.591	$\sum(T_{in} - \overline{T_{\imath n}})^2 =$ 22.021	$\sum(T_{enrolled} - \overline{T_{enrolled}})^2 =$ 13.994	$\sum(T_{in} - \overline{T_{\imath n}})\,(T_{enrolled} - \overline{T_{enrolled}}) =$ 16.805

Table 5
Correlation calculation for ring finger



$T_{in} - \overline{T_{\imath n}}$	$T_{enrooled} - \overline{T_{enrolled}}$	$(T_{in} - \overline{T_{\imath n}})^2$	$(T_{enrooled} - \overline{T_{enrolled}})^2$	$(T_{in} - \overline{T_{\imath n}})$ $(T_{enrooled} - \overline{T_{enrolled}})$
1.598	1.431	2.553	2.048	2.287
-0.109	-0.059	0.012	0.003	0.006
-1.888	-1.838	3.565	3.378	3.470
1.598	1.336	2.553	1.785	2.135
-0.109	-0.090	0.012	0.008	0.010
-1.888	-1.798	3.565	3.232	3.395
1.598	1.928	2.553	3.718	3.081
-0.109	-0.059	0.012	0.003	0.006
-1.888	-1.838	3.565	3.378	3.470
1.598	1.941	2.553	3.768	3.102
-0.109	-0.059	0.012	0.003	0.006
-1.888	-1.838	3.565	3.378	3.470
1.598	0.941	2.553	0.886	1.504
$\sum(T_{in} - \overline{T_{\imath n}}) =$ 48.498	$\sum(T_{enrooled} - \overline{T_{enrolled}}) =$ 48.448	$\sum(T_{in} - \overline{T_{\imath n}})^2 =$ 27.074	$\sum(T_{enrooled} - \overline{T_{enrolled}})^2 =$ 25.589	$\sum(T_{in} - \overline{T_{\imath n}})\,(T_{enrooled} - \overline{T_{enrolled}}) =$ 25.942

Then the correlation between input features (T_{in}) and enrolled features ($T_{enrolled}$) is calculated using equation (1) as follows:

$$\text{Score} = \frac{\sum_m \sum_n \left(\text{Tin}_{mn} - \overline{\text{Tin}}\right)\left(\text{Tenrolled}_{mn} - \overline{\text{Tenrolled}}\right)}{\sqrt{\sum_m \sum_n \left(\text{Tin}_{mn} - \overline{\text{Tin}}\right)^2 \left(\sum_m \sum_n \left(\text{Tenrolled}_{mn} - \overline{\text{Tenrolled}}\right)^2\right)}} \tag{1}$$

The correlation between input features and enrolled features are to be considered as matching scores.

3.6 Match Score Level Fusion

From each biometric matcher the generated scores are amalgamated into a single score by utilizing equation to get the arithmetic mean of matching scores in (2):

$$\text{Mean score} = \frac{1}{n} \sum_{i=1}^{n} \left(\text{Score}\right)_i \tag{2}$$

Where, n- number of biometric matchers and $(Score)_i$ - matching scores obtained from corresponding matching modules.

3.7 Authentication Decision

The final decision is determined on basis of Mean score. The threshold is set such that the genuine user will not be rejected and imposter user will not be accepted. If the mean score is greater than the threshold, the corresponding individual is verified; otherwise, the person is classified as an imposter, and access to the system is prohibited.

3.8 Evaluation Metrics

The suggested approach is assessed using the following evaluation metrics, as depicted in equations (3) to (9), and the outcomes are presented in section IV.

Accuracy is the percentage ratio of correct predictions to total forecasts.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

False Acceptance Rate (FAR) is the ratio of erroneous positive forecasts divided by the total number of negatives.

$$\text{FAR} = \frac{FP}{FP + TN} \tag{4}$$

False Rejection Rate (FRR) is the ratio of erroneous negative estimates divided by the sum of positives.

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \tag{5}$$

Equal Error Rate (EER) refers to the state where the false acceptance and rejection rates are equal

$$\mathrm{EER} = 0.5\left(\mathrm{FAR} + \mathrm{FRR}\right) \tag{6}$$

Precision: The number of positive class forecasts that really belong to the positive class is referred to as precision.

$$\mathrm{Pr\,ecision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

Recall: The recall is the number of correct positive class predictions produced from all correct positive samples in a dataset.

$$\mathrm{Re\,call} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

Specificity: The proportion of actual negatives, which got correctly predicted as the negative is defined in terms of Specificity.

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{9}$$

4 Results and Discussion

The empirical values of the finger vein-based biometric recognition system's proposed technique are provided here. The performance of the unimodal biometric system is given in Table 6.

Table 6

Performance of unimodal system with various learning rates

Learning Rate	**INDEX FINGER**			**MIDDLE FINGER**			**RING FINGER**		
	FAR	**FRR**	**Accuracy (%)**	**FAR**	**FRR**	**Accuracy (%)**	**FAR**	**FRR**	**Accuracy (%)**
0.03	0.28	0.24	97.43	0.22	0.25	97.79	0.26	0.28	97.7
0.025	0.25	0.27	97.71	0.24	0.27	97.72	0.22	0.21	97.5
0.01	**0.22**	**0.26**	**97.74**	**0.24**	**0.22**	**97.82**	**0.22**	**0.24**	**97.8**
0.001	0.23	0.25	97.63	0.23	0.25	97.8	0.28	0.29	97.6

0.0005	0.28	0.24	96.4	0.32	0.24	97.83	0.26	0.28	97.6
0.0002	0.45	0.46	97.12	0.42	0.25	97.8	0.42	0.41	97.5
0.0001	0.45	0.43	96.08	0.44	0.22	96.5	0.38	0.48	96.2

The CNN is tested with various learning rates such as 0.03, 0.025, 0.01, 0.001, 0.0005, 0.0002 and 0.0001 for index, middle and ring finger vein traits distinctly. The results show that the learning rate 0.01 achieves a maximum accuracy, low FAR and FRR compared to other learning rates employed. In the case of unimodal system, for learning rate of 0.01, accuracy of 97.74%, 97.82% and 97.8% are obtained for index, middle, and ring finger vein images respectively. As a consequence, the biometric system with ReLU activation function is examined with a learning rate of 0.01 for finger vein images and the obtained results are presented in Table 7.

Fundamental performance measures such as FAR and FRR guarantee that the matching algorithm does not make any erroneous positive and negative matches for single template comparison attempts. Metrics like as accuracy, specificity, precision, recall, and EER are used to validate the biometric system. The above-mentioned performance metrics are assessed so as to measure the performance of our proposed finger vein-based system and are presented in Table 7. The proportion of genuine positives and negatives in the entire data set is measured by accuracy. For the proposed system, we have obtained a higher accuracy for multimodal systems as 99.83% & 99.76% compared with unimodal systems. The proportion of false acceptance equals the proportion of false rejection, according to EER. The suggested multimodal system has an EER of around 0.125 for both left- and right-hand fingers.

Table 7

Performance analysis of proposed system in unimodal and multimodal modes

Metrics (%)	**Left-hand fingers**				**Right-hand fingers**			
	Index	Middle	Ring	Multi-modal	Index	Middle	Ring	Multi-modal
Accuracy	97.74	97.82	97.8	99.83	97.74	97.8	97.6	99.76
Precision	96.6	96.4	96.8	98.97	96.4	96.6	97.2	98.96
Recall	97.4	96.8	97.2	98.97	97.2	96.4	96.8	98.96
Specificity	97.2	97.8	97.4	98.8	97.4	97.6	97.2	98.8
EER	0.24	0.23	0.23	0.125	0.25	0.25	0.26	0.125
FAR	0.22	0.24	0.22	0.11	0.24	0.22	0.28	0.11
FRR	0.26	0.22	0.24	0.14	0.26	0.28	0.24	0.14

Table 8
Comparative analysis of related work

Authors	Biometric trait	Level of fusion	Methodology	Performance metrics

Yang Jinfeng and Zhang Xu [10]	Finger print & finger vein	Feature level	Supervised local-preserving canonical correlation analysis & nearest neighborhood classifier	FAR = 7.32
				FRR = 5.00

Sengar et al [20]	Palm print & fingerprint	Feature level	Deep Neural Network	FAR = 1.10
				FRR = 3.00

Lin et.al [25]	Palm and dorsal hand vein	Feature level	Hierarchical integrating function & Multi resolution analysis	FAR =1
				FRR = 3.5
				EER = 3.75

Kumar and Prathyusha[26]	Knuckle shape & dorsal hand vein	Score level	Matching vein triangulation and shape features	FAR = 1.14
				FRR = 1.14

Raghavendra et al [27]	Palm print & palm vein	Score level	Log Gabor transform, weighted sum rule	FAR =7.4
				FRR=4.8
				EER=2.2

			Non-standard mask, weighted sum rule	FAR =2.8
				FRR=1.4
				EER=2.2

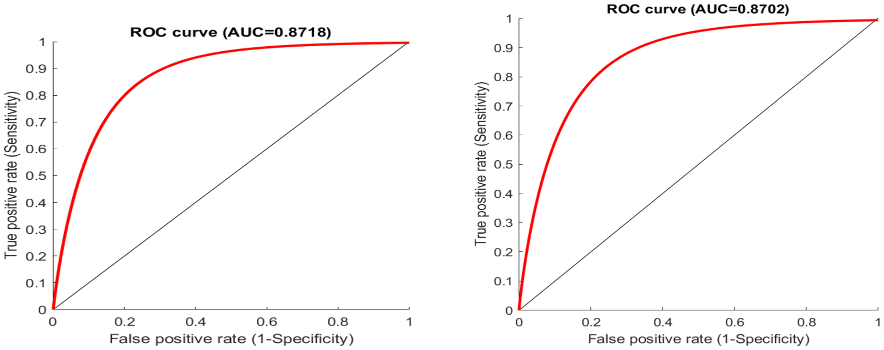
Proposed method	Index, middle and ring finger vein	Score level	CNN & correlation-based matching	FAR = 0.11
				FRR = 0.14
				EER =0.125

The precision of 98.97% and 98.96% is obtained for left and right-hand fingers respectively, specifies that the proposed system perfectly predict the genuine user. At the same time, recall ensures a measure of the proportion of actual genuine users got predicted as genuine. The proposed method achieves a recall of about 98.97% and 98.96% for left- and right-hand fingers. Hence the model is authoritative to the genuine users. The specificity of about 98.8% ensures that the system is faithful in the sense of identifying the imposter user.

The comparative analysis of the proposed work with existing methodologies has been done and is illustrated in Table 8.

The Receiver Operating Characteristic (ROC) curve of the suggested method is depicted in Figure 3. The ROC curve is a depiction of the True Positive Rate versus the False Positive Rate at different thresholds. The area under the curve

(AUC) shows a trade-off between genuine positive and false positive rates. The higher the AUC, the better the model distinguishes between real and impostor users. The suggested technique yields an AUC of around 0.8718 for the left-hand finger vein and 0.8702 for the right-hand finger vein, demonstrating that the proposed method accurately discriminates genuine and impersonating users.



(a) (b)

Figure 3

ROC-AUC curve (a) for Left hand (b) for Right hand

Conclusions

A multimodal, finger vein-based biometric system, that emphasizes on the fusion of index, middle and ring finger vein images of hand at the score level is proposed in this paper. The SDUMLA-HMT finger vein database was used to analyze the performance of the system. The essential traits from the finger vein images were extracted using CNN with the ReLU activation function. The correlation-based matching approach is then utilized to match the input template and the enrolled templates, and the match scores are fused using the arithmetic mean-based approach. The proposed multimodal system has maximum accuracy of 99.83% compared to average accuracy of 97.8% in unimodal case. As a result, the suggested multimodal biometric system provides a better solution to the difficulties in the unimodal biometric system. Moreover, in terms of FAR, FRR, EER, the proposed system outperforms the existing methodologies.

Security of biometric traits is essential, because the risk goes all the way up to the database, this work will be further developed, to strengthen the security of the biometric system by keeping encrypted templates, as a reliable, user-specific identification.

References

[1] Munish Kumar and Priyanka, “Finger print Recognition System: Issues and Challenges”, International Journal for Research in Applied Science & Engineering Technology, Volume 6, Issue 2, pp. 556-561, 2018

[2] Noyes E., Davis JP., Petrov N., Gray K. L. H and Ritchie K. L., “The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers”, . R. Soc. Open Sci. 8: 201169, 2021

[3] A. Dantcheva, C. Chen and A. Ross, "Can facial cosmetics affect the matching accuracy of face recognition systems?," IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, USA, pp. 391-398, 2012

[4] Xie. C., and Kumar.A., “Finger Vein Identification Using Convolutional Neural Network and Supervised Discrete Hashing,” . In: Bhanu, B., Kumar, A. (eds) Deep Learning for Biometrics. Advances in Computer Vision and Pattern Recognition, Springer, 2017

[5] Akintoye. K. A., Mohd Rahim. M. S and Abdullah A. H, “Challenges of Finger Vein Recognition System: A Theoretical Perspective,” Journal of Computational and Theoretical Nanoscience, Vol. 8, No. 2, pp. 196-204, 2018

[6] Khellat-Kihel. S, Abrishambaf. R, Monteiro. J. L and Benyettou. M, “Multimodal fusion of the finger vein, fingerprint and the finger-knuckle-print using Kernel Fisher analysis,” Applied Soft Computing, Vol. 42, pp. 439-447, 2016

[7] Bharathi, S, Sudhakar, R and Balas, VE, ‘Biometric Recognition Using Fuzzy Score Level Fusion’, International Journal of Advanced Intelligence paradigms, Vol. 6, No. 2, pp. 81-94, 2014

[8] Bharathi, S, Sudhakar, R and Balas, VE, ‘Hand vein based Multimodal biometric Recognition,’ Acta Polytechnica Hungarica, Vol. 12, No. 6, pp. 213-229, 2015

[9] Xuekui Yan, Wenxiong Kang, Feiqi Deng and Qiuxia Wu, “Palm vein recognition based on multi-sampling and feature-level fusion,” Neurocomputing, Vol. 151, No. 2, pp. 798-807, 2015

[10] Jinfeng Yang and Xu Zhang, “Feature-level fusion of fingerprint and finger-vein for personal identification,” Pattern Recognition Letters, Vol. 33, No. 5, pp. 623-628, 2012

[11] Yong-Fang Yao, Xiao-Yuan Jing and Hau-San Wong, “Face and palm print feature level fusion for single sample biometrics recognition”, Neurocomputing, Vol. 70, No. 7, pp. 1582-1586, 2007

[12] Ialiang Peng, Ahmed A. Abd El-Latif, Qiong Li and Xiamu Niu, “Multimodal biometric authentication based on score level fusion of finger biometrics,” Optik, Vol. 125, No. 23, pp. 6891-6897, 2014

[13] Maleika Heenaye and Mamode Khan, “A Multimodal Hand Vein Biometric based on Score Level Fusion,” Procedia Engineering, Vol. 41, pp. 897-903, 2012

[14] Gurjit Singh Walia, Tarandeep Singh, Kuldeep Singh and Neelam Verma, “Robust multimodal biometric system based on optimal score level fusion model,” Expert Systems with Applications, Vol. 116, pp. 364-376, 2019

[15] Hiew Moi Sim, Hishammuddin Asmuni, Rohayanti Hassan and Razib M. Othman, “Multimodal biometrics: Weighted score level fusion based on non-ideal iris and face images,” Expert Systems with Applications, Vol. 41, No. 11, pp. 5390-5404, 2014

[16] Udresh Dwivedi and Somnath Dey, “Score-level fusion for cancelable multi-biometric verification,” Pattern Recognition Letters, Vol. 126, pp. 58-67, 2019

[17] Khellat-Kihel. S., Abrishambaf. R, Monteiro. J. L and Benyettou. M, “Multimodal fusion of the finger vein, fingerprint and the finger-knuckle-print using Kernel Fisher analysis,” Applied Soft Computing, Vol. 42, pp. 439-447, 2016

[18] Kun Su, Gongping Yang, Bo Wu, Lu Yang, Dunfeng Li, Peng Su and Yilong Yin, “Human identification using finger vein and ECG signals,” Neurocomputing, Vol. 332, pp. 111-118, 2019

[19] Ammour. B., Boubchir. L., Bouden. T and Ramdani. M, “Face–Iris Multimodal Biometric Identification System,” Information analysis and processing, Vol. 9, No. 1, 2020

[20] Sengar. S. S., Hariharan. U and Rajkumar. K, “Multimodal Biometric Authentication System using Deep Learning Method,” International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp. 309-312, 2020

[21] Balas. V. E., Roy. S. S., Sharma. D and Samui, P. (Eds.). “Handbook of deep learning applications” Springer, Vol. 136, 2019

[22] Macsik. P., Pavlovicova. J., Goga. J and Kajan. S, “Local Binary CNN for Diabetic Retinopathy Classification on Fundus Images” Acta Polytechnica Hungarica Vol. 19, No. 7, 2022

[23] Dobes. M and Sabolova. N, “Emotion Recognition Using Pretrained Deep Neural Networks”, Acta Polytechnica Hungarica Vol. 20, No. 4, 2023

[24] The SDUMLA-HMT database [Online] Available link: http://mla.sdu.edu.cn/sdumla-hmt.html

[25] Lin C. L and Fan K. C, “Biometric Verification Using Thermal Images of Palm-Dorsa Vein Patterns”, IEEE Transaction on Circuits System Video Technology, Vol. 14, pp: 199-213, 2004

[26] Kumar. A and Prathyusha K. V, "Personal authentication using hand vein triangulation and knuckle shape", IEEE Transaction on Image Processing, Vol. 18, pp: 2127-2136, 2009

[27] Raghavendra. R., Imran. M. Rao. A and Kumar. G. H, "Multi modal biometrics: analysis of hand vein and palm print combination used for personal verification", Proceedings of IEEE International Conference on Emerging Trends and Engineering Technology, 2010

Extended Linear Regression and Interior Point Optimization for Identification of Model Parameters of Fixed Wing UAVs

Huda Naji Al-sudany, Béla Lantos

Budapest University of Technology and Economics, Hungary
Magyar tudósok krt. 2, H-1117 Budapest, Hungary
E-mail: alsudany@iit.bme.hu, lantos@iit.bme.hu

Abstract: The paper deals with the identification of the system parameters in the nonlinear dynamic model of fixed wing UAVs. Fixed wings airplanes are popular in long distance applications and have to be modelled accurately to guarantee efficient control properties. Different methods are suggested to solve the parameter estimation beginning with the standard linear regression (LR) and continued with its extension (ELR), the optimization using interior point methods and finally using the FireFly technique, which is a metaheuristic algorithm. The methods illustrate the convergence and the speed of these approaches. The known default parrameter values of a Sekwa UAV were used to demonstrate that the elaborated identification methods can also reconstruct the numerical values of the dimensionless system parameters embedded into the nonlinear model using the physical weighting functions. After the extension of linear regression, MinMax optimization algorithm was used to get the best and optimal solution and reconstruct the parameters of the Sekwa aircraft. FireFly optimization gives also comparable results with minmax method. Flight data was needed for simulation and testing the different approaches. Matlab and toolboxes were used as simulation software. The results showed the estimated parameters are more accurate than linear regression estimation. Even though it is a small improvment this will reflect in all calculations.

Keywords: Aircraft model identification; linear regression;Min Max optimization; FireFly optimization

1 Introduction

Analysis and parameter prediction are vital in achieving the control and stability properties of a somewhat complicated system like an aircraft. The parameter estimation method has been used with great success in the past to predict numerous parameters depending of real data of flight. Considering that the rigid body model is reliable, parameter estimate using flight data is currently employed often when applied to airplanes in the linear flight domain. As a result, the derivation of aircraft model is used to estimate procedure lacks elastic degrees of freedom. High levels of flexibility in an aircraft may make it more susceptible to the dynamics of a system with too many factors that must be evaluated [1] [2].

The identification of an accurate and verified mathematical model of aircraft apparatus is known as aircraft system identification. This is a crucial stage in the development of flight vehicles due to the generated model is vital for:

I. Comprehending the cause-and-effect relationship.

II. Examining the capabilities and characteristics of aircraft.

III. Verifying aerodynamic databases and upgrading flight control law designs are steps in step three.

IV. Supporting the expansion of the flying envelope.

V. Attempting to recreate the flight path, which incorporates incidence analysis and wind calculation.

VI. Running adaptive control and fault diagnosis.

This article is a natural extension of our previous research efforts. Building upon our prior work, titled "Comparison of Adaptive Fuzzy EKF and Adaptive Fuzzy UKF for State Estimation of UAVs Using Sensor Fusion," in [3] where we focused on state estimation techniques, and "Prediction of the Navigation Angles Using Random Forest Algorithm And Real Flight Data of UAVs," in [4] which explored machine learning-based prediction of navigation angles, this current article delves into a new dimension of UAV research. Specifically, it addresses the critical task of model parameter identification for Fixed Wing UAVs, utilizing extended linear regression and interior point optimization techniques.

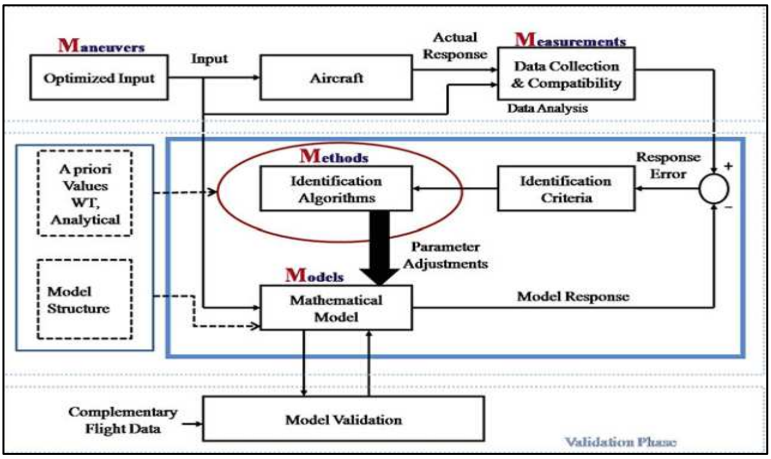


Figure 1

Block diagram of aircraft [5]

By citing our previous work, we establish the continuity and progression of our research, highlighting how our latest contribution complements and broadens the scope of our ongoing efforts in advancing UAV technology.

The method for identifying an aircraft system is described in Figure 1.

1.1 State of Art

The authors in [6] discussed how to use measurable input and output data to estimate parameters in aircraft flight dynamic models, like control and stability derivatives. In this method, the aircraft control effectors are moved using orthogonal phase-optimized multisines, frequency responses of MIMO systems are computed using Fourier analysis, and noise values of the parameters of model are determined using frequency response error (FRE), a maximum likelihood estimator. The T-2 generic transport system and the X-56A aeroelastic model are examples of airplanes whose flight test results are used to illustrate the technique. By using the maximum likelihood estimator, one may easily incorporate prior knowledge and combine data from many motions without additional correction while also giving precise statistical uncertainties for the expected values. However, the tactic still has several shortcomings. Although frequency responses provide physical insight into the dynamics, their usage restricts modeling to linear, time-invariant systems, necessitating flight test data with low disturbances compared to a reference state. While other approaches can benefit from fewer data records, a meaningful frequency response estimate requires steady-state data, which take longer to gather. If there is environmental disturbance, more loops of steady-state data may be needed. The strategy requires the capacity to extend the command path with computerized inputs. Uncertainties or uncertain environment can cause many problems that need to be solved [7].

In [8], Online system identification was covered by the authors; as technology advanced, it became a crucial step in the construction of methods for estimating aerodynamic parameters. In order to estimate the aerodynamic parameters of fixed-wing aircraft in unsteady conditions like stalls, this research proposes two online system identification (SID) algorithms that are based on Kalman filters. The suggested approaches, in contrast to previous SID ones, incorporate aerodynamic features associated with the upset state directly into the aircraft dynamics, such as modification of aerodynamic derivatives or flow separation point. To get the best estimations of the relevant aerodynamic characteristics, the standard or unscented Kalman filter is then applied in real-time. To illustrate the usefulness of the proposed approaches and their superiority to a previously proposed method, real flight data sets from a variety of aircraft are used in testing.

In [9], authors deal on the issue of fixed-wing UAV modeling and control. For control purposes dynamic models for aircraft were given in body, stabilization-axes, and wind-axes coordinate systems. It was possible to define and resolve a typical integral backstepping control issue that ensures stability in closed loops. Integral parts of the control can aid in reducing parameter changing and disruption impacts. The approach ensures that the combined 3D attitude system will remain stable in a closed loop even with changing reference signals. For motion portions that can be joined and smoothed, primitives for path design were developed. The amount of time required for each section can be specified. They worked on Sekwa aircraft.

The techniques for identification are used specifically for accurate representation of system of aircraft. Many studies might use nominal values but this will reflect in their works. Some studied attitude estimation using Kalman filter and artificial intelligence as in [9] [10] [11] and the results are great but still some errors because of not estimation the aircraft parameters. Optimization also has roles in medicine and communications [12] [13] [14].

In this approach, an extension of Linear regression is used to identify Sekwa aircraft parameters accurately with new mathematical representations. The goals of this search is to implement Linear regression with an extension for aircraft parameters estimation for a specific aircraft, the linear regression is implemented with interior point optimization algorithms , but our aim is to extend this to get more accurate results. The structure of this paper is as follows. Section 2 deals with the developed algorithms including the nonlinear UAV models, the basic regression model the interior point optimization, the extended linear regression, and FireFly optimization. Section 3 presents and analyses the identification results based on standard linear (LR) and extended linear regression (ELR), and FireFlytechnique. Since the default parameter values of the Sekwa UAV are available hence the identified parameter values are also compared in the parameter domain. The paper is finished with the Conclusions, and References.

2 Developed Algorithm

2.1 Nonlinear UAV Model

Fixed wing propeller driven aircrafts are nowadays popular for both long distance military and civilian applications. The paper assumes that the reader is familiar with the fundamentals of the nonlinear dynamic model and control of fixed wing aircraft. There are excellent classical books on this field, specialities regarding UAVs can be found in [9] [10]. The paper concentrates on the parameter identification of the models, see here only some details based on [16]. It is assumed that registered flight data are available obtained by data logging either within teleoperation or online control. The identification can be performed using batch-processing or online. Bach-processing makes it easy to use special sotwares, state estimation and filtering and differentiation of the estimated signals before starting the identification process. The notations used are the well spread ones in vehicles and robotics literature.

The use of coordinate systems (frames) are preferred, namely Kn, Kb, Ks and Kw are the Flat-Earth, body, stability axis and wind axis frames, respectively. The parametrisation of the model can often be performed in the stability axies frame or in the wind axis frame. Denote v_b and ω_b the body linear velocity and angular velocity, the relative air speed is $v_r = v_b - R_n^b v_{wind}^n$ where the wind velocity is constant and R_n^b transforms vectors from Kn to Kb, m is the mass and $J = I_c$ is the inertia matrix, and F_B and T_b are the resulting external force and torque satisfying the Newton-Euler equations:

$$\dot{v}_b = -\omega_b \times v_b + F_B / m \text{ and } J\dot{\omega}_b = -\omega_b \times (J\omega_b) + T_B \tag{1}$$

where the force and moment effects are

$$F_B = (F_x, F_y, F_z)^T = F_{BA} + F_{BT} + F_{BG} \text{ and } T_B = (\overline{L}, M, N)^T = T_{BA} + T_{BT} \tag{2}$$

Here the second letters A,T and G in the indexes denote the aeodynamic, trust and gravity effects, respectively. The (nongravity) forces and torques depend on the wing reference area S_{wa}, the free-stream dynamic preassure $\overline{q} = 1/2\rho v_T^2$, different dimensionless coefficients $C_D, C_L, C_Y, C_l, C_m, C_n$ and, in the case of the torques, on the wing span b and the wing mean geometric chord $\overline{c}$. The dimensionless coefficients depend in first line on the angle of attack α and the sideslip angle β, the control surfaces and the Mach-number:

$$\begin{array}{rc}
D_{stab} = \overline{q}S_{wa}C_D & \text{drag} \\
L_{stab} = \overline{q}S_{wa}C_L & \text{lift} \\
Y = \overline{q}S_{wa}C_Y & \text{sideforce} \\
\overline{L} = \overline{q}S_{wa}bC_l & \text{rolling moment} \\
M = \overline{q}S_{wa}\overline{c}C_m & \text{pitching moment} \\
N = \overline{q}S_{wa}bC_n & \text{yawing moment}
\end{array} \tag{3}$$

Notice that the lift force C_L, the drug force C_D, etc. are usually defined in the stability frame, not in the wind-axes frame. The relative air-speed can be transformed using elementary transformations in the wind-axis frame by $v_r = R_w^b v_w = Rot(y, -a)Rot(z, \beta)v_w$ where $v_w = (1,0,0)^T v_T$. For the stability frame $\beta = 0$. The resulting external force in the body frame is

$$F_B = \overline{q}S_{wa}(C_X, C_Y, C_Z)^T + (1,0,0)^T F_T + mg_B \tag{4}$$

where F_T is the thrust force and g_B is the gravity acceleration in the body frame. The drug and lift forces in the stability frame satisfy

$$
\begin{aligned}
&(-C_D, C_Y, -C_L)^T = Rot(y,-\alpha)^T (C_X, C_Y, C_Z)^T \\
&(-C_{DW}, C_Y, -C_{LW})^T = Rot(z,\beta)^T (-C_D, C_Y, -C_L)^T
\end{aligned}
\tag{5}
$$

In order to obtain forces and moments in standard SI dimensions (N and Nm) the dimensionless components should be multiplied by appropriate weighting functions. For example, the angular velocity in the stability frame can appear in the model as weighting function where $\omega_S = Rot(y,-\alpha)^T \omega_b$. Hence, identifying the parameters appearing in the model, the linear parameter estimation is embedded in the nonlinear models through the weighting functions. First, the user chooses the structure of the model, i.e. the number of parameters, the form of the high-level functions and the weighting signals in them:

$$
C_i = C_i(p_{i1}, p_{i2}, \ldots, s_{i1}, s_{i2}, \ldots), \quad i \in \{D, L, Y, C_X, \ldots, C_n\}
\tag{6}
$$

Where p denotes parameters and s denotes weighting signals of the function. For constant weight $s=1$ is allowed.

On the other hand, the structure of the nonlinear model can contain nonlinear relations too, for example the drug may depend on the square of the lift, making the parameter estimation nonlinear or constrained linear. Such a situation is typical for many aircraft in steady state:

$$
\begin{aligned}
&C_L = C_{L0} + C_{L\alpha}\alpha \\
&C_D = C_{D0} + \frac{C_L^2}{\pi A_{sr} e_{Osw}} = \overline{C}_{D0} + \overline{C}_{D1}\alpha + \overline{C}_{D2}\alpha^2
\end{aligned}
\tag{7}
$$

Where A_{sr} and e_{Osw} denote the wings aspect ratio and the Oswald efficiency factor, respectively. Taking the square, the introduced new parameters $\overline{C}_{D0}, \overline{C}_{D1}, \overline{C}_{D2}$ make the problem formally similar to linear parameter estimation in these parameters but it is evident that they are in relation with C_{D0}, C_{L0} and $C_{L\alpha}$ generating constraints amongst them. Such a situation appears for the Sekwa fixed wing UAV causing problem in the identification. Neglecting the constraints for the parmeters $\overline{C}_{D0}, \overline{C}_{D1}, \overline{C}_{D2}, C_{D0}, C_{L0}, C_{L\alpha}$ the usual least square parameter esttimation is not necessarily convergent in $C_{D0}, C_{L0}, C_{L\alpha}$ to the correct values. Notice that this property is in force also for other methods if in them dominate the linear character together with some noise extension. Fortunately, the constraints are simple and sparse, hence using nonlinear optimization with constraints (see for example fmincon with options, starting from LS generated initial values) give an extra chance for parameter improvement. Returning back to the linear parameter estimation problem, one can assume that from the flight data and state estimation the states in the differential equations are available and the differentiation of the signals has already been performed. The LS problem can be considered in the form

$y(t)=\varphi^{T}(t)\vartheta+e(t)$ where $y(t)$ is a vector coming from the differential equations and ϑ is the (full) parameter vector. In case of Sekwa UAV the set of weighting functions may be

$$(1,\alpha,\alpha^{2},\beta,\hat{p},\hat{q},\hat{r},\delta_{e},d_{a},\delta r,\delta_{th}) \tag{8}$$

and in case of $C_D=(\bar{C}_{D0},\bar{C}_{D1},\bar{C}_{D2})$ this component can be represented by

$$(1,\alpha,\alpha^{2},\beta,\hat{p},\hat{q},\hat{r},\delta_{e},d_{a},\delta r,\delta_{th})\times(1,1,1,0,\ldots,0)^{T} \tag{9}$$

Similar relations can be found for the other components in the differential equations. In this way the derivatives of the state equations subtracted by signals not containing identifiable parameters in the state equation playes the role of $y(t)$ and the right side is as above for C_D. Then, the components can be collected for every sampling time multiple t and the LS problem can be built up and solved. A similar technique can be used for the examined more complex problems in the paper.

2.2 Problem Statement

After defining parameters and basic coefficients, the main problem of this research is how to estimate and predict these coefficients depending on real data. In regard, three main method are illustrated which will be explained in the next sections: Linear Regression (LR), Extended Linear Regression (ELR) and Optimized ELR by Firefly optimization algorithms. We deal with the solution of the parameter identification problem at two levels.

1) Normally, the results can be tested only to demonstrate that the output signals in the flight data can be well matched if the simulated model, inside with the identified parameters, is driven with the input (actuator) signals of the flight data. This can be checked in open loop or in closed loop under control. Notice that integral components in closed loop can reduce the effect of errors while matching during open loop testing is more difficult. Here, dominates signal comparison.
2) In rare situations the internal parametrs of the aircraft may also be known and we can test whether the default parameters can also be reconstructed using identification. Here, dominate parameter comparision, which is more general.

The studied aircraft was the Sekwa UAV with available known default model parameters. The details can be found in [15] [16] [17]. Hence. it was also possible that beside the convergence of the parameter identification also the parameter errors could have been tested. The comparison can be found later in table. Next section will explain the general regression model and Firefly optimization algorithm.

2.3 Basic Regression Model

The main idea of our work is re-modelling the mathematical equation of aircraft model from a new view, suppose the next model:

$$f_j = \sum_{i=1}^{M} H_j(x_i) + b_j + c_j + d_j, \ \ j \in [1, n] \tag{10}$$

Where: n: is the number of all samples. M number of variables to be estimated. a, b, c are known values, H_j is nonlinear function in general (it might be linear). f is the dependent variable that can be measured or observed. x_i are cofficients to be estimated (it is vector). As an example, f might be the position and x_i might be the velocity and a, b, c are parameters that are initial values or noise parameters. In linear case, they reduce to next (* is the normal multiplication):

$$f_j = \sum_{i=1}^{M} H_{ji} * x_i + b_j + c_j + d_j, \ \ j \in [1, n] \tag{11}$$

Whereas H_{ji} are independent known variables. (H could be represented by vector)

$$\sum_{i=1}^{M} H_{ji} * x_i = \left[H_{j1}, H_{j2}, \ldots, H_{jM} \right] * \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_M \end{bmatrix}$$

To solve this issue

$$f_j - \left(b_j + c_j + d_j \right) = \sum_{i=1}^{M} H_{ji} * x_i, \ \ j \in [1, n] \tag{12}$$

And it became as a normal regression model; note that H_i might contain old values of f_i, as an example $H_5 = h_1 x_1 + f_2 x_3 + \cdots$. To solve this type of regression model, we need to build equation as follows

$$\begin{aligned} &Y = f_j - \left(b_j + c_j + d_j \right), \ \ A \cdot X = \sum_{i=1}^{M} H_{ji} * x_i \\ &Y = A \cdot X \\ &X = (A)^{-1} \cdot Y \end{aligned} \tag{13}$$

Where, $\boldsymbol{A}$ is a matrix with n rows (number of samples) and M columns (number of parameters to be estimated. The matrix $\boldsymbol{A}$ is not square. The next equation show dimensions

$$\underset{n*1}{Y} = \underset{n*M}{A} \cdot \underset{M*1}{X} \tag{14}$$

Pseudo inverse matrix is needed here to calculate the vector X. The Sekwa model equations will be used to find the aircraft coefficients (Force equations, Torque equations and Navigation equations).

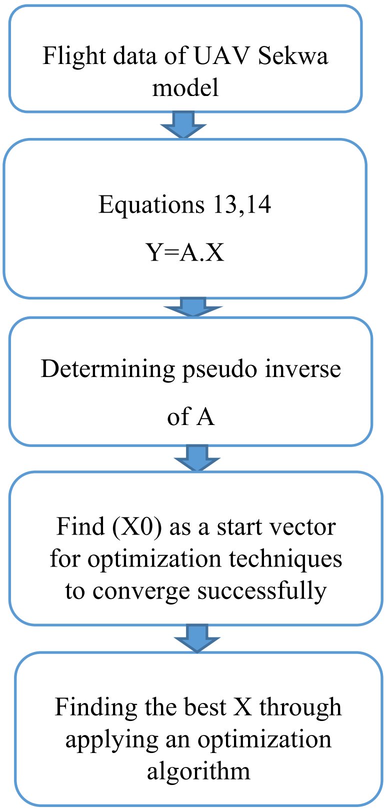

Flight data of UAV Sekwa model
Equations 13,14
Y=A.X
Determining pseudo inverse of A
Find (X0) as a start vector for optimization techniques to converge successfully
Finding the best X through applying an optimization algorithm

Figure 2
Flowchart of Sekwa model parameters estimation

Linear regression depends on Interior point optimizsation, which will be explained next.

2.3.1 Interior Point Optimization Function

Interior point algorithms are a certain class of algorithms that solve linear and nonlinear convex optimization problems. It enabled solutions of linear programming problems that were beyond the capabilities of the simplex method. Contrary to the simplex method, it reaches a best solution by traversing the interior of the feasible region. The method can be generalized to convex programming based on a self-concordant barrier function used to encode the convex set.

After we get the equation:

$$\underset{n*1}{Y} = \underset{n*M}{A} \cdot \underset{M*1}{X} \tag{15}$$

We run an optimization function to find best value of X , we need this step since we got number of equations larger than number of coefficients. The optimization algorithm in [15] is to find minimum of constrained nonlinear multivariable function, it is available in Matlab [18].

2.4 Extended Linear Regression

The above representation of equation is accurate for ideal situations and ideal flight environment, so new term is added to each equation to get close to accurate calculations, but this added term is not for noise. Hence, if we want to simulate noise state too, we have to add n term for n sample. For aircraft, each coefficient might be dependent from all other ones; so if the number of coefficients is M, another M terms will be added to extend the regression as follows:

$$f_j = \sum_{i=1}^{M} H_j(x_i) + b_j + c_j + d_j + \sum_{i=1}^{M} G_j\left(eps_i\right), j \in \left[1, n\right] \tag{16}$$

The aim is to estimate $x_1, x_2, \ldots, x_M$ so we add $eps_1, eps_2, \ldots, eps_M$ term. Whereas G_j are independent known variables . To solve this issue in linear case:

$$G_j(eps_i) = G_{ji}.f_j.esp_i$$

$$f_j - \left(b_j + c_j + d_j\right) = \sum_{i=1}^{M} H_j \cdot x_i + \sum_{i=1}^{M} G_{ji}.f_j.eps_i\,, j \in [1, n] \tag{17}$$

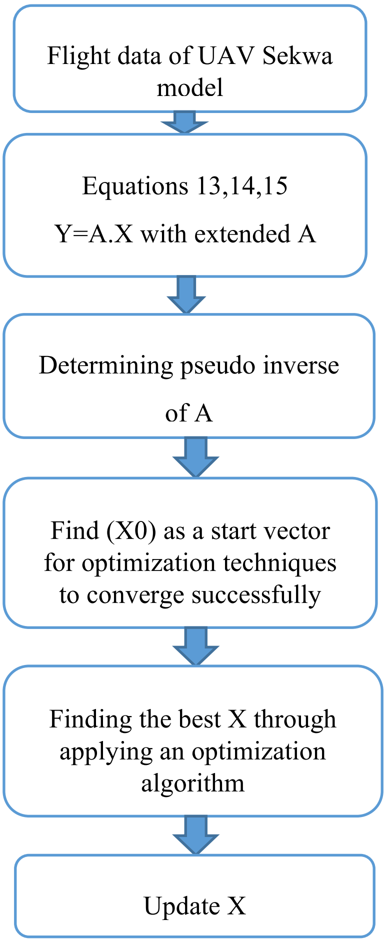
Suppose the term $\tilde{X} = [x_1, x_2, \ldots x_M, eps_1, eps_2, \ldots eps_M]$ whereas

$$\underset{n*1}{Y} = \underset{n*2M}{A} \cdot \underset{2M*1}{X} \tag{18}$$

This will extend the matrix and will end up with more accurate and reality representation of the aircraft model. The G_{ji} are zeroes or ones; this according to the existence if x_i. This means if coefficient with index depends on coefficients with index i then $G_{ji} = 1$ and then eps_i should be estimated

$$G_{ji} = \begin{cases} 1 & if\ x_i\ is\ exist\ \ in\ \ \ equation \\ 0 & \ \ if\ x_i\ is\ \ not\ \ in\ \ equation \end{cases} \tag{19}$$

The final desired X might be $X = \left[x_1, x_2, \ldots x_M\right]$; this means that the calculated eps_i compensates noise. Figure below shows these stages:


Flight data of UAV Sekwa model
Equations 13,14,15
Y=A.X with extended A
Determining pseudo inverse of A
Find (X0) as a start vector for optimization techniques to converge successfully
Finding the best X through applying an optimization algorithm
Update X

Flowchart of Sekwa model parameters estimation with extended regression

For ELR, MinMax optimization function is used, this will be explained next. It is a ready function in matlab toolbox.

2.4.1 MinMax Optimization Function

The optimization function here is “Solve minimax constraint problem”. fminimax function searches for the best solution that minimizes the maximum of a set of objective functions. The problem includes any type of constraint. The optimization function needs an objective function to be minimized or in some cases to maximized (minimize loss or error, maximize accuracy) [19] [20] [21]. All optimization algorithms need a loss or an objective function to be minimzed during optimization phase (searching for optimal solution, this will be explained next).

2.4.2 Objective Function

For linear problems the tak is to find the solution of linear regression [22] [23] for equation (18):

$$Y = A \cdot X$$

To solve this problem, new objective function can be defined for minimizing:

$$\begin{aligned} fun = \min_{X} \|Y - A \cdot X\|^2 &= \left(Y - A \cdot X\right)^T \left(Y - A \cdot X\right) \\ &= X^T A^T AX - 2X^T A^T Y + Y^T Y \end{aligned} \tag{20}$$

X can be simply determined minimizing fun. In our case, it is an optimization problem and need to search for a solution to minimize the objective function. Best solution would be when *fun*=0. Matmatically, $(Y - A \cdot X)^T(Y - A \cdot X) = X^T A^T AX - 2X^T A^T Y + Y^T Y$ can also be solved, especially in case of constraints, by using the preferable wayof numerical optimum seeking.

Constraints may be the lower and upper bound for X to defined. Then the defined values have to be satisfied during the algorithm.

Next, FireFly mehtod will be discussed to help ELR in searching for the best solution.

2.4.3 FireFly Optimization Algorithm

One of the newest metaheuristic algorithms for optimization issues is the firefly method. The program takes its cues from firefly flashing behavior. The flashlight is utilized as a warning system to keep the fireflies from potential predators [24]. The program will treat randomly produced solutions as firefly, and brightness will be allocated based on how well they perform on the objective function. They can divide into smaller groups due to their attractiveness, and each group converges around the local models [25]. The following three rules are the fundemantals of firefly described as follows [26]:

1. Fireflies come in both genders.
2. Attractiveness is proportional directly to their brightness.
3. The landscape of the objective function controls a firefly's brightness.

When compared to other algorithms, firefly offers two key advantages: automated subdivision and the capacity to handle multimodality.

The key parameters of the Firefly Algorithm are as follows:

Population Size (n): The number of fireflies. We used n=150.

Light Intensity (I): This represents the objective function value or fitness of a solution. Fireflies are attracted to brighter fireflies, meaning that solutions with higher light intensity values are considered better.

Absorption Coefficient (γ): This parameter represents the light absorption during the propagation of light. It's used to reduce the attractiveness of a firefly based on the distance between them. If the distance is high, this means that this solution is not close, so will decrease attractive parameter. Where **Absorption Coefficient (γ)=0.99.**

Maximum Generations (MaxGen): The number of iterations or generations the algorithm will run before terminating. It controls the stopping criterion for the optimization process. Where **Maximum Generations=500.**

Objective Function: The mathematical function that represents the problem to be optimized. This represents the function in equation (20).

Search Space and Bounds: This represents the constraints. It should be mentioned the initial values were determined after many experiments.

The problem is searching for best X that minimizes the function in equation (20). Firefly will be adopted to search for this X by its fireflies. Simple flowchart is shown below:

```
Begin
  1) Objective function $f(x)$ to be minimized.  As in equation (20)
  2) Generate an initial population of fireflies (fireflies number must be defined); Each firefly will search for the best value of X values in Equation (20), they will attract other fireflies when they found minimum values to repopulate the fireflies near the minimum zone and search again till they reach minimum value of function in Equation (20)
  3) Formulate light intensity $I$ so that it is associated with $f(x)$
   4) Define absorption coefficient $\gamma$
   while (t < MaxGeneration)
      for i = 1 : n (all n fireflies)
         for j = 1 : i (n fireflies)
            if ($I_j > I_i$),
               Vary attractiveness with distance $r$ via $e^{-\gamma r}$ ;
               move firefly i towards j;
               Evaluate new solutions and update light intensity; (the new solution here is the vector X)
            end if
         end for j
      end for i
      Rank fireflies and find the current best loss function that in equation (20);
   end while
end
```

This algorithm is applied after extended regression is done. After implementing FireFly, the initial input for this algorithm is

$$x_0 = \alpha * x_{elr} + (1-\alpha) * x_{erlm} \tag{21}$$

Where x_{elrm} is the modified extended linear regression output by using the extended parameters, x_{elr} is a vector with 52 parameters (first 26 parameters are the basic parameters and the other 26 parameters are the extended parameters).

$$x = \left[x_{elr}, x_{erlm}\right]$$

If we consider that the type of the x_{erlm} is not additive; the equation will be the following:

$$x_0 = \alpha * x_{elr} + (1-\alpha) * x_{elr} * x_{erlm} \tag{22}$$

This is known as multiplicative error. Then, Y is recalculated according to the initial values:

$$\underset{n*1}{Y} = \underset{n*2M}{A} \cdot \underset{2M*1}{X} \tag{23}$$

The algorithm stops when it reaches max generation. The algorithm saves the best solution each iteration, the solution contains the values of X and the value of function in Equation (20). The next section will illustrate the results of the developed algorithm. The X in Equation (20) is known for the aircraft, the developed algorithm will result also new X values, then there is a comparison to evaluate the developed algorithm to be used with another aircraft.

3 Identification Results

3.1 Flight Data

The used flight data is sampled by 0.01 seconds, it contains a real data of trip with all controls, and it contains all sensors data: accelerometers, gyroscopes and angular velocities, position data, velocity data and quaternion data.

3.2 Parameters Reconstruction

The code will result in a vector of length M ; it contains all target parameters that are explained in Table 1. These parameters used exist in [9, 10]. The flowchart in Figures 2 and 3 illustrated estimation of Sekwa model parameters.

3.3 Extended Linear Regression Results

To check the quality of the algorithm for later unknown airplanes, the X in Equation (20) was assumed here to be known (it has 26 parameters). The developed algorithm will result also new $\tilde{X}$ values that are solutions of the optimization algorithm. Comparing them, the quality of the method can be judged for later use if the correct parameters are unknown. The real values of Sekwa UAV parameters are from [15]. LR and ELR were implemented with interior point optimizations. The figures below show the results of LR and ELR regression. The y axis in next plots refers to true values of coffiecients that was explained in the introduction.

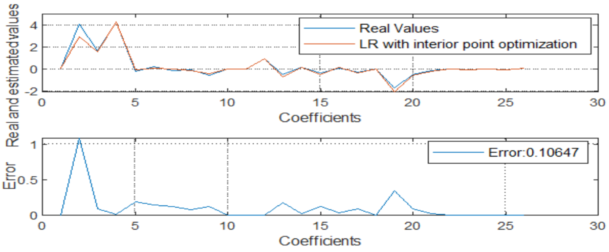


Figure 4

LR with interior point results

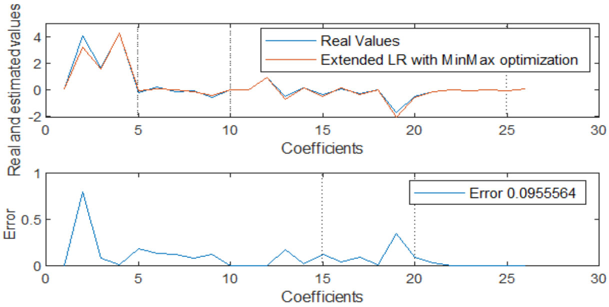


Figure 5

Extended LR with Min Max results

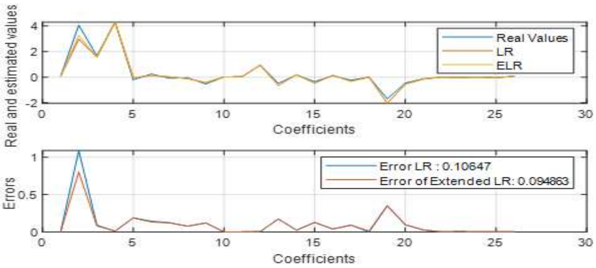


Figure 6

Compare results of LR and Extended LR

The formula for calculating the improvement ratio is used to minimize the relative error of two methods. It refers to the absolute value of the difference of the errors of (LR&Extended LR) divided by the error of LR.

$$improve\ ratio = \frac{abs(0.10647 - 0.094863)}{0.10647} = 10.9\%$$

3.4 FireFly Results

The estimation is shown below, where the figure contains our developed method Firefly ELR, linear regression and Firefly without the regression.

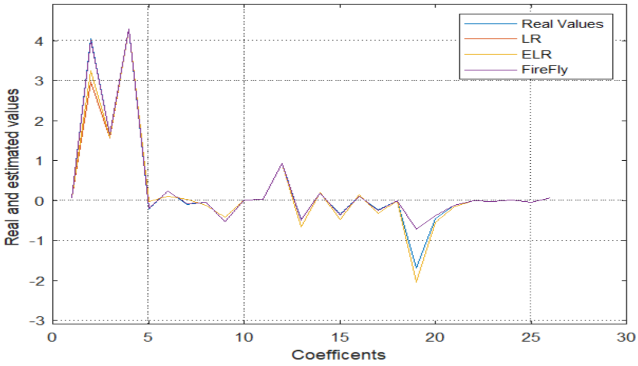


Figure 7
FireFly results

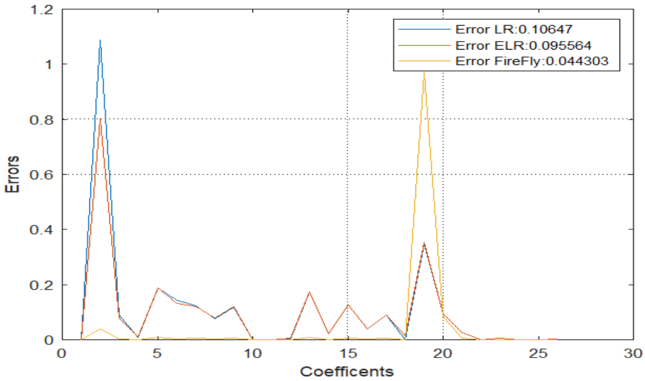


Figure 8
All methods results

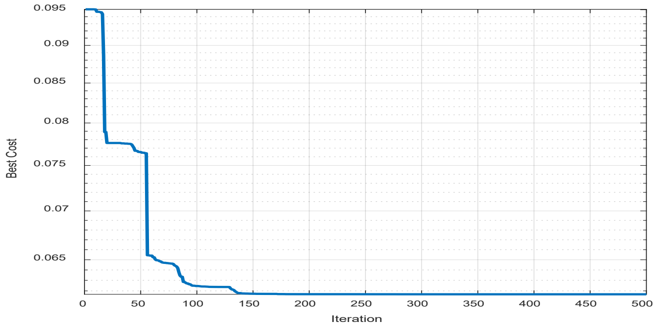


Figure 9

FireFly Cost function

3.5 Comparison of Parameter Estimation Results

The table below shows the comparison between our approach, Basic LR, only FireFly and FireFly ELR comparing to the true values.

Table 1

Results comparison

True value	FireFly	LR (Interior point)	Firefly ELR (Min Max)	name
0.0633	0.0633	0.0629	0.0625	C_{L0}
4.0543	4.0324	2.9675	3.2533	C_{Lq}
1.6524	1.6529	1.5611	1.5732	$C_{L\delta e}$
4.3	4.3006	4.2918	4.2882	$C_{L\alpha}$
-0.2114	-0.2021	-0.0236	-0.0244	C_{Yp}
0.2409	0.2345	0.0977	0.1096	C_{Yr}
-0.094	-0.0879	0.0284	0.0272	$C_{Y\delta a}$
-0.0478	-0.0516	-0.1231	-0.1251	$C_{Y\delta r}$
-0.5401	-0.5342	-0.4216	-0.419	$C_{Y\beta}$
-0.4848	-0.4934	-0.6578	-0.6578	C_{lp}
0.1704	0.1715	0.1933	0.193	C_{lr}
-0.352	-0.3584	-0.4787	-0.4787	$C_{l\delta a}$
0.1056	0.1075	0.1443	0.1443	$C_{l\delta r}$
-0.2381	-0.2426	-0.3282	-0.3282	$C_{l\beta}$
0	0	0	0	C_{m0}
-1.6945	-0.7247	-2.0426	-2.0426	C_{mq}
-0.4583	-0.3731	-0.5528	-0.5528	$C_{m\delta e}$
-0.1288	-0.1238	-0.1552	-0.1552	$C_{m\alpha}$
-0.0021	-0.0021	-0.0029	-0.0029	C_{np}

-0.0354	-0.0351	-0.0296	-0.0299	C_{nr}
0.0018	0.0018	0.0016	0.0016	$C_{n\delta a}$
-0.0478	-0.0478	-0.0489	-0.0489	$C_{n\delta r}$
0.0658	0.0659	0.0679	0.0679	$C_{n\beta}$
0.0001* [185,275,934.4]	0.0001* [185,275, 934.2]	0.0001* [185,273,930.9]	0.0001* [185,271,929.3]	C_D

The results clearly show that our approach is accurate and the plots above present also the improvement ratio. The final results demonostrate that the order of the improvement in the parameter estimation results is 10.9% of the parameters and the error between real and estimated parameter values decreased around 6.3%.

Conclusions

This work showed the importance of the mathematical representation of extended linear regression algorithm for aircraft system model. The idea is to represent the equation from another view to get the model to more accurately represent reality. The estimation of parameters is improved, this can be used later in aircarft with unkown parameters to use them in other topics of aircraft applications. The improvement ratio declared that the Extended Linear Regression with FireFly method is better than normal Linear Regression and this can be used in many problems of model analysis in different fields of applications.

References

[1] M. Mohamed: System identification of flexible aircraft in frequency domain, Aircraft Engineering and Aerospace Technology, Vol. 89, No. 6, pp. 826-834, 2017

[2] M. K. Samal, A. Singhal, and A. K. Ghosh: Estimation of equivalent aerodynamic parameters of an aeroelastic aircraft using neural network, Journal of the Institution of Engineers (India), Aerospace Engineering Division, Vol. 90, pp. 3-9, 2009

[3] H. N. Al-sudany, B. Lantos: Comparison of Adaptive Fuzzy EKF and Adaptive Fuzzy UKF for State Estimation of UAVs Using Sensor Fusion. Periodica Polytechnica Electrical Engineering and Computer Science, Vol. 66, No. 3, pp. 215-266, 2022

[4] H. N. Al-sudany, B. Lantos: Prediction of the Navigation Angles Using Random Forest Algorithm and Real Flight Data of UAVs, IEEE 20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY), pp. 000097-000102, 2022

[5] M. Mohamed, V. Dongare: Aircraft Aerodynamic Parameter Estimation from Flight Data Using Neural Partial Differentiation, Springer Nature, 2021

[6] J. A. Grauer, M. J. Boucher: Aircraft system identification from multisine inputs and frequency responses. Journal of Guidance, Control, and Dynamics, Vol. 43, No. 12, pp. 2391-2398, 2020

[7] L. I. N. Chun-Yueh: Fuzzy AHP-based prioritization of the optimal alternative of external equity financing for start-ups of lending company in uncertain environment. Sci. Technol, Vol. 25, No. 2, pp.133-149, 2022

[8] G. G. Seo, Y. Kim, S. Saderla: Kalman-filter based online system identification of fixed-wing aircraft in upset condition. Aerospace Science and Technology, Vol. 89, pp. 307-317, 2019

[9] Z. Bodó, B. Lantos: Modeling and control of fixed wing UAVs. IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 332-337, 2019

[10] R. F. Stengel: Some effects of parameter variations on the lateral-directional stability of aircraft. Journal of Guidance and Control, Vol. 3, No. 2, pp. 124-131, 1980

[11] A. Assad, W. Khalaf, I. Chouaib: Radial basis function Kalman filter for attitude estimation in GPS-denied environment. IET Radar, Sonar & Navigation, Vol. 14, No. 5, pp. 736-746, 2020

[12] G. Rigatos, P. Siano, D. Selisteanu, R. E. Precup: Nonlinear optimal control of oxygen and carbon dioxide levels in blood. Intelligent Industrial Systems, Vol. 3, pp. 61-75, 2017

[13] D. Singh, A. Shukla: Manifold optimization with MMSE hybrid precoder for Mm-Wave massive MIMO communication. Sci. Technol, Vol. 25, No. 1, pp. 36-46, 2022

[14] C. A. Bojan-Dragos, R. E. Precup, S. Preitl, R. C. Roman, E. L. Hedrea, A. I. Szedlak-Stinean: GWO-based optimal tuning of type-1 and type-2 fuzzy controllers for electromagnetic actuated clutch systems. IFAC-PapersOnLine, Vol. 54, No. 4, pp. 189-194, 2021

[15] Blaauw, Deon: Flight control system for a variable stability blended-wing-body unmanned aerial vehicle. Diss. Stellenbosch: University of Stellenbosch, 2009

[16] Broughton, B. A., and R. Heise. :Optimisation of the Sekwa blended-wing-Body research UAV,2008

[17] Z. Bodó: Modern control methods for unmanned aerial and ground vehicles, (Doctoral dissertation, Budapest, Hungary), 2021

[18] R. H. Byrd, J. C. Gilbert, J. Nocedal: A trust region method based on interior point techniques for nonlinear programming. Mathematical programming, Vol. 89, No. pp. 149-185, 2000

[19] Z. Drezner: On minimax optimization problems. Mathematical programming, Vol. 22, pp. 227-230, 1982

[20] M. Milovančević, D. Milčić, B. Andjelkovic, L. Vračar: Train Driving Parameters Optimization to Maximize Efficiency and Fuel Consumption. Acta Polytechnica Hungarica, Vol. 19, No. 3, pp. 143-154, 2022

[21] R. E. Precup, E. L. Hedrea, R. C. Roman, E. M. Petriu, A. I. Szedlak-Stinean, C. A. Bojan-Dragos: Experiment-based approach to teach optimization techniques. IEEE Transactions on Education, Vol. 64, No. 2, pp. 88-94, 2020

[22] Tian Z: Backtracking search optimization algorithm-based least square support vector machine and its applications. Engineering Applications of Artificial Intelligence. 2020 Sep 1;94:103801

[23] B. Lantos, L. Márton: Nonlinear control of vehicles and robots. Springer Science & Business Media, 2010

[24] Johari NF, Zain AM, Noorfa MH, Udin A: Firefly algorithm for optimization problem. Applied Mechanics and Materials. 2013 Dec 12; pp. 512-517

[25] S. K. Sarangi, R. Panda, S. Priyadarshini, A. Sarangi: A new modified firefly algorithm for function optimization. IEEE international conference on electrical, electronics, and optimization techniques (ICEEOT), pp. 2944-2949, 2016

[26] Ezzeldin R, Zelenakova M, Abd-Elhamid HF, Pietrucha-Urbanik K, Elabd S: Hybrid Optimization Algorithms of Firefly with GA and PSO for the Optimal Design of Water Distribution Networks. Water. 2023 May 17;15(10), 1906

The Changing Business Ethics and Etiquette, in Slovakia and Hungary, due to Globalization

Zsuzsanna Tóth, László Józsa, Erika Seres Huszárik

J. Selye University, Bratislavská cesta 3322, 945 01 Komárno, Slovakia,
tothz@ujs.sk; jozsal@ujs.sk; huszarike@ujs.sk

Kim-Shyan Fam

Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary,
kimfam@magscholar.com

Mohamad-Noor Salehhuddin Sharipudin

Faculty of Modern Languages and Communication, Universiti Putra Malaysia,
43400 Serdang, Selangor, Malaysia, salehhuddin@upm.edu.my

Suffian Hadi Ayub

School of Communication and Media, College of Computing, Informatics and
Media, Universiti Teknologi MARA (UiTM) 40450 Shah Alam, Selangor Darul
Ehsan, Malaysia, suffianhadi@uitm.edu.my

Abstract: There are significant differences between the culture, ideology, and values of different nations, so it is not surprising that there are differences in business ethics and etiquette. Therefore, it is essential when establishing and maintaining business relations that the parties get to know each other's ethnic customs or general international etiquette to approach each other with greater understanding and develop more successful business negotiations and business relationships. This paper aims to illustrate the similarities and differences in etiquette behavior, through the examples of two neighboring countries (Slovakia and Hungary). Based on the results of our primary, questionnaire-based, cross-national research, we conclude that companies in both Hungary and Slovakia consider that the manifestations of business ethics in the countries have improved over the last ten years due to globalization.

Keywords: globalization; business ethics; business etiquette; business relations

1 Introduction

At the beginning of the 21st Century, it must seem cliché to say that "we live in a global environment". As the global economy deepens and the importance of internationalization becomes apparent to businesses, their relationships with foreign companies and their agents will increase. Cross-cultural negotiations are often integral to the process, whether buying or selling. The trend towards globalization has led different cultures to enter the world stage. In business, culture is seen as an essential contributor to success.

Organizations, employees and teams are increasingly operating in multicultural, multinational environments. More and more companies are exporting work, not just goods, worldwide. Physical distance or time differences are no longer barriers to foreign investment. Local companies need foreign investment to compete with global companies. The acceleration of global business development is accompanied by a growing interest in cross-cultural management research, as evidenced by the increasing presence of international studies in leading journals.

1.1 The Impact of Globalization on National Cultures

t is generally accepted that the American Theodore Levitt first used the concept of globalization in an article in the Harvard Business Review. He used the term to describe the merging of markets for individual products produced by transnational corporations [9].

Globalization has a powerful impact on our society [15], dividing but also unifying [50]. Sparke [47] defined the concept as a combination of countries' political, social and economic concepts. He drew particular attention to globalization's positive and negative effects on companies. Ten Brink [49] also pointed out that significant firms in some countries have shown an apparent openness to developing and deepening international cooperation. However, companies' motivation to support collaboration is not only economic in origin but also reveals political and cultural aspects behind these processes. The business's costs remain significant in the company's profitability and competitiveness. Rowley and Warner [45] pointed out that the business environment has changed significantly due to globalization. They argue that firms have had to rethink their production and supply chain strategies. Dunning's [14] research also confirmed Rowley and Warner's claims and pointed out that reducing operating costs is a form of international pressure on firms.

At the turn of the 21st Century, globalization has been ascribed a more prominent, comprehensive role than ever before [44]. According to Kobrin [29] and Witt [53], globalization in international business has reached unprecedented proportions. Globalization has triggered the integration of markets and trade [20]. In addition to the political changes it has brought, the process has also brought technological innovations and novel management practices, which have increased knowledge transfer, reformed production times, and increased the use of goods and services, which have also increased international economic cooperation [40]. In a positive sense, as outlined above, one of the main benefits of globalization is the improvement of countries' development indicators [20].

However, the negative and positive sides of globalization in business have been explored by many authors. For example, the share of inequalities between countries has increased. Several authors have also addressed the issue of how physical workers leave their home country for countries offering higher wages [21] [30] [32] [33].

Nowadays, after COVID-19, business leaders have to prepare for the world's leading countries to repatriate part of their activities to their home countries [22]. Trade wars between the world's major powers in the East and West may slow down the pace of globalization [5] [27]. Currently, business actors are trying to find their way in this fast-changing environment. Some globalization researchers have already predicted the end of the process [38]. Charron [7] argued that the bureaucracy and corruption in these countries determine economic divergence in European countries due to globalization. At the same time, as globalization slows down, we can observe technological innovations such as 3D printing that could fundamentally affect global value chains by changing the role of sourcing, production, and further activities in the supply chain [26]. The new international environment is gradually making it more challenging for individuals, economics and societies to operate in the information age. In such an environment, there will be new frameworks and rules of the game according to which the new civilization will operate [28]. Global production plants must have the ability to react quickly, which can be crucial to their survival [16].

The new tasks of business ethics thus become a critical reflection of the realities of the new global environment and aim to regulate entrepreneurial activity through moral norms. Therefore, to support all social issues directly or indirectly related to and support sustainable life on Earth. Business ethics can help mitigate negative phenomena and consequences in a new global environment.

1.2 Business Ethics and Etiquette

Business ethics influences the managers and employees of a company, helping them to make the right and wrong decisions in business situations [10]. In business situations, the question often arises: What should I do? What is the right thing to do? In answering this question, the individual's business principles, values, and emotional intelligence influence daily life and social relationships [34]. Researchers who have studied international business negotiations have primarily examined the impact of culture on negotiations from an ethical perspective [23] [46]. The number of studies focusing on the background of ethical business decisions is increasing yearly [3] [39] [51]. Entrepreneurs in the business sector may often find themselves in a situation where they must answer ethical and moral questions about a decision [55]. Normative ethical rules can help to answer ethical questions related to business [19] [31]. Ethics is nothing more than a set of specific norms of behavior and action [4]. The application of business ethics not only to customers but also to supplier relationships is essential. Ethical business policies, procedures, and equal standards are hallmarks of fair business relationships [42]. Many business actors prefer the stability of long-term business relationships [1] [18].

However, one area that has received less attention so far is business ethics. This area needs to be examined to help businesses cope in a global environment. In their study, Cook and Cook [11] point out that the globalization of the world has made it increasingly important for managers to be aware of different cultures, business etiquette rules and management styles. Their study examines managerial manners and professional behavior in international business. This includes, for example, respectful behavior, how to introduce oneself, business card rules, tone of communication, knowledge of different cultural customs, recognition of position, business gift-giving habits and non-verbal communication. In their study on corporate communication, Reynolds and Valentine [43] found that a positive correlation can be established between the internal communication of companies and the success of their international communication.

Business etiquette in business relationships is based on culturally established protocols, manners, rules of conduct, and a shared code of conduct within a given locality [13]. Business etiquette includes expected manners, acceptable behavior, and the principles that guide such behavior (e.g., courtesy, prudence). Knowledge of business etiquette is part of an individual's "soft skills" that are key to working with employees, customers, business partners, and other stakeholders [52].

Researchers in the field of communication emphasize the vital role of etiquette in the management context of international business and communication situations between people of different cultures, as leadership skills and the multinational business environment require appropriate behavior from managers [36]. In today's rapidly changing and competitive societies, etiquette may be seen as an outdated concept, but it undeniably impacts people's perceptions and decisions in global management. It can always be emphasized that how managers behave towards themselves and their colleagues determines the company's credibility. Alongside effective cross-cultural communication and negotiation skills, knowledge of business etiquette is responsible for managers' success at home and globally. In their recent research, Chaney and Martin [6] agree that leadership behavior, neatness, appearance, and attitude are critical for evaluating corporate performance and overall success in a global workplace. It is widely accepted that etiquette includes specific aspects of behaviors, habits, and non-verbal communication that can influence the effectiveness of a leader's performance [37].

Yu [54], in a comprehensive assessment of current trends in the global business world, point out that the vast expansion of multinational businesses over the past decade has made it vital to learn more about different cultures and human behaviors to reduce the chances of leadership failure.

1.3 Business Ethics and Etiquette in Slovakia and Hungary

Business ethics is not a long tradition in the Slovak Republic. With the change in the political climate and economic conditions after 1989, the space for developing business ethics opened up. The acquisition of foreign experience was an essential factor. Large international companies opened their businesses in Slovakia and introduced elaborate ethical rules (e.g. the Code of Ethics). In addition, banks, insurance companies, and other service-oriented firms also had direct contact with their clients and contributed to this process. The exchange of information between Slovakia and the international arena, the common European market, globalization, and the merging of business conditions in national and international markets, resulting from interdependence, accelerate and increase the role of business ethics.

Slovaks are more temperamental and emotional than Czechs in business meetings and are more impulsive when solving problems. Slovaks are easy to start negotiations with and remain open and adaptable throughout the process. Most Slovak negotiators are friendly, open, and spontaneous. They value personal contact and often rely on recommendations and references from people they know. In general, Slovak negotiators could be more attentive to etiquette and formalities. However, if a woman is present, she is always treated with courtesy and gallantry. Some researchers say Slovaks are somewhat insecure and very optimistic during negotiations. Those who come prepared with realistic and clear objectives have an advantage in business negotiations. The Slovak partner will appreciate this attitude, making closing the deal more manageable. The introductory phase can be skipped during the first meeting and go straight to business. However, closing the deal immediately at the first or second meeting is not advisable, as Slovaks need more time to consider all the commitments. If pressed, they may even walk away from the deal [24].

A Hungarian survey of 1,300 managers from 325 companies in 1996 [8] shows a mixed picture of ethical behavior. According to the authors, the institutional management of business ethics was in its infancy in the mid-1990s, as only slightly more than 10% of companies had a code of conduct, for example. The companies with the most ethical behavior sought to compete in developed country markets. However, it was found that the ethical behavior of 'Western' companies operating in domestic markets was not outstanding in those years. Later, Szegedi [48] argues that the responsibility of Hungarian firms is essential toward their customers and owners. He adds that ethical sensitivity is expected to improve [41].

András et al. also found that CSR is an existing concept among Hungarian companies today and is often not a matter of money [2]. To put it a bit simplistically, CSR involves ethical behavior, which should manifest not only in externalities and spectacular actions but also in conscious strategies. Győri [25], based on a review of various case studies, argues that market, governance, and ethical responsibility systems should complement each other in business. In our review of the Hungarian literature, we found a cultural specificity in that in Hungary, authors often focus on unethical business conduct and, within that, corruption [17] when discussing business ethics and fair conduct [12].

2 The Aim of the Study and the Applied Research Method

We aimed to synthesize and summarize the academic findings on the characteristics of business behavior, with particular emphasis on the different cultures, and to examine the behavior of companies operating in Slovakia and Hungary from the perspective of business ethics and etiquette based on theoretical foundations.

To achieve our goal, we have defined our main research question:

In today's globalized world, are there any differences, identities, or peculiarities in Slovak and Hungarian business ethics and etiquette?

Our hypothesis related to our research question is the following:

H1: Globalization has improved business and business ethics in both countries over the last ten years.

The empirical research was based on an online questionnaire survey attended by representatives of companies in Hungary and Slovakia. The questionnaire investigates the participants' behavior; thus, it also has the disadvantages of the survey method, i.e. it is not sure that the respondents are willing and able to provide the exact information to the question asked. In addition, there may also be disadvantages in answering personal and sensitive questions [35]. The questionnaire survey was followed by data cleaning and evaluation. To test our hypotheses, we chose a single descriptive analysis as the primary research method, all the more so as our data were obtained in a single session on a single sample [35].

Our empirical research was part of a more extensive international study. Within the framework of an international project, Marketing in Asia Group, New Zealand, Slovakia, and Hungary were studied in terms of business communication, ethics, and etiquette. The questionnaire was developed and tested with the participants by Professor Kim-Shyan Fam and Dr James E Richard, the research leaders from Victoria University, Wellington. The questionnaire provided was translated from English into Hungarian and Slovak and then back into English by another party to ensure that the cross-cultural comparison was an accurate translation.

In order to collect the data, we needed to create a database of companies operating in Slovakia and Hungary. The size of the companies and the industry was not decisive. The address list compiled using the collection pages contained contact details of 938 companies. Our online questionnaire was sent out in the spring of 2018. Due to invalid, non-functional email addresses, we received 22 replies. We used a random sampling method, the snowball method, and simultaneously, as compiling the database, we collected known company managers to whom we forwarded our online questionnaire and asked them to forward it to company managers with whom they were in contact. After data cleansing, we had 257 completed questionnaires, with a response rate of 28.05% over three months. This result leads us to conclude that respondents are unlikely to be willing to participate in surveys of this kind. 103 respondents from Hungary and 154 from Slovakia participated in our survey. Completed questionnaires were coded, and the values obtained were recorded in the SPSS statistical program table. The evaluation was also carried out using this program: univariate, bivariate, and multivariate analyses were performed.

To test our research hypotheses, we first used univariate analyses using variance indicators (standard deviation) on the one hand and positional indicators (mean, mode) on the other. To search for deeper correlations, we chose the methods of bivariate and multivariate analyses. Analysis of variance examines the effect of one or more factors on one or more factors. Analysis of variance involves comparing the means of more than two sets of variables on a sample basis. This is why it is called a generalization of the two-sample t-test. To decide the null hypothesis, we use the squares of the variances, hence the name analysis of variance. It can be used for problems where the value of a probability variable depends on one or more systematic effects and chance. In correlation analysis, we examine the relationship between two variables measured on a metric scale. The analysis examines the Pearson correlation coefficient (r) value, ranging from -1 to +1. The coefficient sign indicates the relationship's direction, while its absolute value indicates the strength of the relationship. It is important to stress, however, that a value of r=0 does not automatically mean that there is no relationship between the two variables, but only that there is no linear relationship. It is also important to note that correlation analysis is unsuitable for finding causal relationships, as it does not distinguish between dependent and independent variables. In our analysis, we followed the steps of correlation analysis, first filtering outliers from the data table and then running the analyses to interpret the results.

3 Our Research Results

As a first step in presenting our research results, we compare the perspectives of Slovak and Hungarian respondents on business ethics using univariate analyses.

Therefore, business etiquette is a set of behaviors often maintained by custom and enforced by members of society to provide an environment where members feel comfortable and secure in their social and business relationships. Respondents with an insight into the business world considered different elements of business ethics to be of varying importance.

We can now use statistical measures if we consider the responses as quantitative scales. The importance of the given items was rated on a seven-point scale. Personal appearance and professional behavior were considered the essential elements by our Slovakian respondents (both with a mean of 5.92), but punctuality (5.84) and respect (5.75) were also highlighted in the business etiquette question. Respondents rated cultural sensitivity and giving gifts less highly by respondents (mean of 4.74 and 4.94, respectively).

The importance of the elements of business etiquette was estimated with relatively low variance by the respondents. The variance was highest for reciprocity (1.49).

Table 1

Importance of elements of business etiquette with statistical measures (N=151-154)

	Mean	Deviation	Mode
Communication	5.38	1.17	very important
Cultural sensitivity	4.74	1.20	slightly important
Gift-giving	4.94	1.27	slightly important
Appearance	5.92	1.08	very important
Professional behavior	5.92	1.14	very important
Punctuality	5.84	1.12	very important
Respect	5.75	1.09	very important
Social behavior	5.66	1.07	very important
Trust	5.61	1.35	particularly important
Reciprocity	5.33	1.49	very important

Source: Author´s editing

The responses thus show that all the elements of business etiquette are considered very important by the "median" businessperson, except cultural sensitivity and gift-giving, which are considered somewhat important.

Moving on to Hungary's business etiquette practices, we find that the most crucial element for Hungarian companies was respect (mean: 6.26). Respondents were less likely to consider cultural sensitivity and giving gifts (mean: 5.04 and 4.40).

Respondents estimated the importance of the elements of business etiquette with relatively low variance. The standard deviation was highest for the perception of giving gifts (1.21), i.e., the most minor consensus among the elements listed.

Table 2

Importance of elements of business etiquette with statistical measures (N=85-87)

	Mean	Deviation	Mode
Communication	6.07	0.96	very important
Cultural sensitivity	5.04	1.17	slightly important
Gift-giving	4.40	1.21	slightly important
Appearance	5.77	1.03	very important
Professional behavior	6.00	1.12	very important
Punctuality	6.11	1.08	particularly important
Respect	6.26	0.90	very important
Social behavior	5.83	0.99	very important
Trust	6.23	0.83	very important
Reciprocity	5.85	0.89	very important

Source: Author´s editing

The responses thus show that all the elements of business etiquette are considered very important by the "median" businessperson, except cultural sensitivity and gift-giving, which are considered somewhat important.

As a first step in the multivariate analysis of the elements of business etiquette, the co-movement of responses to each characteristic is examined. For example, it is possible to identify whether those who consider one element of business etiquette, such as respect, punctuality, or trust, to be necessary hold similar views on other attributes. The result of the correlation analysis is most easily illustrated by a correlation table showing all possible correlation coefficients.

The correlation table shows that social behavior and respect move together the most among the elements of business etiquette (correlation coefficient: 0.710). The correlation analysis also highlights the co-movement between punctuality and professionalism (correlation coefficient: 0.622). Interestingly, reciprocity shows no significant correlation with any other characteristic in the sample except trust.

Elements of business etiquette can be examined to see if there are significant differences between men's and women's perceptions and priorities. In the gender breakdown, the mean of the responses differed little in many cases.

Table 3

Correlation between elements of business etiquette (N=151-154)

	Communication	Cultural sensitivity	Gift giving	Appearance	Professionalism	Punctuality	Respect	Social behavior	Trust	Reciprocity
Communication	1									
Cultural sensitivity	0.272	1								
Gift-giving	0.113	0.233	1							
Appearance	0.325	0.124	0.417	1						
Professionalism	0.311	0.300	0.167	0.435	1					
Punctuality	0.230	0.104	0.045	0.433	0.622	1				
Respect	0.268	0.155	0.150	0.378	0.299	0.330	1			
Social behavior	0.199	0.234	0.143	0.365	0.367	0.356	0.710	1		
Trust	0.352	0.242	0.117	0.306	0.145	0.237	0.256	0.172	1	
Reciprocity	0.220	0.116	-0.006	-0.009	-0.057	-0.005	0.065	0.024	0.598	1

Source: Author´s editing

Overall, women consider the listed elements of business etiquette somewhat more critical. The difference between men's and women's ratings on a seven-point scale can be statistically tested using a t-test. In this way, it can be determined whether women or men consider a characteristic statistically significantly more or less important than another characteristic in the list of elements of business etiquette.

Table 4

Gender differences in perceptions of elements of business etiquette (N=151-154)

	Man	Woman	Difference	p-value of t-statistic
Communication	4.98	5.59	-0.61	0.004
Cultural sensitivity	4.40	4.95	-0.55	0.011
Gift-giving	5.11	4.88	0.22	0.312
Appearance	5.83	5.91	-0.08	0.687
Professional behavior	5.81	5.86	-0.05	0.810
Punctuality	5.77	5.86	-0.09	0.647
Respect	5.47	5.86	-0.39	0.047
Social behavior	5.47	5.71	-0.24	0.203
Trust	5.21	5.79	-0.58	0.013
Reciprocity	5.00	5.51	-0.51	0.044

Source: Author´s editing

The t-test results show no significant gender difference in the perception of the importance of gift-giving, personal appearance, professionalism, punctuality, and social behavior. The most considerable difference was in the perception of communication, which was significantly more critical among female participants than men. On the other hand, respect, trust, reciprocity, and cultural sensitivity were already rated significantly more important by the women surveyed when the data were tested at the 5% significance level.

Table 5

Perception of elements of business etiquette based on business experience (N=151-154)

	under 1 year	1-5 years	6-10 years	11-20 years	more than 20 years	p-value of ANOVA
Communication	5.89	5.26	5.00	5.57	5.35	0.413
Cultural sensitivity	4.56	4.48	4.75	4.86	4.79	0.790
Gift-giving	4.11	4.87	4.00	4.46	5.34	0.000
Appearance	5.11	6.00	4.92	5.64	6.22	0.000
Professional behavior	4.56	5.87	5.50	5.50	6.28	0.000
Punctuality	5.22	5.48	5.33	5.89	6.07	0.023
Respect	5.56	6.26	5.25	5.57	5.77	0.070
Social behavior	5.00	5.74	5.67	5.54	5.76	0.339
Trust	5.44	5.55	5.17	6.14	5.49	0.151
Reciprocity	5.67	5.55	5.17	5.88	4.95	0.027

Source: Author´s editing

When assessing the elements of business etiquette, it is worth looking at how the importance of specific attributes changes throughout the business experience. This can be tested using the analysis of variance, which shows whether statistically significant differences exist between experience groups.

The variance analysis results show no significant differences in the perceptions of trust, social behavior, cultural sensitivity, and communication between people with different business experiences. There is heterogeneity in the perceptions of respect at the 10 percent significance level. At the 5 percent significance level, the perceptions of reciprocity and punctuality differ between the groups studied for gift-giving, personal appearance, and professionalism.

Our hypothesis for the study can be compared with the empirical observations, i.e. the results of the questionnaire survey, using statistical methods and analysis.

The research hypothesis is as follows:

H1: Due to globalization, market players have improved business ethics in Hungary and Slovakia over the last ten years.

Slovakia and Hungary joined the European Union in 2004, accelerating both countries' already significant globalization processes. Accession to the common market meant that companies in Slovakia and Hungary no longer focused on their home markets but had the opportunity to tap into E.U. markets and cooperate at the European level. This has also meant that business people in Slovakia and Hungary have become familiar with other business cultures through their various interactions, impacting business ethics and business ethics skills.

Globalization has been further facilitated by the ubiquity of the internet and the rapid and widespread spread of info-communications tools. Fresh graduates were already exposed to this process, so their business ethics behavior and business etiquette skills have likely remained the same from previous generations.

Respondents were asked whether business ethics and conduct had improved in the country ten years before their survey. They were also asked whether there had been an improvement in business ethics among graduates.

A handy tool for multivariate analysis is the correlation calculation, or correlation table of results, which shows the extent to which perceptions of improvements in business ethics and etiquette and graduates' skills and attitudes coincide.

Table 6

Correlation coefficients of perceived improvements in business ethics and etiquette among general and graduate students (N=134)

	Business ethical behavior has improved in the country over the past 10 years.	Business ethics have improved in the country over the last 10 years.	Business ethical behavior of graduates has improved in the country over the last 10 years.	The business ethics skills of graduates have improved in the country over the last 10 years.
Business ethical behavior has improved in the country over the past 10 years.	1			
Business ethics have improved in the country over the last 10 years.	0.738	1		
Business ethical behavior of graduates has improved in the country over the last 10 years.	0.438	0.512	1	
The business ethics skills of graduates have improved in the country over the last 10 years.	0.352	0.411	0.713	1

Source: Author´s editing

Unsurprisingly, the correlation matrix of the four questions contains only positive elements. The results show that those who think ethical behavior has improved in the country over the last ten years are likely to hold similar views on business ethics (correlation coefficient: 0.738). Graduates' views on ethical business behavior and business ethics skills also show a relatively high correlation (0.713).

In the rest of our analysis, we compare our respondents' views on the changes in Slovakia and Hungary. According to Slovakian respondents, business ethics in the country have improved slightly over the last ten years, with an average score of 4.93 on a seven-point scale. The average response for business ethics was 4.57, representing a slight improvement. For recent graduates, the average response for improvement in business ethics behavior was 4.39, while the average response for improvement in business ethics skills was 4.40. Both correspond to a slight improvement in these attributes and skills in the country over the previous ten years. In addition to the average, examining the median scores may also be interesting. According to the median businessperson, business ethics and etiquette have improved slightly in Slovakia over the last ten years. In contrast, the business ethics behavior of graduates has not changed according to the median opinion. The perception of business ethics skills of those entering the Slovakian labor market is unclear, as the median is just between "about the same" and "slightly improved".

Hungarian respondents to the questionnaire rated both the improvement in business ethics behavior and the development of business ethics skills at 4.48 on a scale of 7, which is broadly equivalent to a slight improvement in business ethics and etiquette in the country over the past ten years. Slightly higher scores were recorded for recent graduates' perceptions of business ethics and etiquette, with an average of 4.51 for the former and 4.56 for the latter. These scores represent a slight improvement. In the case of Hungary, it is also worth considering the medians in addition to the averages. The "median businessperson" perceives a slight improvement in business ethics behavior over the previous ten years and a similar perception of business ethics skills in the country. However, the median for graduates' business ethics and etiquette is that this area has remained the same.

The results of our questionnaire survey show that respondents in both Hungary and Slovakia experienced a slight improvement in business ethics behavior and business ethics skills. This slight improvement is evident in the overall business culture in the two countries: there is a slight improvement in graduates' business ethics behavior and business etiquette skills.

Table 7

Differences between Slovakia and Hungary in perceptions of improvements in business ethics and etiquette (N=209)

	Slovakia	Hungary	difference	p-value of t-statistic
Business ethical behavior has improved in the country over the past 10 years.	4.93	4.48	0.45	0.012
Business ethics have improved in the country over the last 10 years.	4.57	4.48	0.09	0.652

Business ethical behavior of graduates has improved in the country over the last 10 years.	4.39	4.51	-0.12	0.512
The business ethics skills of graduates have improved in the country over the last 10 years.	4.40	4.56	-0.16	0.343

Source: Author´s editing

The results show a similar trend in Slovakia and Hungary. However, similarities and differences between the two countries can be formally tested using a t-test. The t-test results indicate that only the change in general business ethical behavior shows a significant difference between the two countries. In contrast, the results are not statistically different for the other questions. The improvement in business ethics was also slightly more statistically significant in Slovakia than in Hungary. Based on the analyses performed, we can conclude that hypothesis H1 of our research can be accepted, so our thesis is as follows:

T1: Due to globalization, market players have seen a slight improvement in business and business ethics in Hungary and Slovakia over the past ten years.

Summary and Conclusions

The aim of our research was, on one hand, to summarize and synthesize the scientific results dealing with business ethics and on the other hand, generate new scientific results, by using empirical research to investigate the specificities of business ethics in Slovak and Hungarian companies. The study and research of business ethics in both countries have been given little emphasis, with few studies and research being found almost in isolation. There needs to be adequate literature on business ethics in Slovakia and Hungary, as well as comparative literature on business ethics in the two countries. Our questionnaire survey also proves that it is worthwhile to research the peculiarities in business ethics in both countries, as it is undoubtedly helpful to know and use the research results to choose a strategy. The results of our research can be helpful for negotiators who can gain insight into the evolution of Slovak-Hungarian cultural norms in negotiation behaviors and better identify factors that improve negotiation outcomes, which can serve as a valuable tool for negotiators to facilitate agreements between the two cultures or possibly other cultures. In addition, they can use our findings to plan their approach and use of strategy and thus, build more strategic, successful relationships for the long term.

References

[1] AKROUT, H., DIALLO, M. F., AKROUT, W. - CHANDON, J.-L. (2016) Affective trust in buyer-seller relationships: A two-dimensional scale. *The Journal of Business & Industrial Marketing*, 31(2), 260-273

[2] ANDRÁS, I. – RAJCSÁNYI-MOLNÁR, M. – FÜREDI, G. „A vállalatilag felelős vállalat – A CSR és a cafeteria-metszet értelmezési lehetőségei a gyakorlatban” In Metamorfózis – Glokális dilemmák három tételben (Szerk.: András István – Rajcsányi-Molnár Mónika) Új Mandátum Könyvkiadó. 2013, pp. 127-139

[3] BAZERMAN, M. H. - SEZER, O. (2016) Bounded awareness: Implications for ethical decision making. Organizational Behavior and Human Decision Processes, 136, 95-105

[4] BLUMENTHAL, S.-F. (2011) Wissenschaftsbezogene Ethikinitiativen supra-/nationaler Organisationen im europäischen Forschungsraum. In *E. Jantscher, L. Neuhold, & B. Pelzl (Eds.). Ethik in Forschung und Technik: Annäherungen*. Bölau Verlag: Wien

[5] BREWSTER, R. (2019) WTO dispute settlement: Can we go back again? UJIL Unbound, 113, 61-66, Bruno, G. S. F

[6] CHANEY, L. H. - MARTIN, J. S. Intercultural business communication (4th ed.) Upper Saddle River, NJ: Pearson Prentice Hall, 2011

[7] CHARRON, N. (2013) Diverging cohesion? Globalisation, state capacity and regional inequalities within and across European countries. European Urban and Regional Studies, 23(3), 355-373

[8] CHIKÁN, A. Vállalatok és funkciók integrálója - Folyamatjellegű irányítás - alprojekt záró tanulmánya. Budapesti Közgazdaságtudományi Egyetem, Vállalatgazdaságtan Intézet, 1997

[9] COLLINS, M. (2018) The Pros and Cons of Globalization, Saving American Manufacturing. Forbes Media LLC.

[10] CRANE, A., MATTEN, D., GLOZER, S. - SPENCE, L. (2019) Business ethics: Managing corporate citizenship and sustainability in the age of globalization (Fifth edition. ed.). Oxford: Oxford University Press.

[11] COOK, R. A. - COOK, G. O. (2011) Guide to business etiquette (2nd ed.) Upper Saddle River, NJ: Prentice Hall publishing.

[12] CZIBIK, Á. – HAJDU, M. – MAKÓ, Á. – TÓTH, I. J. – VÁRHALMI, Z. „Integritás és korrupciós kockázatok a magyar vállalati szektorban”. MKIK Gazdaság- és Vállalkozáskutató Intézet, Budapest, 2011

[13] DEALE, C. S., & Lee, S.-H. (2019) An exploratory study of hospitality and tourism stakeholders’ perceptions of professional etiquette. Journal of Hospitality & Tourism Education, 1-14

[14] DUNNING, J. H. (2014) The Globalisation of Business (Routledge Revivals): The Challenge of the 1990s. Routledge

[15] ERIKSEN, T. H. (2007) Globalization: The Key Concepts, 1st ed., Berg Publishers, Oxford

[16] ERIKSSON, T., NUMMELA, N. - SAARENKETO, A. (2014). Dynamic capability in a small global factory. *International Business Review*, 23, 169-180

[17] FAZEKAS, M. – TÓTH, I. J. – LAWRENCE, P. K. „Anatomy of grand corruption: A composite corruption risk index based on objective data". Corruption Research Centre, Budapest, 2013

[18] FONFARA, K., RATAJCZAK-MROZEK, M. - LESZCZYŃSKI, G. (2018) Change in business relationships and networks: Concepts and business reality. *Industrial Marketing Management*, 70, 1-4

[19] FRANKENA, W. K. (2016) In *N. Hoerster (Ed.). Ethik: Eine Analytische Einführung* (6 ed.) Wiesbaden: Springer VS

[20] GOLDIN, I. - REINERT, K. (2007) „Globalization for development: trade", Finance, Aid, Migration and Policy, The World Bank and Palgrave Macmillan, Washington, DC.

[21] GÖRG, H. - GÖRLICH, D. (2015) „Offshoring, wages and job security of temporary workers", *Review of World Economics*, Vol. 151, No. 3, pp. 533-554

[22] GRAPPI, S., ROMANI, S. - BAGOZZI, R. P. (2020) Consumer reshoring sentiment and animosity: Expanding our understanding of market responses to reshoring. *Management International Review*, 60(1), 69-95

[23] GROVEWELL (2005) „Introduction to the GLOBE research project on leadership worldwide", available at: www.grovewell.com/wp-content/uploads/pub-GLOBE-intro.pdf

[24] GULLOVÁ, S. Mezinárodní obchodní a diplomatickýprotokol, Grada Publishing, Praha, 2010

[25] GYŐRI, ZS. „Első- és másodfajú etikai kudarcok". 2012, Vezetéstudomány XLIII. ÉVF. 10. SZÁM / ISSN 0133-0179. 56-63. o.

[26] HANNIBAL, M. - KNIGHT, G. (2018). Additive manufacturing and the global factory: Disruptive technologies and the location of international business. *International Business Review*, 27, 1116-1127

[27] HODA, A. (2019). Where is U. S. trade policy headed under the Trump administration? In R. Kathuria, & P. Kukrega (Eds.), 20 years of G20 (pp. 81–92). Singapore: Springer

[28] KLINEC, I., PAUHOFOVÁ, I. - STANĚK, P. (2009) Nové globálne prostredie, a zmena parametrov rozdeľovania bohatstva v 21. storočí. Munkadokumentumok 21 [online] [cit. 2016/01/17] ISSN 1337-5598. [cit. 2019.02.10.] Forrás: http://ekonom.sav.sk/uploads/journals/WP20.pdf

[29] KOBRIN, S. J. (2017) „Bricks and mortar in a borderless world: globalization, the backlash, and the multinational enterprise", *Global Strategy Journal*, Vol. 7, No. 2, pp. 159-171

[30] KORCSMÁROS, E., MACHOVÁ, R. (2021) Challenges of burnout prevention in Slovak SMEs– focus on optimal employment In: *Acta Polytechnica*, Vol. 18, No. 2 (2021), pp. 87-104

[31] LACZNIAK, G. R. - MURPHY, P. E. (2019) The role of normative marketing ethics. *Journal of Business Research*, 95, 401-407

[32] LAFFINEUR, C. - MOUHOUD, E. M. (2015) „The jobs at risk from globalization: the French case", *Review of World Economics*, Vol. 151, No. 3, pp. 477-531

[33] LEE, H. - LEE, J. (2015) „The impact of offshoring on temporary workers: evidence on wages from South Korea", *Review of World Economics*, Vol. 151, No. 3, pp. 555-587

[34] MACHOVA, R., ZSIGMOND, T., LAZÁNYI, K., BENCSIK, A. Generations and Emotional Intelligence A Pilot Study. Acta Polytechnica Hungarica. 2020, 17 (5): 229-247

[35] MALHOTRA, N. K. Marketingkutatás, Akadémiai Kiadó Rt., Budapest, 2005

[36] NICOLAU C., ALSATI K. - HERANU A. (2017) Using Business Etiquette Nowadays. Qualitative Research on Business Phoning, Emailing and Meetings. Ovidius University Annals, Economic Sciences Series, Ovidius University of Constantza, Faculty of Economic Sciences. 2017. Vol. 0(2). P. 444-448

[37] OKORO, E. (2012) Cross-Cultural Etiquette and Communication in Global Business: Toward a Strategic Framework for Managing Corporate Expansion. *International Journal of Business and Management,* 7 (16), 130-138

[38] OZTURK, A. - CAVUSGIL, S. T. (2019) Global convergence of consumer spending: Conceptualization and propositions. *International Business Review*, 28, 294-304

[39] PAIK, Y., LEE, J. M. - PAK, Y. S. (2017) Convergence in international business ethics? A comparative study of ethical philosophies, thinking style, and ethical decision-making between US and Korean managers. *Journal of Business Ethics,* 1-17

[40] PERRATON, J. (2019) "The scope and implications of globalisation", In The Handbook of Globalisation, 3rd ed., Edward Elgar Publishing

[41] PIRICZ, N. (2016) The relationship among ethics and conflict management in Hungarian metal and machinery supply chains. *International Journal of Engineering and Management Sciences*, *1*(1), 1-18

[42] QIU, T. (2018) Dependence concentration and fairness perceptions in asymmetric supplier-buyer relationships. *Journal of Marketing Management*, 34(3-4), 395-419

[43] REYNOLDS, S. - VALENTINE, D. (2011) Guide to Cross-Cultural Communication (2nd ed.). Upper Saddle River, NJ: Prentice Hall publishing

[44] ROBERTS, J. - DÖRRENBäCHER, C. (2016) „Renewing the call for critical perspectives on international business", *Critical Perspectives on International Business*, Vol. 12, No. 1

[45] ROWLEY, C. - WARNER, M., (2013) Globalisation and competitiveness: big business in Asia. Routledge

[46] SAORÍN-IBORRA, M. C. - CUBILLO, G. (2016) „Influence of time pressure on the outcome of intercultural commercial negotiations", *Journal of Promotion Management*, Vol. 1, pp. 1-15

[47] SPARKE, M. (2013) Introducing globalisation: Ties, tensions, and uneven integration. John Wiley & Sons

[48] SZEGEDI, Krisztina „A magyar vállalatok etikai érzékenysége". PhD. értekezés. Miskolci Egyetem, 2001

[49] TEN BRINK, T. (2013) Paradoxes of prosperity in China's new capitalism. *Journal of Current Chinese Affairs*, 42(4), 17-44

[50] THE GUARDIAN (2003) „Passion and pessimism", available at: www.theguardian.com/books/2003/apr/05/society

[51] TREVIñO, L. K., DEN NEIUWENBOER, N. A. - KISH-GEPHART, J. (2014) (Un)ethical behavior in organizations. *Annual Review of Psychology*, 65, 635

[52] WEBER, M. R., Finley, D. A., Crawford, A., & Rivera, D. (2009) An exploratory study identifying soft skill competencies in entry-level managers. Tourism and Hospitality Research, 9(4), 353-361, doi:10.1057/thr.2009.22

[53] WITT, M. A. (2019) „De-globalization: theories, predictions, and opportunities for international business research", *Journal of International Business Studies*, Vol. 50, No. 7, pp. 1053-1077

[54] YU, T. - CANNELLA, A. Rivalry between multinational enterprises: An event history approach. Academy of Management Journal, 2007, 50(3): 665-686, http://dx.doi.org/10.5465/AMJ.2007.25527425

[55] ZSIGMOND, T., MACHOVA, R. - KORCSMÁROS, E. (2021) The Ethics and Factors Influencing Employees Working in the Slovak SME Sector, *Acta Polytechnica Hungarica*, 18 (11): 171-190

The Presence of Cybersecurity Competencies in the Engineering Education of Generation Z

Judit Módné Takács, Monika Pogátsnik

Óbuda University, Alba Regia Technical Faculty
Budai u. 45. H-8000 Székesfehérvár, Hungary,
modne.t.judit@amk.uni-obuda.hu, pogatsnik.monika@amk.uni-obuda.hu

Abstract: In the context of 21st Century work in cyberspace, soft skills, and cybersecurity competencies are essential for young engineers in preparation for a career in engineering. The primary objective of this pilot study is to assess the effectiveness and level of security awareness training in the context of digital literacy education, considering the soft skills, educational experiences, and attitudes of the youth. The research uses an innovative methodology. The questionnaire-based quantitative survey is complemented by an alternative qualitative method. In addition to the measurement of attitudes supported by a focus group interview mixed with an imagery association technique, the level of cyber-competence development of the N=130 participating engineering students will be measured by a partially adapted questionnaire. The results of the research will provide insights into the level of awareness, knowledge, and ways of dealing with cyberspace threats among young engineering students, as well as highlight the gaps and strengths of education in terms of skills development. In conclusion, although young engineering students are aware of cyberspace threats, they are not well equipped to deal with them, especially in terms of password habits, security settings, and the use of online social platforms.

Keywords: cybersecurity; education; generation z; competency

1 Introduction

Security awareness combines technical aspects of security with motivations, emotions, behavior, culture, and fears [1]. The term cybersecurity has been coined for the definition of the relationship between cyberspace and security [2]. Nowadays, people need to be prepared and informed about the cyberworld in their personal and professional lives because cyberspace has become an integral part of our lives. Probably due to cost efficiency and changing habits, the communication channel for people in the 21st Century is increasingly shifting to cyberspace [3]. Cyberspace has become part of people's everyday lives, alongside the real, physical world. Like the mechanisms in physical space that have ensured human survival over thousands of years of evolution, our cybersecurity awareness and protection mechanisms in online space are constantly evolving [4].

A key issue in cyberspace is that people often know enough about cyberthreats to answer certain questions, but they don't know how to apply them in practice [5]. The number of data breaches is increasing rapidly, according to recent trends and cybersecurity statistics [6]. Flexible working has become the norm. 84% of employees can work from home at least part-time. In addition, more than half of employees say they would consider a change of job if they could no longer do their job remotely [7]. Cyberattacks are not only aimed at organizations but also at individuals who work from home or who use unfamiliar software to take part in online meetings [8]. And if someone is not aware of the security of their own devices and systems in cyberspace, it is reasonable to assume that the same user will not be aware of the security of their workplace [9]. The level of security awareness at work is not always the same as at home. Whereas the workplace has multiple levels of control, the home has none, and failures in the home, such as social engineering attacks, present a very serious threat to corporate security [10]. Preventive measures, particularly training, and awareness-raising, are essential. Raising security awareness will lead to higher security of information systems.

As awareness is a combination of motivational, emotional, behavioural, and cultural aspects of security, developing cybersecurity awareness is critical in the field of cybersecurity. Adequate cybersecurity awareness helps individuals and organizations prepare for cyberthreats and the risks they pose. The development of soft skills is a priority area for the 21st Century, given their key role in human relations, communication, and problem-solving. To effectively develop cybersecurity awareness, a combination of different skills and practical knowledge is essential. A review of the literature shows that a multifaceted approach and complex teaching methods are key to successful outcomes in the development of cybersecurity awareness.

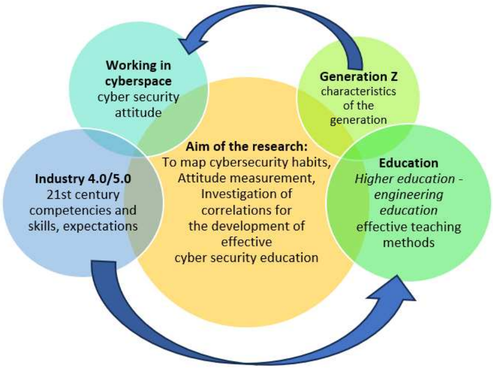

Working in cyberspace
cyber security attitude
Generation Z
characteristics of the generation
Aim of the research:
To map cybersecurity habits, Attitude measurement, Investigation of correlations for the development of effective cyber security education
Industry 4.0/5.0
21st century competencies and skills, expectations
Education
Higher education - engineering education
effective teaching methods

Figure 1

An experimental research strategy for digital education that effectively develops 21st Century skills

Source: Author's construction

With a particular focus on the link between industry and education in the development of digital skills, this research aims (see Figure 1) to explore how the development of cybersecurity awareness, a digital skills cluster, can be integrated into public education from a student's perspective. In addition, the research will analyze the impact of other soft skills that influence the level of cybersecurity awareness, as well as develop a measure to assess the level of cybersecurity awareness among students. The research will also look at the impact of the teaching methods used on the development of cybersecurity awareness. Our aim is to investigate the skills that underpin cybersecurity awareness and to look for correlations between the skills and the effectiveness of the educational methods. The project aims to measure the cybersecurity awareness of students entering university through a pilot, complex, and specific skills evaluation. The research questions are as follows:

Q1: How does developing cybersecurity awareness in public education relate to industry and education in supporting the skills needed for the 21st Century?

Q2: How do soft skills and pedagogical methods are used to influence the development of cybersecurity awareness, and what's the relationship between developing skills, using pedagogical methods, and students' cybersecurity awareness?

Regarding the structure of the article, Chapter 2 analyses the emergence of cybersecurity awareness development in education, analyzing the industry's expectations in the labor market to perform working processes efficiently and safely in cyberspace. The methodology of the experimental research and the innovative combined methods used to measure cybersecurity attitudes are presented in Chapter 3. Chapter 4 details the quantitative survey results and Chapter 5 presents the qualitative focus group interview results using associative methods. Chapter 6 provides a comparison of the research findings with the literature and answers the research questions. Finally, a summary of the research and recommendations for future research directions are provided in Chapter 7.

2 The Importance of the Development of Cybersecurity Awareness in Education in the Light of the Industry

Due to the rapid development of digital transformation in the industry, the development of cybersecurity awareness plays a crucial role in education. The risk and complexity of cyberthreats are increasing with the rise of digitalization and information technologies in the industrial sector. To ensure that students, pupils, and employees are adequately aware of cybersecurity, educational institutions, and professionals need to actively engage in cybersecurity education and training. All disciplines need to be developed with this in mind, not just focusing on training cybersecurity professionals. Furthermore, modern teaching methods and interactive learning approaches are essential to developing effective cybersecurity awareness and soft skills.

2.1 Reviewing Cybersecurity Skills and Attitudes Related to Industry 4.0/5.0

Cybersecurity has become a fundamental issue for the industry in the 21st Century [11] with the emergence of the Industrial Internet of Things (IIoT), where many smart devices are connected to the web, computers, and people. Cybersecurity skills are a complex, diverse, and long list. Relevant cybersecurity skills and attitudes toward digital competence and the labor market are reviewed below. Digital competence is one of the 8 key competencies of the European Union Reference Framework [12]. Security, including cybersecurity, can be seen as part of digital competence, although its importance has been heightened by the pandemic period. In the present era, the stage of human life has shifted towards cyberspace (work, education, economic sector) [10]. Some terms, for example, have recently gained prominence, such as online education, digital curriculum, industry 4.0, and home office. Industry 4.0 stakeholders have identified a range of skills gaps in digital and cybersecurity (data security, cyberattacks, secure communications, ethical hacking, mobile security, and legal issues) [13]. The shortcomings include low levels of applied digital literacy [14], [15], insufficient critical and analytical thinking [16], lack of ability to adapt to new situations, low flexibility, and resilience [17], and weak cybersecurity skills [18], [19]. Concern about occupational health and safety has been expressed by the European Economic and Social Committee [20]. Digital development is associated with a high-level of psychosocial risks in the workplace, such as work overload, inadequate communication, and an increase in work intensity [21]. A decrease in the sense of security at work and an increase in sources of stress are among the negative effects of digitalization.

2.2 The Education-Job Market Relationship and the Role of Cybersecurity

Companies already need employees who can adapt flexibly and continuously to complex needs because of Industry 4.0 and digitization [22]. Those who can successfully adapt and keep up with the very fast pace of digitalization will be able to prevail. Organizational security training will be successful if it is the creation and improvement of security for all employees within the organization [23]. The demand for cybersecurity professionals is growing. Education cannot currently fill the shortage of skilled labor. In addition, an increasing number of studies and surveys indicate that women are under-represented or under-skilled in STEM and cybersecurity occupations. In 2021, only 11% of the global cybersecurity workforce will be women, according to one study [24], down from the current expectation of 50%. A 2020 study [25] reported similar results, with 30% of cybersecurity workers under 30 years old being women, 24% in the 30-38 age group, and only 12-14% in the 39-60 age group. The increased involvement of women in cybersecurity will only increase with the growing demand for cyber-security and the general shortage of professionals. However, research shows that men are more aware of cybersecurity and have more favorable habits than women, [26] [27] so women are considered a kind of risk factor in the field of cybersecurity. The factors influencing the development of the necessary skills must therefore be considered in education. There is a close relationship between security awareness at the level of the organization and security awareness at the level of the individual [25].

Generally, a cybersecurity professional is responsible for the organization's, company's, and employees' security in cyberspace. They are constantly checking that the systems in use are operating securely. Training employees, sharing experiences, and raising security awareness are also part of their job. They need to keep up to date with the latest trends in cyberattacks and train themselves continuously. This is just a short list of the responsibilities of a cybersecurity officer, who often needs a lot of soft skills as well as digital skills, a lifelong learning attitude, motivation, and commitment. A high-level of stress tolerance and problem-solving skills are essential, as is the ability to work in 'emergencies'. Good social skills, the ability to adapt to the needs and attitudes of colleagues, and flexibility are important in carrying out tasks in cooperation with all employees of the organization. The ability to react quickly, analytical thinking, and continuous learning is essential in this field due to the emergence of ever-changing technologies and innovations. These multidisciplinary skills are the keys to the success of cyberspace professionals; they are the "Swiss army knives of the digital world".[28] Future workers must be prepared for the consequences of the widespread introduction of new technologies, in particular robotics, high-levels of automation, and the malicious use of artificial intelligence and machine learning, through education, training, and skills development. This includes aspects such as cyber-attacks, system vulnerabilities, data manipulation, hacking autonomous systems, and the mass collection of personal data, which are paramount to security.

2.3 Presence and Development of Digital Literacy and Security Awareness in Education

Since 2013, the development of security awareness from primary school to higher education must be integrated into the educational process of developing IT and digital competencies, as already stated in the Hungarian Government Decree 1139/2013 (III.21.) on the National Cybersecurity Strategy [29]. Promoting the development of cybersecurity competencies is one of the most important areas of higher education, according to a 2016 research report commissioned by the Swiss

Federal Department of Foreign Affairs (FDFA) [30]. Developing online hygiene and safety competencies in education and protecting minors from cyberabuse and radicalization is a focus of the European Commission's Communication on the Digital Agenda for Education 2018 [31]. 1163/2020. (IV.21.) The Government Decision on the National Security Strategy of Hungary states that the main task concerning cyberspace in the country, in addition to the identification and monitoring of actual risks and threats, is the promotion of security-conscious behavior among users [32]. In this way, cybersecurity education has become one of the defining fields of the 21st Century. Cybersecurity education aims to develop the competencies and skills necessary for effective participation in the online environment. The development of appropriate and targeted cyber security education and awareness can make a significant contribution to the prevention of cyberthreats [33]. By raising awareness, security costs can be reduced and the risks to which users are exposed can be minimized. As behavioral culture plays an important role in the development of security awareness [34], parental and teacher awareness is crucial in the education sector. Cybersecurity can be improved, and cyberthreats can be more effectively prevented by actively involving users in awareness raising [35].

Children have access to information and communication technology (ICT) tools even before the school age. The European Commission has published its new Digital Agenda for Education 2021-2027 [36], which defines two strategic directions. Improving the performance of the digital education ecosystem, including the development of the necessary infrastructure, further developing, and strengthening the digital skills of educators [37], as well as securing educational platforms and facilitating the use of high-quality educational content, are key priorities for the coming years.

Education's focus on knowledge transmission is hindered by limited ICT tools and opportunities, hindering students' access to up-to-date knowledge. Students need to enter the world of work with the attitudinal and competency skills to engage in continuous self-improvement, deal flexibly with possible barriers, and maintain motivation for informal learning [38]. School infrastructure is key to developing digital literacy, but access is currently limited to laptops, computers, and projectors. Technological advances such as smartboards, Lego robots, and VR glasses could help improve competence [39]. According to the 2019 OECD Survey, 39% of educators feel minimally prepared to use digital technologies. More than 20% of young people lack basic digital literacy skills [21]. There are gender differences based on the OECD 2022 survey. Research shows that boys on average get higher marks and achieve more than girls [40]. The aim of education is not only the transfer of knowledge but also the development of skills and the raising of awareness.

Correlations between the different levels of digital literacy were one of the findings of the 2020 survey [41] on the digital literacy of teacher educators. For example, the levels of digital literacy and reflective literacy were strongly correlated among respondents. Digital literacy alone is not sufficient for an adequate training process according to the expectations of the 21st Century, so cross-competences and different skills need to be examined together.

The generational needs of the information society and young people cannot be met by outdated educational methods and conservative educational approaches. Although the national curriculum has prioritized the development of digital literacy since 2007, it can still be seen as an area to be developed in today's classrooms [39]. A lack of cybersecurity awareness and sometimes poor teacher motivation, teaching methods, and digital literacy are also major problems in quality public education [35]. Different levels of education focus on developing online and digital competencies and skills to participate effectively in the online world [42]. Teaching cybersecurity can be challenging when most teachers lack skills, methodologies, and tools. Summarising, the adequate development of students' digital literacy requires qualified teaching staff with digital skills, digital pedagogy, continuous learning, and self-improvement, adequate institutional support through infrastructure and training, and an understanding of the changing roles of students. This will address both technical and behavioral issues to reduce digital illiteracy. [37]

2.4 Generation Z's Digital Literacy and Cybersecurity Awareness and their Measurement in Education Settings

The Youth Digital Skills Index (yDSI) [43], a unique internationally validated measure, was developed based on the following skills dimensions. Technological, operational, and engineering competencies, information navigation and processing competencies, communication and interpersonal competencies, and content generation competencies. The ySKILLS measurement tool was deficient in problem-solving skills, a dimension of digital literacy that has been identified in other studies but is not included in the ySKILLS concept [44]. Among other things, problem-solving skills, which were not included in ySKILLS, are included in the digital skills measurement tool developed by DigComp. DigComp [45] is a self-assessment tool designed to guide individual users in learning and improving. The framework identifies 5 domains of competence (information literacy, communication/collaboration, digital content creation, security, and problem-solving). Under security, it tests skills to protect identity, personal data and privacy, data security, and digital identity [12].

In higher education, and subsequently, in the labor market, the existence and need for cybersecurity skills and cybersecurity awareness is changing [46]. The number of attacks on businesses is increasing every year, with very serious financial consequences [47]. Remote working has had an impact on all sectors of the economy. Cybersecurity skills for almost all workers who use ICT tools in some ways have therefore become important, not just in specific IT security areas. Prevention is the most effective tool, and appropriate education, skills development, and awareness raising are the only way to ensure that a sufficiently skilled and aware workforce will enter the labor market [48]. Since the principle of security by design can be seen as a preventive technique in the design of modern industrial systems [49], a new approach to engineering education is particularly important. Generation Z grew up in the digital age and is highly digitally literate. However, they do not always have a sufficient level of cybersecurity awareness, which is necessary for them to behave safely in the digital environment. Developing and measuring cybersecurity awareness has become a priority in the educational environment. Frameworks for measuring digital literacy and cybersecurity awareness are presented in yKILLS and DIGCOMP.

3 Purpose and Methods of Research for an Experimental Cybersecurity Skills Assessment

To measure the effectiveness of cybersecurity competence development integrated into IT education, qualitative and quantitative data were collected among a small group of first-year BSc engineering students.

3.1 Aim and Methodology of the Research

The research focuses primarily on the skills of students, the skills that 21st Century workers need. It focuses on security awareness, cybersecurity knowledge, and habits within digital competencies. The direction of inference is inductive. The correlations between cybersecurity competencies and soft skills and the relationship between the methodology of teaching these skills and the cases identified in the pilot research can be used to formulate hypotheses. The reliability of the survey is dependent on the combination and validity of the techniques employed. The use of combined methods is paramount, as no single research method alone can provide complete and reliable results, especially when measuring attitudes [50]. Using mixed methods allows researchers to use multiple data sources, approaches, and analyses to confirm and better understand findings on an issue [51]. The full complexity of the behavior or habit being studied cannot be captured by a single measurement or research method. Ensuring the reliability of research findings requires examining and evaluating the collected data and findings from different perspectives and using different methodologies. Such methodological diversity adds to the validity and reliability of the research so that the research findings can be considered more reliable and comprehensive. [52]

The research relies on the use of paper-based questionnaires and random sampling. A total of 130 first-year undergraduate engineering students were participants. Responses to self-report questionnaires should be treated with caution, as they may not reflect true habits and reality in the case of inadequate self-awareness. Due to the self-reflective nature of the questionnaire, a face-to-face focus group discussion is justified from a research perspective [51]. The research was complemented by a focus group discussion with an interpretive photo interview [50]. The selection of participants for the focus group discussion was among the respondents to the questionnaire. Six students were selected at random to participate in the focus group. Their habits in cyberspace, which were the focus areas of the questionnaire, were evaluated. The creative and interactive technique used in the focus group discussion helps the interviewees to express their hidden thoughts and spontaneous answers in a freer way. The image association technique makes the interview process more effective and helps the respondents to understand the interview by using an alternative approach.

3.2 Quantitative Research - Self-Reflection Questionnaire

The skills assessment part of the questionnaire was modeled on existing DigComp [45] and yDSI [44] framework survey tools. The questions focus on cyber-security awareness, with some rephrasing and addition of questions from existing questionnaires. The result is a 40-item self-reflection questionnaire that contributes to the understanding of the effectiveness of the preventive education process in terms of digital literacy, security awareness, and cybersecurity skills acquisition. The aim of the questionnaire is not only to measure skills but also to explore the diversity of education methods and sources of knowledge acquisition. The soft skills of the students are also measured by the questionnaire. The questionnaire consists of explicit, closed-ended questions, using a 6-point Likert scale that requires a certain degree of separation from the interviewee. A score of 1 means 'strongly disagree', a score of 4 means 'strongly agree' and a score of 2-3 means 'tend to disagree' - 'tend to agree'. 5 means "I don't understand the question, I don't know what it means" and 6 means "I don't want to answer". Response category 5 represents both inadequate skills and inadequate knowledge.

3.3 Focus Group Interview with Image Association Technique

The relevant research methodology of the focus group discussions is content analysis [53], so this was also chosen in this research. The research questions were summarized in the form of alternative qualitative research to deepen the research. The focus group discussion is conducted using an interpretive photo interview, and image association technique. The subjects of the focus group discussion, 6 persons (1 female, 5 males in gender distribution), were selected randomly from the questionnaire survey. The focus group discussion searches for answers to the attitudes, skills, and competencies of Generation Z youth in cyberspace. The discussion also reveals the methods that teachers used in the previous secondary and primary education processes to help students acquire competencies that they had already gained or even lacked. The image association technique is used as a way of stimulating and motivating reflection on the topic [50]. Verbal questions are followed by viewing photos and images, followed by a spontaneous interpretation. It is used to support human attitudes, the way individuals think, and the accurate and detailed collection of their experiences.

4 Results of the Cybersecurity Competency Assessment

The results obtained from the questionnaire were subjected to descriptive statistics, looking at the average scores and the distribution of the data. The demographic composition of the sample is 88% male (N=114) and 12% female (N=16), with women underrepresented in our sample. The presented pilot research results are, therefore, not generalizable due to the sampling method used. The survey studied the types of the educational background of the participants. In terms of the percentage distribution, 32% of survey participants entered engineering higher education from general education and 25% from technical secondary school, so 1/4 of them were able to study an engineering subject in-depth during their secondary education. Examining the educational background of the survey respondents and the field of their previous education, slightly more than 50% of the students who were admitted had some type of prior technical education. Figure 2 presents the results for each skill group based on the average of all survey respondents, highlighting the larger negative differences in the areas of safety, stress management, and group.

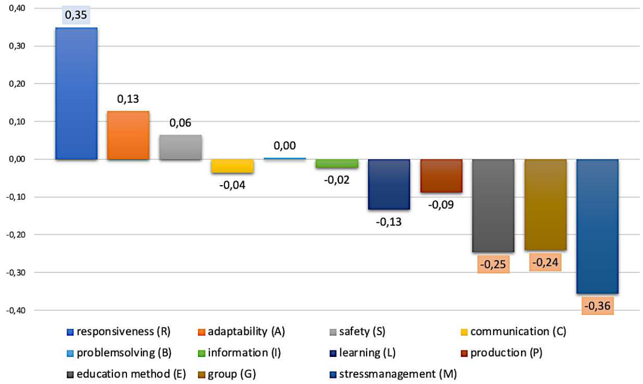


Figure 2

Results for each skill group averaged across survey respondents Source: Author's construction

Participants self-reported performing exceptionally well on the rapid response questions but underperformed in several areas. One critical group is the area of stress management. Figure 2 shows that among the areas that are more strongly below the second and third averages are the ability to work in groups and to evaluate experienced teaching methods. Categorizing the skills tested, the best mean score=2.01 is for safety skills, while the lowest mean score=1.76 is for teaching and learning skills. The results of the focus group discussion presented later are consistent with these results. The interviews reveal a definite problem with the teaching methods used, which are ineffective and students report having limited support in learning to understand their learning habits and to apply appropriate methods. The Bartlett test and the KMO (Kaiser-Meyer-Olkin) measure were used to check that the conditions for factor analysis were met. The value of the KMO measure was 0.622. The result of Bartlett's test was statistically significant ($\chi2 = 911.684$, df = 351, $p < 0.001$), indicating a significant difference between the variances of the responses to the cybersecurity awareness questions. Different correlations between statements indicating cybersecurity awareness were found based on the results. Awareness of the ability to block unsolicited pop-ups and attention to the consequences of online activity were negatively correlated (r=-0.315). A positive correlation was found between the awareness of the use of copyright-protected content and the knowledge of copyrights and licences (r=+0.512). There is also a positive correlation between awareness of security settings and awareness of the protection of sensitive data (r=+0.585) and between awareness of trustworthy websites and knowledge of security settings (r=+0.596). Finally, there is a positive correlation between the useful knowledge of cyber-security acquired during the IT training and the attention paid to the security of the Internet (r=+0.606).

In the rest of the research, the article examines in more detail the different aspects of safety skills and their interrelationships. In terms of the scope of the questions, a total of N=3510 evaluable responses were received from participants during the data collection. The following characteristics are collected in Table 1.

Table 1

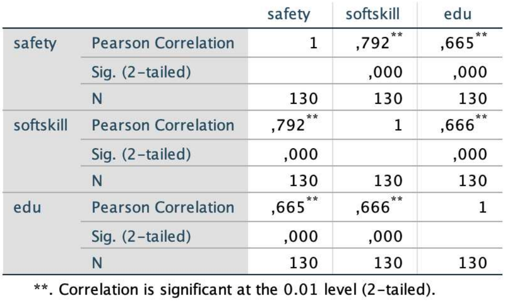
Summary table of the different distributions of security and cybersecurity issues

Source: Author's construct based on the answers to the security questions of the self-reflection questionnaire (N=3510)

	Absolute frequency distribution	**Relative frequency distribution**	**Cumulative frequency distribution**
	f(a)	***f(%)***	***f(c)***
0 - not	284	8,09	284
1 - rather not	632	18,01	916
2 - rather yes	1208	34,42	2124
3 - yes	1321	37,64	3445

Analyzing further the safety skills of the respondents, the results of female participants (Mean=1,99 Mode=3 Std. Deviation=0,938) and male respondents (Mean=2,04 Mode=3 Std. Deviation=0,948) do not show significant differences, apparently not affected by gender in terms of safety skills and related habits.

From a different perspective, the questions can be divided into two broad domains, the emotional, physiological, and environmental aspects, and the engineering and technical aspects. The safety aspect also shows no significant difference in the respondents' results when the questions are divided into groups. The results for the emotional, physiological, and environmental aspects (mean=2.07 mode=3 std. deviation=0.967) and the engineering and technical aspects (mean=2.00 mode=3 std. deviation=0.922) are almost identical. Categorizing the responses to the survey, the deviation from the mean and standard deviation of the security responses indicates that students in business, public administration, and law performed exceptionally well overall. Additionally, students in STEM and Arts and Humanities were particularly outstanding for safety questions related to emotional, physical, and environmental aspects. Figure 3 shows the correlation coefficients between cybersecurity awareness, soft skills, and the effectiveness of education methods.



		safety	softskill	edu
safety	Pearson Correlation	1	,792**	,665**
	Sig. (2-tailed)		,000	,000
	N	130	130	130
softskill	Pearson Correlation	,792**	1	,666**
	Sig. (2-tailed)	,000		,000
	N	130	130	130
edu	Pearson Correlation	,665**	,666**	1
	Sig. (2-tailed)	,000	,000	
	N	130	130	130

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 3

Correlation between each skill Source: Author's construction

A Pearson correlation was conducted to determine the relationship between cybersecurity awareness, soft skills, and education, and learning skills. A strong positive correlation was found between security awareness and soft skills. This correlation is statistically significant (r = 0.792, n = 130, p < 0.01). This means that participants' cybersecurity awareness is reciprocally enhanced by more advanced communication, problem-solving, and collaboration skills. There is also a moderately strong positive correlation between awareness and education (r = 0.665, n = 130, p < 0.01), meaning that knowledge and skills acquired through innovative teaching methods and appropriate learning skills contribute to cybersecurity awareness development.

5 Interpretative Focus Group Photo Interview Results

Content analysis was used to analyze the results of the focus group interviews. The research involved interviewing six respondents, and the interviews were mixed using an associative technique to enhance the dynamics of the conversation. The interviews were recorded and then transcribed. The transcripts included the interviewees' responses to the research questions. The survey included open-ended questions focusing on online threats, how to set and use security when browsing, how to protect and set personal devices, and how to share information on the Internet. In addition, the survey covered the useful cybersecurity knowledge that was acquired during the studies and its practical use, as well as the educational methods and approaches. The research aimed to get a comprehensive picture of the respondents' cybersecurity awareness and behavior online, as well as the role of education in developing cybersecurity knowledge. Content analysis involved categorizing the responses given based on the questions asked, and then evaluating these categories and their relationships. The choice of methodology allowed for a deeper understanding of the students' opinions, attitudes, and experiences [52]. Participants enjoyed sharing their own experiences but tended to share the experiences of others. The pictures were a great help in revealing deeper thoughts and connections, they were happy to turn to the pictures for help when they got stuck or found it difficult to open up about the topic.

According to the results of the reflective questionnaire, respondents are generally aware of the dangers of cyberspace, but this awareness was only partially confirmed during the focus group. They consider cyberbullying to be the most dangerous form of bullying, but they also recognize that people of all ages can be at risk. Respondents are aware of software that can protect them from threats, but they do not actively use these solutions because they do not have a strong sense of threat. In general, they don't use many security settings in the equipment they use. Biometric identification, anti-virus software, and ad blockers are sometimes used on the computer or mobile phone, but no other protective measures are installed. Based on the data collected in the interview, chat apps, online friends, online dating, collecting likes, Facebook, sharing personal data, manipulation, anonymity, gender neutrality, and Tinder are the aspects that influence the preferred information-sharing habits of Generation Z. The online space is of particular importance to them because of these social interactions, especially the social networking sites that they use. They want to make themselves visible in cyberspace, although they are careful about whom they share content with.

The quantity and quality of the educational process and knowledge transfer was the second area of research identified during the interviews. They had not received any useful cybersecurity knowledge in their classes, although some of them had attended technical secondary schools. They lacked useful knowledge about cybersecurity, information security, and defenses. They learn about various incidents and defenses from the news. Most of their existing knowledge and skills were gained from personal experience. Students encountered few methodological innovations during the interview. When it comes to pedagogical methods, they experienced traditional pedagogical methods while studying, and sometimes they were only taught using innovative, modern methods. They believe that they will need to learn throughout their lives to be successful and that internal motivation will be particularly important in the future. Money and multiple career opportunities are their motivators for learning. There was a clear desire for a change in the way people were taught, a need for a different, more practical way of teaching, and a reduction in the amount of theoretical knowledge they had been taught.

6 Discussion

Based on the results of the research, it can be concluded that developing cybersecurity awareness in public education is significantly related to industry in supporting the skills needed in the 21st Century. Cybersecurity education and awareness in educational institutions effectively contribute to developing students' digital literacy and security awareness. Gen Z students place a high value on the labor market benefits of a degree while showing little interest in new learning opportunities. When it comes to getting a good job, they consider the importance of relationships and confidence to be paramount. At the same time, the concept of changing people's mindsets, retraining, and lifelong learning is particularly important, as the rapidly changing labor market makes it important for everyone to be prepared for career change and adaptability [54].

Workplace culture affects the effectiveness of responding to cybersecurity incidents, according to 68% of employees [25]. Organizations need to recognize changes in the labor market and adapt to cybersecurity challenges, as workplace culture and employee satisfaction have a significant impact on employee efficiency. Students are motivated by money and career opportunities, thus increasing their learning activities while working and leaving public education. In the coming years, professionals will have to face a shortage of workers and the management of risks generated by new technologies, especially because of working from home [25]. These findings underline the importance of a close relationship between public education and industry to be successful in providing the skills that will be needed in the 21st Century.

Based on the results of the research, it can be concluded that the development of cybersecurity awareness is closely related to soft skills and applied educational methodologies. Defense reflexes and security awareness are already in place for physical threats, but further development is still necessary in cyberspace [9]. Education has a key role to play in raising awareness of cybersecurity, and experience and practical skills are of key importance for recruits. Relevant IT experience (29-35%), strong problem-solving skills (38-44%), and proper cybersecurity experience (31-35%) are emphasized [25]. There is a lack of response, prevention, and mitigation of already existing problems and attacks. For this reason, in many places, annual training courses, e-learning materials or exams may not be enough to create a real awareness of security [10]. Findings indicate that security awareness is strongly positively related to soft skills and moderately positively related to education. Commitment, self-awareness, and the ability to change are the basis for successful self-education, for which students need to be prepared [9].

Soft skills such as stress management, problem-solving, communication skills, working with others, and conflict management should be developed in students when they enter higher education [55] [56]. The students had acquired most of their existing knowledge and skills from their own experience. There were few methodological innovations encountered by the students during the interviews. In terms of pedagogical methods, students reported similar experiences in surveys of similar age groups, where they had experienced traditional pedagogical methods during their learning and in some cases had only been taught using innovative, modern methods [57]. Cybersecurity awareness is also enhanced by the innovative use of teaching methods and appropriate learning skills. Gender does not affect security skills and habits, as further analysis of respondents' security skills shows no significant difference between female and male respondents. This has led to a contradictory result, as based on the results of several studies [26] [58] [59], men are more aware of security issues. Further research with a larger number of participants would be necessary for confirmation or rejection of this finding.

As far as the habits of the generation are concerned, we can confirm the following results. According to a 2021 survey, over 84% of Hungarians participate in social networking sites, the highest rate in the EU [60]. For Generation Z, the online space, especially social networking sites, plays an important role in socializing and sharing information. Chat apps, online friends, online dating, collecting likes, Facebook, sharing personal information, manipulation, anonymity, gender neutrality, and Tinder are the aspects that influence Generation Z's preferred information-sharing habits, according to the data collected during the interview. Because of these social interactions, especially the social networking sites they use, the online space is particularly important to them. Based on the results, it is important to teach cyber-security awareness and soft skills to support a safer online presence for young people.

Summary, Further Research Directions

In the 21st Century, competencies, and skills in the field of cybersecurity are essential for young professionals. The present exploratory pilot study was able to process 130 usable responses through a quantitative method, which already allowed statistical processing of the data, but the results cannot be generalized. The main objective of the research was to explore the cybersecurity competencies, soft skills, educational experiences, and mindsets of the students.

Based on the research results, it can be concluded that soft skills and applied pedagogical methods are closely related to the development of cybersecurity awareness. Cyber-security awareness shows a strong positive correlation with soft skills, meaning that the more developed the communication, problem-solving, and collaboration skills of the students, the higher their cybersecurity awareness. The research also showed that knowledge and skills acquired through innovative pedagogical methods as well as appropriate learning skills contribute to developing cybersecurity awareness. These findings highlight the role of soft skills and applied pedagogical methods in increasing students' cybersecurity awareness and emphasize the importance of cybersecurity education and learning. There is a clear justification for the continuation of the research in the future, and its results can be useful in promoting the renewal of education.

References

[1] T. Butler-Bowdown, “Psychology in a nutshell 50 basic psychological works” (Pszichológia dióhéjban 50 pszichológiai alapmű), *HVG Könyvek*, 2007

[2] A. Beláz, D. Berzsenyi, “Cybersecurity Strategy 2.0: Issues for strategic cybersecurity governance” (Kiberbiztonsági Stratégia 2.0: A kiberbiztonság stratégiai irányításának kérdései), *Center for strategis and defense studies analyses* (*Stratégiai védelmi kutató központ (elemzések))*, pp. 1-15, 15 p., 2017

[3] M. Alshaikh et al., “Toward Sustainable Behaviour Change: An Approach for Cyber Security Education Training and Awareness”, *27th European Conference on Information Systems (ECIS),* Stockholm & Uppsala, Sweden, 2019

[4] A. P. Bodó et al., “Targeted cyber-attacks. Annual training for staff involved in the security of electronic information systems” (Célzott kibertámadások. Éves továbbképzés az elektronikus információs rendszer biztonságával összefüggő feladatok ellátásában részt vevő személy számára), Budapest, Hungary: *Nemzeti Közszolgálati Egyetem*, 2018

[5] A. Szarvák, V. Póser, “Information Technology Safety Awareness – a review of regularly used terms and methods” *15th International Symposiumon Applied Informatics and Related Areas organized in the frame of Hungarian Science Festival 2020:* AIS 2020 Székesfehérvár, Magyarország: Óbudai Egyetem, pp. 107-111, 5 p., 2020

[6] W.-H. So, H. Kim, “A Study on the Online School Violence of Teenagers in Cyberspace”, *Asia-Pacific Journal of Convergent Research Interchange*, Vol. 7, No. 1, pp. 105-114, 2021, doi: 10.47116/apjcri.2020.01.10

[7] D. Berzsenyi, “The human side of cybersecurity” (A kiberbiztonság humán oldala), *Nemzet és Biztonság–Biztonságpolitikai Szemle 10.2*, pp. 54-67, 2017

[8] S. Baraković, J. B. Husic, “Cyber hygiene knowledge, awareness, and behavioral practices of university students”, *Information Security Journal: A Global Perspective*, pp. 1-24, 2022, doi: 10.1080/19393555.2022.2088428

[9] I. Dobák, S. Babos, “Security awareness opportunities in the light of 21st Century platforms” (A biztonságtudatosítás lehetőségei a 21. századi platformok fényében), *Nemzetbiztonsági Szemle*, Vol. 9, No. 4, pp. 18-34, 2021, doi: 10.32561/nsz.2021.4.2

[10] R. Gyaraki, “The role of security awareness, or questions about cybersecurity” (A biztonságtudatosság szerepe, avagy kérdések a kiberbiztonságról), *Magyar Rendészet*, Vol. 22, No. 2, pp. 245-261, 2022, doi: 10.32577/mr.2022.2.16

[11] A. Corallo et al., “Cybersecurity awareness in the context of the Industrial Internet of Things: A systematic literature review”, *Computers in Industry,* 137, 103614, 2022, https://doi.org/10.1016/j.compind.2022.103614

[12] EU Science Hub, European Commission, *The Digital Competence Framework 2.0*, 2021, URL: https://tinyurl.hu/33WQ (last retrieved: 2021.10.21.)

[13] P. Leitão et al., “Analysis of the Workforce Skills for the Factories of the Future”, *IEEE Conference on Industrial Cyberphysical Systems (ICPS2020),* Vol. 1, pp. 353-358, 2020, doi: 10.1109/ICPS48405.2020.9274757

[14] N. Soukupová et al., “Industry 4.0: an Employee Perception” (Case of the Czech Republic), *Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis,* Vol. 68, No. 3, pp. 637-644, 2020, doi: 10.11118/actaun202068030637

[15] N. Obermayer et al., “Companies on Thin Ice Due to Digital Transformation: The Role of Digital Skills and Human Characteristics”, *International and Multidisciplinary Journal of Social Sciences,* Vol. 11, No. 3, pp. 88-118, 2022, doi: 10.17583/rimcis.10641

[16] W. Puriwat, S. Tripopsakul, “Preparing for Industry 4.0 - Will youths have enough essential skills?: An Evidence from Thailand”, *International Journal of Instruction*, Vol. 13, No. 3, pp. 89-104, 2020, doi: 10.29333/iji.2020.1337a

[17] N. Obermayer et al., “Influence of Industry 4.0 technologies on corporate operation and performance management from human aspects”, *Meditari Accountancy Research,* Vol. 30, No. 4, pp. 1027-1049, 2022, doi: 10.1108/MEDAR-02-2021-1214

[18] S. Von Solms, L. A. Futcher, "Adaption of a Secure Software Development Methodology for Secure Engineering Design", *IEEE Access,* Vol. 8, pp. 125630-125637, 2020, doi: 10.1109/ACCESS.2020.3007355

[19] F. Iniesto et al., "When industry meets Education 4.0: What do Computer Science companies need from Higher Education?", *TEEM'21: Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality,* pp. 367-372, 2021, doi: 10.1145/3486011.3486475

[20] European Parliament 2021/C 56/02, "Opinion of the European Economic and Social Committee on 'Industrial transition towards a green and digital European economy: regulatory requirements and the role of social partners and civil society' (exploratory opinion)", 2021, URL: https://eur-lex.europa.eu/ (last retrieved: 2022.03.17)

[21] S. Vandekerckhove et al., "Musculoskeletal disorders and psychosocial risk factors in the workplace — statistical analysis of EU-wide survey data, Report", *European Agency for Safety and Health at Work,* Publications Office of the European Union, 2021, doi: 10.2802/39948

[22] B. Fregan, I. Kocsis, Z. Rajnai, "IPAR 4.0 and the risks of digitalisation" (Az IPAR 4.0 és a digitalizáció kockázatai), *Műszaki Tudományos Közlemények (HU) 9:* 1 pp. 87-90, 4 p., 2018

[23] Computerworld, "Cybersecurity is not just an IT problem" (A kiberbiztonság nem csak az informatikusok problémája), 2019, URL: https://tinyurl.hu/OkCd (last retrieved: 2021.08.17)

[24] Cybersecurity Guide, A guide for women in cybersecurity, 2021, URL: https://tinyurl.hu/14Cb (last retrieved: 2021.08.20)

[25] ISC, Cybersecurity Workforce Study: A critical need for cybersecurity professionals persists amidst a year of cultural and workplace evolution, 2022, URL: https://tinyurl.hu/LpqC (last retrieved: 2023.02.14.)

[26] T. Palicz et al., "Results of the 2020 National Population Survey on Security Awareness in Cyberspace" (Biztonságtudatosság a kibertérben – a 2020-as országos lakossági felmérés eredményei), *Belügyi Szemle*, Vol. 70, No. 2, pp. 395-418, 2022, doi: 10.38146/bsz.2022.2.11

[27] S. M. Kennison, E. Chan-Tin, "Taking Risks with Cybersecurity: Using knowledge and personal characteristics to predict Self-Reported Cybersecurity Behaviors", *Frontiers in Psychology,* Vol. 11, 2020, doi: 10.3389/fpsyg.2020.546546

[28] Cybersecurity Guide, How to Become a Cybersecurity Specialist, 2021, URL: https://tinyurl.hu/r1m4 (last retrieved: 2021.09.23)

[29] 1139/2013. (III. 21.) "Government Decision on the National Cyber Security Strategy of Hungary" (Kormányhatározat Magyarország Nemzeti

Kiberbiztonsági Stratégiájáról), *Magyar Közlöny Lap- és Könyvkiadó Kft.,* 2013

[30] V. Radunovic, D. Rüfenacht, "Cybersecurity competence building trends Research report Commissioned by the Federal Department of Foreign Affairs of Switzerland", *DiploFoundation,* 2016

[31] European Comission, the European Economic and Social Committee and the Committee of the Regions on the Digital Education Action Plan COM(2018) 22 final, 2018

[32] 1163/2020. (IV.21.) "Government Decision on the National Security Strategy of Hungary " (Kormányhatározat Magyarország Nemzeti Biztonsági Stratégiájáról), 2020

[33] K. Fekete-Karydis, B. Lázár, "Development of cyber defense strategies, cyber defense challenges, current events" (A kibervédelmi stratégiák fejlődése, kibervédelmi kihívások, aktualitások), *HSZ-HDR*, köt. 147, sz. 5, o. 60-72, 2021

[34] R. Stohl, ""How to Train Your Dragon!" – About the training and learning habits of Generation Z" („Így neveld a sárkányodat!" – A Z generáció képzési és tanulási szokásairól), *Honvédségi Szemle – Hungarian Defence Review*, pp. 116-127, 2021, doi: 10.35926/hsz.2021.2.9

[35] Z. Nyikes, "Information security enhancement with user support options" (Az információbiztonság növelése a felhasználó támogatásának lehetőségeivel), Studia Doctorandorum Alumnae II.: Válogatás a DOSz Alumni Osztály tagjainak doktori munkáiból II. Budapest, Magyarország : *Doktoranduszok Országos Szövetsége (DOSZ),* 2021, 964 p. pp. 637-806, 170 p.

[36] European Comission, Digital Education Action Plan 2021-2027, Resetting education and training for the digital age, 2020

[37] Z. Balogh et al. "The impact, characteristics and challenges of digital literacy and digital culture on society and education" (A digitális kompetencia és a digitális kultúra társadalomra és oktatásra gyakorolt hatásai, jellemzői, kihívásai), *Civil Szemle 17:* 2 pp. 69-88, 19 p., 2020

[38] A. Kálmán, B. G. Kálmán, "The impact of industry 4.0 competence requirements on school system education" (Az ipar 4.0 kompetenciaigényeinek hatása az iskolarendszerű oktatásra), *Iskolakultúra*, Vol. 32, No. 12, pp. 57-73, 2022

[39] T. L. Nyitrai, "The home position of teacher digital competence in public education before COVID-19" (A tanári digitális kompetencia helyzete a közoktatásban a COVID-19 előtt),*jATES,* Vol. 11, No. 2, pp. 124-136, 2021

[40] National competence measurement 2022 (Oktatási Hivatal, Országos kompetenciamérés 2022), URL: https://tinyurl.hu/RvR6 (last retrieved: 2023.03.19.)

[41] L. Horváth et al., "Measuring the digital competence of teacher educators - adapting DigCompEdu to the domestic higher education environment" (Tanárképzők digitális kompetenciájának mérése – a DigCompEdu adaptálása a hazai felsőoktatási környezetre), *Neveléstudomány: Oktatás Kutatás Innováció 8:* 2, pp. 5-25, 21 p., 2020

[42] K. Thiyagu et al., "Cyber safety and security education", *Lulu Publication,* 2019

[43] E. J. Helsper et al., "The youth Digital Skills Indicator", *Zenodo*, 2021, doi: 10.5281/zenodo.4476540

[44] E. J. Helsper et al., "The Youth Digital Skills Indicator: Report on the conceptualisation and development of the ySKILLS digital skills measure", 2021, URL: https://osf.io/m84pe/ (last retrieved: 2021.10.21.)

[45] Cs. Kvaszingerné Prantner, "DIGCOMP 1.0 and DIGCOMP 2.0, The impact of culture change on individual competences: models of digital competence" (A DIGCOMP 1.0 és a DIGCOMP 2.0, A kultúraváltás hatása az egyéni kompetenciákra: a digitális kompetencia modelljei), Eger, Magyarország : *EKE Líceum Kiadó*, 143 p. pp. 59-74, 16 p., 2020

[46] J Novák, "Methods to increase safety awareness in higher education" (Biztonságtudatosság növelésének eszközei a felsőoktatásban), *Műszaki Tudományos Közlemények (HU) 9* : 1, pp. 183-186, 4 p., 2018

[47] Cs. Kollár, J. Poór, "Organisations in the digital age - Information security aspects of the digital workplace" (Szervezetek a digitális korban – A digitális munkahely információbiztonsági aspektusa), *Kiberbiztonság - Cyber Security: Tanulmánykötet a Biztonságtudományi Doktori Iskola kutatásaiból,* Budapest, Magyarország: Óbudai Egyetem, Bánki Donát Gépész és Biztonságtechnikai Mérnöki Kar, 366 p. pp. 95-107, 13 p., 2018

[48] Z. Nyikes, "Possibilities for developing security awareness" (A biztonságtudatosság fejlesztésének egyes lehetőségei), *Műszaki Tudományos Közlemények* (HU) 7 pp. 327-330, 4 p., 2017

[49] P. Bóna, "User awareness is the first line of defence" (A felhasználók biztonságtudatossága az első védelmi vonal) - Videó, *Comforth.hu,* 2020, URL: https://bit.ly/3Ek0WGX (last retrieved: 2021.07.08)

[50] D. Horváth, A Mitev, "Alternative qualitative research manual" (Alternatív kvalitatív kutatási kézökönyv), *Aliena Kiadó,* 2015

[51] A. Kelemen-Erdős, A. Á. Mészáros, "Ethics and Social Responsibility of Information Intermediaries in International Businesses", *Arab Journal of Administration* 41 pp. 239-248, 10 p. 2021

[52] Á. Szokolszky, “Research work in psychology: methodology, methods, practice” (Kutatómunka a pszichológiában: metodológia, módszerek, gyakorlat), *Osiris tankönyvek*, 2004

[53] A. Kelemen-Erdős, “Selection Listing Decisions: New Product Adoption of Food Retailers”, *Journal of Research in Business, Economics and Management 10*: 3 pp. 1905-1917, 13 p. 2018

[54] I. C. Papp et al., “Study preferences in higher education”, *Acta Polytechnica Hungarica*, Vol. 20, No. 4, pp. 229-248, 2023, doi: 10.12700/aph.20.4.2023.4.13

[55] Gy. Molnár, B. Orosz, “Current issues of digitisation processes in a changing digital environment: reflections on some pedagogically relevant contexts in Hungary” (Digitalizációs folyamatok aktuális kérdései változó digitális környezetben: Reflexiók néhány magyarországi pedagógia-releváns kontextusra), Komárno, Szlovákia: *International Research Institute*, 378 p. pp. 120-131, 12 p., 2020

[56] J Módné Takács, M. Pogátsnik, “Examining the stress management techniques of university students” (Az egyetemi hallgatók stresszkezelési technikáinak vizsgálata), *Módszertani újítások és kutatások a szakképzés és a felsőoktatás területén: X. Trefort Ágoston Szakképzés- és Felsőoktatás-pedagógiai Konferencia Tanulmánykötet,* Budapest, Magyarország : Óbudai Egyetem, 424 p. pp. 262-278, 17 p., 2021

[57] G. Farkas et al., “Quality Improvement in Education, based on Student Feedback”, *Acta Polytechnica Hungarica*, Vol. 20, No. 6, pp. 215-228, 2023, doi: 10.12700/aph.20.6.2023.6.12

[58] S. M. Kennison, E. Chan-Tin, “Taking Risks with Cybersecurity: Using knowledge and personal characteristics to predict Self-Reported Cybersecurity Behaviors”, *Frontiers in Psychology*, Vol. 11, 2020, doi: 10.3389/fpsyg.2020.546546

[59] T. McGill, N. Thompson, “Gender differences in information security perceptions and behaviour, in University of Technology”, *Sydney eBooks*, 2018, doi: 10.5130/acis2018.co

[60] EUROSTAT, Digital society statistics at regional level, 2022, URL: https://tinyurl.hu/eOpI (last retrieved: 2023.03.19.)

Finding Maximum Tolerated Dose in Phase I Oncology Clinical Trials with Bayesian Methods

Johanna Sápi[1]

[1]John von Neumann Faculty of Informatics, Biomatics and Applied Artificial Intelligence Institute and University Research and Innovation Center, Physiological Controls Research Center, Óbuda University, Bécsi út 96/b, Budapest, H-1034, sapi.johanna@uni-obuda.hu

Abstract: Maximum tolerated dose (MTD) is a maximal amount of drug or radiation resulting relatively acceptable dose-limiting toxicity (DLT). Accurate value of MTD should be found in Phase I trials in order to create the possibility to conduct successful Phase II (pilot efficacy and safety evaluation) and Phase III (comparative efficacy) trials. The aim of this paper is to review the difficulties of the dose-finding methods including multi-agent problems and late-onset toxicities, and to discuss Bayesian adaptive dose-finding methods which can handle these issues.

Keywords: maximum tolerated dose (MTD); dose-limiting toxicity (DLT); Continual Reassessment Method (CRM); Time-To-Event Continual Reassessment Method (TITE-CRM); copula regression model; logistic regression model; delayed toxicities; late-onset toxicity model

1 Introduction

The main aim of evidence-based medicine is to collect, analyze and critically evaluate research data, and translate systematically collected and evaluated medical knowledge into practice in order to obtain optimal health outcomes [1]. Nowadays, evidence-based approach is fundamental in the field of oncology as well, and biostatistics is an important tool for this.

The aim of Phase I clinical cancer studies from oncological point of view is to find the maximum tolerated dose (MTD) of a drug or radiation which refers to a maximal amount of drug resulting relatively acceptable (typically grade 3) dose-limiting toxicity (DLT) [2]. Knowing the precise value of MTD has a key role in oncological treatment design [3].

Phase I design methods can be divided into three groups: algorithm-based designs, model-based designs and model-assisted designs [4–6].

Algorithm-based designs are conventional designs in the sense that there are prespecified rules to decide on the dose escalation and de-escalation. Algorithm-

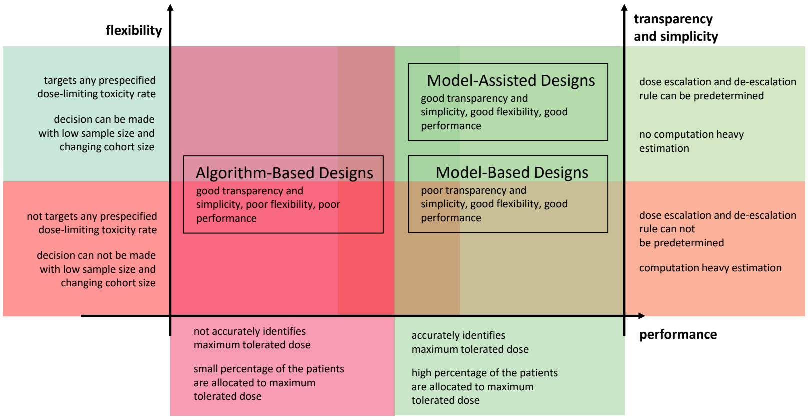

flexibility
targets any prespecified dose-limiting toxicity rate
decision can be made with low sample size and changing cohort size
not targets any prespecified dose-limiting toxicity rate
decision can not be made with low sample size and changing cohort size
Algorithm-Based Designs
good transparency and simplicity, poor flexibility, poor performance
Model-Assisted Designs
good transparency and simplicity, good flexibility, good performance
Model-Based Designs
poor transparency and simplicity, good flexibility, good performance
transparency and simplicity
dose escalation and de-escalation rule can be predetermined
no computation heavy estimation
dose escalation and de-escalation rule can not be predetermined
computation heavy estimation
performance
not accurately identifies maximum tolerated dose
small percentage of the patients are allocated to maximum tolerated dose
accurately identifies maximum tolerated dose
high percentage of the patients are allocated to maximum tolerated dose

Figure 1
Characteristics of Phase I Design groups.
Transparency and simplicity criterion is defined based on whether dose escalation and de-escalation rule can be predetermined, and the estimation is computation heavy or not. Flexibility refers to the ability that a design targets prespecified dose-limiting toxicity rate, and decision can be made with low sample size and changing cohort size. The good performance criterion is met when the design accurately identifies maximum tolerated dose, and high percentage of the patients are allocated to maximum tolerated dose. In terms of these criteria, algorithm-based designs are the least applicable; model-based designs have good perfomance and flexibility, but these designs can be overly complex; while model-assisted designs combine all the good properties.

based designs group contains the most common Phase I design method, the ”$3+3$” design which can be used in a single-agent trial. Albeit it is a simply and easy to use model, it has been widely criticized due to its poor efficiency in terms of treating too many subjects at a suboptimal dose and weakly estimating MTD [7].

Taking into account not a single-agent but a drug combination trial, the process is more challenging due to the complex drug–drug interactions [8]. However, in oncology, combination therapy is often used due to its synergistic treatment effect.

Model-based designs are adaptive designs where a statistical model is used in order to quantify the dose-toxicity relationship, and describe the dose-toxicity curve [9, 10]. Model-based designs have the following dose-finding strategy. The first step is creating a probability model (that can be parametric or non-parametric) in order to quantify the dose-toxicity relationship. The second step is the collection of data from the treated patients, and based on that, the model continuously updates the estimate of the model after each cohort, and this updated estimation is used to find the dose for the next cohort. The final step is the identification of the maximum tolerated dose based on the estimated toxicity probabilities of the dose combinations. Model-based designs group includes the Continuous Reassessment Method (CRM), and the Bayesian copula regression and logistic regression model.

Model-assisted designs group is a relatively new class of trial designs which were developed in order to combine the advantages of algorithm-based and model-based designs. Before the onset of the trial, dose escalation and de-escalation rule can be predetermined (like in algorithm-based designs), and a statistical model is used in order to quantify the dose-toxicity relationship, and describe the dose-toxicity curve (like in model-based designs). Model-assisted designs group contains e.g. the Bayesian Optimal Interval (BOIN) design [11, 12] and the keyboard design [13] for single-agent dose finding.

According to Yuan et. al [5], Phase I design characteristics can be evaluated on three criteria. Transparency and simplicity criterion is defined based on whether dose escalation and de-escalation rule can be predetermined, and the estimation is computation heavy or not. Flexibility refers to the ability that a design targets prespecified dose-limiting toxicity rate, and decision can be made with low sample size and changing cohort size. The good performance criterion is met when the design accurately identifies maximum tolerated dose, and high percentage of the patients are allocated to maximum tolerated dose. In terms of these criteria, algorithm-based designs are the least applicable; model-based designs have good perfomance and flexibility, but these designs can be overly complex; while model-assisted designs combine all the good properties (Fig. 1). Besides Yuan's evaluation, there are other comparative reviews discussing the pros and cons of algorithm-based designs, model-based designs and model-assisted designs (e.g. [6]).

The paper is organized as follows. The second section discusses the most common dose-finding solution for single-agent trials, namely the Continual Reassessment Method. In the third section, two Bayesian adaptive dose-finding methods for multi-agent trials are shown, a copula-type regression model and a logistic regression model. The fourth section considers the question of late-onset toxicities and presents two different methods to handle this problem. Time-To-Event Continual Reassessment Method offers a solution for single-agent trials, while Bayesian data augmentation approach can be used in multi-agent trials. The paper ends with the conclusion section.

2 Continual Reassessment Method (CRM) for Single-Agent Dose-Finding Trials

In dose-finding studies, the typical procedure is that a sequence of doses is investigated in order to find DLT and the corresponding MTD. The main assumption in Continual Reassessment Method [14–16] is that by increasing drug dose, the probability of therapeutic efficacy is monotonically increasing, as well as the probability of toxicity. Hence, the main purpose of Phase I trials is to find a trade-off solution, viz. finding the most efficacious therapy which results in tolerable toxicity risk. Steps of the CRM are the following [14].

Step 1. Choosing of an *a priori* dose-toxicity model. There are two main groups of *a priori* dose-toxicity curve models: one-parameter and two-parameter models. In the case of one-parameter models, the intercept of the curve is fixed, and trial data update the slope (s) of the curve from cohort to cohort. In contrast, using two-parameter models, both intercept (i) and slope (s) of the curve is re-estimated step by step. Advantage of the one-parameter models is that they require less information; however, their accuracy is limited due to the fixed intercept parameter. Using two-parameter models, the accuracy can be improved, but a bigger data set is required for good estimation. In the following, the mostly used dose-toxicity models are listed:

- hyperbolic tangent model

  $$p_{toxicity}(dose) = \left(\frac{\tanh(dose)+1}{2}\right)^{s}, \tag{1}$$

  where s is the slope of the curve.

- one-parameter logistic model

  $$p_{toxicity}(dose) = \frac{\exp(c+s\cdot dose)}{1+\exp(c+s\cdot dose)}, \tag{2}$$

  where s is the slope of the curve, and c is a constant, typically $c=-4$ or $c=3$.

- two-parameter logistic model

  $$p_{toxicity}(dose) = \frac{\exp(i+s\cdot dose)}{1+\exp(i+s\cdot dose)}, \tag{3}$$

  where s is the slope, and i is the intercept of the curve. In Fig. 2 (blue solid line), a two-parameter logistic curve is chosen as an *a priori* dose-toxicity model.

Step 2. Choosing of a target toxicity level. By target toxicity level, we can describe what percentage of the investigated patients would be acceptable to have dose-limiting toxicity (DLT). In oncology trials, investigating chemotherapeutic agents which may cause serious side-effects and usually applied in short treatment period, the target toxicity level is typically chosen to be between 0.2 and 0.3. However, if the purpose of the study is to examine the clinical response and efficacy rate of a drug, target toxicity level can be chosen form a wider range, e.g. $[0.3, 0.9]$. In Fig. 2 (gray solid line), the target toxicity level is 0.5, meaning that it is acceptable that 50% of the patients have DLT.

Step 3. Dose levels and mapping. Physiologically relevant dose levels should be chosen for the dose-toxicity model. A typical choice for dose levels is calculated by using the modified Fibonacci sequence. In this case, $dose1$ is chosen based on preliminary data, and the next doses are calculated as follows: $dose2 = 2\cdot dose1, dose3 = 1.67\cdot dose2, dose4 = 1.5\cdot dose3, dose5 = 1.4\cdot dose4, dose6 = 1.33\cdot dose5, dose7 = 1.33\cdot dose6, dose8 = 1.33\cdot dose7$. Finally, the correspond-

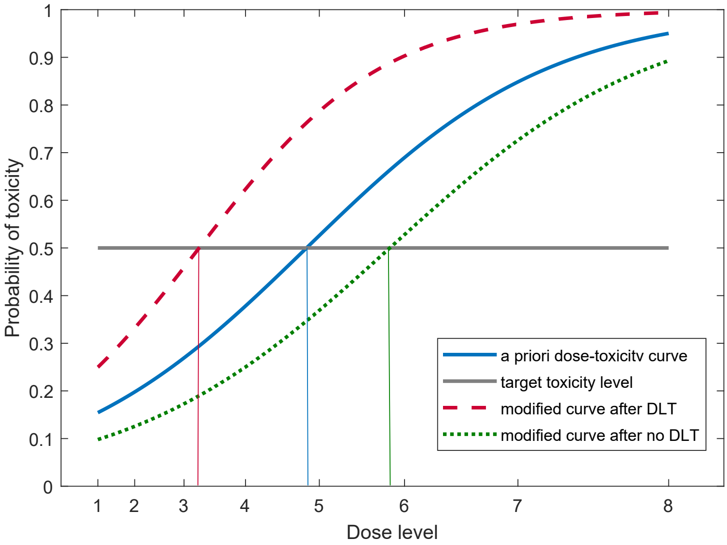


Figure 2
Dose-toxicity models for Continual Reassessment Method.
The *a priori* dose-toxicity curve is shifted up if at least one of the patients from the previous cohort experienced dose-limiting toxicity; if no patient from the previous cohort experienced DLT, the curve is shifted down. The treatment dose of the next cohort is the closest following dose level to the intersection of the dose-toxicity curve and the target toxicity level.

ing toxicity risks should be estimated for every mapped dose value. In Fig. 2 (x axis), modified Fibonacci sequence-based mapped dose levels are shown.

Step 4. Find the optimal starting dose. The optimal starting dose should be chosen based on the intersection of the *a priori* dose-toxicity curve and the target toxicity level. The starting dose is the closest following dose level to the intersection. In Fig. 2, the optimal starting dose is $dose5$.

Step 5. Re-estimation of model parameters of the dose-toxicity curve. Using the optimal starting dose, a given number of patients are treated in the first cohort. Based on the observed toxicity data from this cohort, and using the *a priori* dose-toxicity model, parameters of the original dose-toxicity curve are re-estimated. This method applies the Bayesian approach, i.e. statistically combines *a priori* assumptions with observed data. As a result, the dose-toxicity curve is shifted up or down based on whether the patients experienced DLT in the given cohort or not. Finally, using the updated dose-toxicity curve, the treatment dose of the next cohort can be calculated. From cohort to cohort, as the number of patients involved in the trial is increasing, the dose-toxicity curve is almost only estimated from the observed data, the originally chosen *a priori* dose-toxicity model is substantially changing. In Fig. 2, dashed red curve shows the modified dose-toxicity model if at least one of the patients from the previous cohort

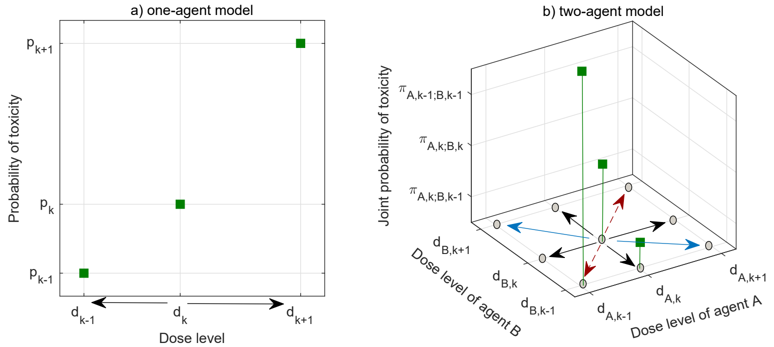


Figure 3
Probability of toxicity as a function of dose level in dose-finding trials.
In one-agent models, there are maximum two adjacent doses for a given dose level, and the probability of toxicity is monotonically increasing as the drug dose is increased. In two-agent models, there are eight adjacent doses; diagonal movements where doses are not changing in the same direction (blue solid arrow) are allowed, but diagonal movements where both doses are changing in the same direction (red dashed arrow) are not allowed. The monotonic order of toxicity is not guaranteed in the case of multi-agent models, the joint toxicity probability is unknown.

experienced DLT (the curve is shifted up, meaning that doses are presumably associated with higher toxicity risks). The treatment dose of the next cohort in this case is *dose*4. In contrast, green dotted curve in Fig. 2 represents the case when the dose-toxicity model was updated due to no patient from the previous cohort experienced DLT (the curve is shifted down, meaning that doses are presumably associated with lower toxicity risks). The treatment dose of the next cohort is *dose*6.

Step 6. Stopping CRM and finding MTD. After each cohort, dose escalation or de-escalation takes place as it is described in Step 5. CRM stops when a predefined stopping criterion is met. In a typical stopping criterion, the total number of the patients who have been treated at a given dose (during the different cohorts) is specified, and an additional condition could be that the next cohort would give the same dose level. This dose is the MTD that was being sought.

3 Bayesian Adaptive Dose-Finding Method for Multi-Agent Trials

Beside the traditional frequentist biostatistical designs, a specific model-based design group, namely Bayesian methods gain more and more importance. Bayes' theorem establishes the relationship between the conditional probability of A given B with the conditional probability of the reverse, i.e. B given A [17]. The advantages of the use of Bayesian biostatistics in clinical oncology are manifold [18–20]. On the one hand, *a priori* knowledge can be incorporated into the trial design and complex statistical methods can be expeditiously handled.

On the other hand, a probability can be assigned directly to the efficiency of the treatment. Also, Bayesian methods have the capacity to naturally integrate evidence from multiple sources [21] and Bayesian methods can provide better results by minimizing risk and maximizing utility [22]. Besides this, Bayesian methods can be used in minimum effective dose (MinED) finding problems as well [23]. However Bayesian methods have significant computational complexity, it is not an obstacle anymore due to modern computing power and available software [24]. Ewings et al. discusses a practical recommendations for implementing a Bayesian adaptive phase I design during a pandemic using AGILE trial which is a randomised seamless phase I/II trial platform [25].

Other important question in Phase I trials is the fact that in most of the cases, oncology protocols recommend multi-modal therapies where a given combination of drugs is used. In these cases, it should be determined which drug is causing the observed toxicity, which is a significant challenge [26].

As we have discussed previously, in single-agent dose-finding trials the main assumption is that the probability of toxicity is monotonically increasing as the drug dose is increased. For a given dose level, there are maximum two adjacent doses where – based on the dose-finding algorithm – the current dose can be escalated or de-escalated (Fig. 3 a) one-agent model), and the order of toxicity level corresponding to the new dose is known (i.e. is it higher or lower than the previous one).

In contrast, using a two-agent model [27, 28], the doses span a 2-dimensional space where for a given dose, there are eight adjacent doses, including diagonal movements when both doses are changing in one step (Fig. 3 b) two-agent model). Such diagonal movements where both doses are changing in the same direction (i.e. both agent doses are increasing or both are decreasing in one step) is not allowed [29]. A special case in two-agent models is when a discrete dose space is used, i.e. several doses of one agent are fixed. A solution for this case can be the parsimonious working model for the dose–toxicity relationship where the aim is to find the MTD of the other agent to be used in combination with each of the doses of agent one [30, 31].

The main problem using multiple agents is that monotonicity of the dose-toxicity curve is not an always valid assumption, namely the monotonic order of toxicity is not guaranteed [32, 33], the joint toxicity probability is unknown, hence deciding on dose escalation or de-escalation is not trivial [34].

3.1 Copula regression model

Yin et al. [29] published a copula-type drug combination regression model where *a priori* information comes from trials in which drugs were investigated individually as single agents. The developed copula-type Bayesian adaptive dose-finding method reduces to the Continual Reassessment Method when a single-agent is investigated.

The (individual) toxicity probability of agent A in the case of the jth dose (A_j) is p_j, and the investigated sequence is $p_1 < p_2 < ... < p_j < ... < p_J$, where

p_J is the toxicity probability of the MTD of agent A (i.e. A_J). Similarly, the (individual) toxicity probability of agent B in the case of the kth dose (B_k) is q_k, and the investigated sequence is $q_1 < q_2 < ... < q_k < ... < q_K$, where q_K is the toxicity probability of the MTD of agent B (i.e. B_K). The individual toxicity probabilities are known. As it was mentioned before, however the individual toxicity probabilities are ordered, the joint toxicity probabilities are not trivially ordered; for instance the relationship between $\pi_{j,k}$ (joint toxicity probability of (A_j, B_k)) and $\pi_{j-1,k+1}$ (joint toxicity probability of (A_{j-1}, B_{k+1})) is not known.

In the next step, a power parameter is assigned to the *a priori* toxicity probabilities in order to reduce the uncertainty of the probabilities; the ”true” toxicity probabilities are p_j^α and q_k^β, where $\alpha > 0$ and $\beta > 0$ are unknown parameters with prior means centered at 1. To calculate the joint toxicity probabilities, the following conditions have to be satisfied:

- if $p_j^\alpha = 0$ and $q_k^\beta = 0 \Rightarrow \pi_{j,k} = 0$,
- if $p_j^\alpha = 0 \Rightarrow \pi_{j,k} = q_k^\beta$; and if $q_k^\beta = 0 \Rightarrow \pi_{j,k} = p_j^\alpha$,
- if either $p_j^\alpha = 1$ or $q_k^\beta = 1 \Rightarrow \pi_{j,k} = 1$,

where $j = 1, ..., J$ and $k = 1, ..., K$.

In a copula-type model, the joint toxicity probability distribution can be calculated using the marginal distributions and a dependence parameter. The dependence function in the Archimedean copula family is

$$C_\gamma(u, v) = \psi_\gamma \left\{ \psi_\gamma^{-1}(u) + \psi_\gamma^{-1}(v) \right\}, \tag{4}$$

where $0 \leq u, v \leq 1$, and γ is an association parameter, C_γ is a distribution function on $[0,1]^2$, and ψ_γ is the copula generator with the following properties: $0 \leq \psi_\gamma \leq 1$, $\psi_\gamma(0) = 1$, $\psi_\gamma' < 0$ and $\psi_\gamma'' > 0$. Taking into account a specific type from the Archimedean copula family (Clayton copula), the proposed regression model copula is

$$\pi_{j,k} = 1 - \left\{ \left(1 - p_j^\alpha\right)^{-\gamma} + \left(1 - q_k^\beta\right)^{-\gamma} - 1 \right\}^{-\frac{1}{\gamma}}, \tag{5}$$

where $\gamma > 0$ describes the drug–drug interaction. This model is a multivariate generalization of the Continual Reassessment Method, allowing internal learning from other combinations of dose levels.

In the case of multi-agent models, target toxicity level has an intersection curve with the joint toxicity probability surface, this curve defines the required maximum tolerated dose. As a consequence, there could be more than one discrete MTD solution. The final MTD combination should be selected based on the recommendation of medical experts.

The likelihood function can be calculated based on a binomial distribution

$$L(\alpha,\beta,\gamma|\text{data}) \propto \prod_{j=1}^{J}\prod_{k=1}^{K} \pi_{j,k}^{x_{j,k}}(1-\pi_{j,k})^{n_{j,k}-x_{j,k}}, \tag{6}$$

where $n_{j,k}$ represents the patients who are treated with (j,k) dose level combination, and $x_{j,k}$ represents the patients who experienced dose-limiting toxicity.

Assuming independent *a priori* distributions, viz. $f(\alpha,\beta,\gamma) = f(\alpha)f(\beta)f(\gamma)$, joint posterior distribution is

$$f(\alpha,\beta,\gamma|\text{data}) \propto L(\alpha,\beta,\gamma|\text{data})f(\alpha)f(\beta)f(\gamma). \tag{7}$$

After each cohort, Gibbs sampler is used to find the unknown parameters, and hence the $\pi_{j,k}$ joint toxicity probability can be calculated, on which the dose escalation or de-escalation decision for the following cohort can be done.

The dose-finding algorithm has the following steps (denotations: ϕ is the target toxicity level, c_e is the probability cut-off for dose escalation, c_d is the probability cut-off for dose de-escalation, $c_e + c_d > 1$):

- Starting the first cohort: patients are treated with the lowest combination: (A_1,B_1)
- Start-up rule in order to obtain reliable posterior estimates:
  - first, the dose of agent A is fixed, while the dose of agent B is continuously increased based on the predescribed sequence, until the first DLT is experienced: $\{(A_1,B_2),(A_1,B_3),...,(A_1,B_{DLT})\}$
  - second, the dose of agent B is fixed, while the dose of agent A is continuously increased based on the predescribed sequence, until the first DLT is experienced: $\{(A_2,B_1),(A_3,B_1),...,(A_{DLT},B_1)\}$
  - if one patient experiences DLT in both agents, the start-up period is finished
- Investigating joint toxicity probabilities:
  - if $P\left(\pi_{j,k} < \phi\right) > c_e$, then dose escalation takes place to an adjacent dose combination where the corresponding joint toxicity probability is higher than the current one; if the current dose combination is (A_J,B_K) (viz. the individual MTD for both agents), no more dose escalation takes place, dose combination stays at the same level
  - if $P\left(\pi_{j,k} > \phi\right) > c_d$, then dose de-escalation takes place to an adjacent dose combination where the corresponding joint toxicity probability is lower than the current one; if the current dose combination is (A_1,B_1) (viz. the lowest dose for both agents), no more dose de-escalation takes place, the trial is terminated

- otherwise (when no cut-off dose for escalation or de-escalation is reached), the next cohort continues with the same dose combination
- when the predefined maximum cohort size is reached, the trial ends; the dose combination which has the closest value to the target toxicity level is set to be the joint MTD for the investigated agents

For an integrated Bayesian Phase I/II adaptively randomized oncology trial design based on the copula model, see [35].

3.2 Logistic regression model

Riviere et al. [2] proposed a drug combination–toxicity relationship logistic regression model

$$\text{logit}(\pi_{j,k}) = \beta_0 + \beta_1 u_j + \beta_2 v_k + \beta_3 u_j v_k, \tag{8}$$

where $\pi_{j,k}$ is the joint toxicity probability of a two agent drug combination, β_1 is the the toxicity effect of agent A, β_2 is the the toxicity effect of agent B, and β_3 is the interaction between the two agents ($\beta_0 ... \beta_3$ are unknown parameters). Variable u_j represents the standardized dose of the jth level of agent A, v_k is the standardized dose of the kth level of agent B. Standardized doses are defined individually (as if they are administered as a single-agent) using the *a priori* estimates of the toxicity probabilities of the jth dose level of agent A (p_j) and the kth dose level of agent B (q_k)

$$u_j = \log \frac{p_j}{1-p_j} \tag{9}$$

$$v_k = \log \frac{q_k}{1-q_k}. \tag{10}$$

In this model, the likelihood function is a product of the Bernoulli probabilities

$$L(\beta_0, \beta_1, \beta_2, \beta_3 | \text{data}) \propto \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{j,k}^{x_{j,k}} (1-\pi_{j,k})^{n_{j,k}-x_{j,k}}, \tag{11}$$

where $n_{j,k}$ represents the patients who are allocated at combination (j,k), and $x_{j,k}$ represents the patients who experienced dose-limiting toxicity.

The posterior distribution is sampled using Gibbs sampler, and the *a posteriori* toxicity probabilities are estimated using Monte Carlo simulation

$$\tilde{\pi}_{j,k} = \frac{1}{L} \sum_{l=1}^{L} \frac{\exp\left(\beta_0^{(l)} + \beta_1^{(l)} u_j + \beta_2^{(l)} v_k + \beta_3^{(l)} u_j v_k\right)}{1 + \exp\left(\beta_0^{(l)} + \beta_1^{(l)} u_j + \beta_2^{(l)} v_k + \beta_3^{(l)} u_j v_k\right)}, \tag{12}$$

where $\left(\beta_0^{(l)}, \beta_1^{(l)}, \beta_2^{(l)}, \beta_3^{(l)}\right)_{l=1,...,L}$ are the L posterior samples, assuming that β_0 and β_3 are normal *a priori* distributions ($N(0,10)$), and β_1 and β_2 are exponential *a priori* distributions (Exp(1)).

The dose-finding algorithm is the one that was proposed in the copula regression model [29]; however, MTD level is found in a different way. Target toxicity level is extended to a target toxicity interval using parameter δ

$$\phi_{interval} = [\phi - \delta; \phi + \delta], \tag{13}$$

where ϕ is the target toxicity level.

Using the target toxicity interval, the *a posteriori* densities of the toxicity probability can be divided into three groups:

- if the toxicity probability is in the $[0; \phi - \delta]$ interval, the corresponding cumulative density is the probability of under-dosing,
- if the toxicity probability is in the $[\phi - \delta; \phi + \delta]$ interval, the corresponding cumulative density is the probability of target toxicity,
- if the toxicity probability is in the $[\phi + \delta; 1]$ interval, the corresponding cumulative density is the probability of over-dosing (for a specific solution of the over-dosing problem, see e.g. [11] where the Bayesian Optimal Interval (BOIN) design is introduced).

For each (A_j, B_k) dose combination, the probability of being in the targeted toxicity interval can be calculated. The dose combination that has the highest *a posteriori* probability, and have been used previously to treat at least one cohort of the patients, should be chosen as MTD

$$(A_{MTD}, B_{MTD}) = \max\left\{P\left(\pi_{j,k} \in [\phi - \delta; \phi + \delta]\right)\right\}. \tag{14}$$

For another Bayesian dose-finding method which use logistic regression model, see e.g. [36] where the approach allows the inclusion of covariates.

4 Bayesian Dose-Finding Methods for Trials with Delayed Toxicities

Administering different radiations to the patients, late-onset toxicities can be observed which affect the dose escalation or de-escalation decisions. In order to conduct a complete follow-up after each cohort, in some cases several weeks or even months are required. Late-onset toxicity is an important problem in the non-conventional cancer therapies like Targeted Molecular Therapies (TMTs) [37].

4.1 Time-To-Event Continual Reassessment Method (TITE-CRM) for single-agent trials

The time-consuming nature is a strong limit to the use of Continual Reassessment Method in the case of late-onset toxicities. A short-cut for this problem is to allow patients to enter to the trial in a staged fashion, which extends the CRM to Time-To-Event Continual Reassessment Method (TITE-CRM) that can handle late-onset toxicities [38, 39].

In the original CRM, the decision of the dose level of the next cohort can be formulated as

$$F\left(d_{n+1},\hat{p}_n\right)-\phi \leq F\left(d_k,\hat{p}_n\right)-\phi \text{for } k=1,...,K, \tag{15}$$

where $F(d,p)$ is the dose-toxicity model, $d_1,...,d_K$ are the dose levels, p is the probability of toxicity, n is the number of observations, $\hat{p}_n$ is an estimate of p, and ϕ is the target toxicity level.

The likelihood function in this case is

$$L_n(p)=\prod_{i=1}^{n} F\left(d_i,p\right)^{y_i}\left\{1-F\left(d_i,p\right)\right\}^{1-y_i}, \tag{16}$$

where y_i is the indicator of toxic response for the ith patient.

In the TITE-CRM, there is a weighted dose-toxicity model $G(d,\omega,p)$ that has the following properties: $G(d,0,p)=0$ and $G(d,1,p)=F(d,p)$, where ω $(0 \leq \omega \leq 1)$ is a function of the time-to-event of a patient, and it is linear in F. The dose escalation or de-escalation decisions are the same as in the original CRM (i.e. (15)), but the likelihood is weighted as well

$$\tilde{L}_n(p)=\prod_{i=1}^{n} G\left(d_i,\omega_{i,n},p\right)^{y_{i,n}}\left\{1-G\left(d_i,\omega_{i,n},p\right)\right\}^{1-y_{i,n}}, \tag{17}$$

where $y_{i,n}$ is the indication of toxic response for the ith patient prior to the entry time of the $(n+1)$th patient, and $\omega_{i,n}$ is the corresponding weight. Using TITE-CRM, patients who have not experienced DLT are weighted by the proportions of their follow-up times compared to the full period of the trial, and patients who have experienced DLT are weighted by 1.

4.2 Late-onset toxicity model for multi-agent trials

Liu et al. [37] proposed a late-onset toxicity model for multi-agent trials using the Bayesian data augmentation approach, treating the late-onset toxicity as missing data. The dose-toxicity model is described by the Finney model

$$\text{logit}(\pi_{j,k})=\beta_0+\beta_1\log\left(a_j+\rho b_k+\gamma(a_j\rho b_k)^{\frac{1}{2}}\right), \tag{18}$$

where β_1 is the slope of the regression ($\beta_1>0$), ρ is the relative potency of agent B versus agent A to induce toxicity (if $\rho>1 \Rightarrow$ agent B is more likely to cause toxicity than agent A), and γ is the synergy-antagonism parameter describing the drug-drug interaction between the agents ($\gamma<0 \Rightarrow$ antagonism effect, $\gamma=0 \Rightarrow$ dose additivity effect, $\gamma>0 \Rightarrow$ synergy effect). In this model, β_0, β_1, γ and ρ are unknown parameters. The Finney model reduces to the standard logistic model when a single-agent is investigated.

If no late-onset toxicity takes place, the toxicity outcomes are fully observed and hence the complete-data likelihood function can be described for the ith patient as

$$L(\theta|y) = \prod_{i=1}^{n} \frac{\exp\left\{y_i\beta_0 + y_i\beta_1 \log\left(a_{j_i} + \rho b_{k_i} + \gamma\left(a_{j_i}\rho b_{k_i}\right)^{\frac{1}{2}}\right)\right\}}{1+\exp\left\{\beta_0 + \beta_1 \log\left(a_{j_i} + \rho b_{k_i} + \gamma\left(a_{j_i}\rho b_{k_i}\right)^{\frac{1}{2}}\right)\right\}}, \quad (19)$$

where n is the total number of patients, y_i is the binary toxicity outcome, a_{j_i} is the jth dose of agent A for the ith patient, and b_{k_i} is the kth dose of agent B for the ith patient. The *a posteriori* distribution of $\theta = (\beta_0, \beta_1, \gamma, \rho)$ in this case is

$$f(\theta|y) \propto f(\theta)L(y|\theta), \quad (20)$$

where $f(\theta)$ is the *a priori* distribution of θ.

If late-onset toxicities take place, the toxicity outcomes are not fully observed due to the missing binary toxicity outcome values. These missing values can be handled using data augmentation which contains two iterative steps:

- imputation (I):
  - in this step, the missing data is imputed by drawing samples from their posterior predictive distribution using Bernoulli probability

    $$f(y_i|t_i > s_i, \theta) = \text{Bernoulli}\left(P\left(y_i = 1|t_i > s_i, \theta\right)\right), \quad (21)$$

    where t_i is the time to toxicity for the ith patient, and s_i is the actual follow-up time;
- posterior (P):
  - in this step, the posterior samples of unknown parameters are simulated based on imputed data
  - here – due to the imputation of the missing data in the previous step – the standard Markov chain Monte Carlo method [26] can be used as in the case of complete-data when no late-onset toxicity takes place.

Iteration-based techniques are well-known not only in model-baesd dose-toxicity modelsbut in fixed point, iteration-based controls as well [40].

5 Discussion and Conclusion

Finding the maximum tolerated dose in Phase I oncology clinical trials is an important and not trivial problem. Beside the safety criterion of the patients (viz. avoiding over-dosing), cost-effectiveness viewpoints should be taken into account as well (e.g. avoiding unnecessary under-dosing experiments), and sometimes even extreme circumstances such as a pandemic [25].

A promising solution of the dose-finding problem is the use of Bayesian methods. In every case, the method is based on an *a priori* dose-toxicity model which gives a preliminary estimation of the toxicity probability for the investigated drug dose levels. In the following steps, these *a priori* assumptions are statistically combined with observed data. The trials consist of cohorts; from cohort to cohort, dose escalation or de-escalation takes place. The trial ends when a pre-defined stopping criterion is met, and maximum tolerated dose can be found which results in relatively acceptable dose-limiting toxicity.

For the simplest dose-finding problem, viz. single-agent trials with fully observed data, the most common solution is the Continual Reassessment Method. In this case the toxicity is monotonically increasing as the drug dose is increased. However, using multiple agents, the monotonic order of toxicity is not guaranteed, the joint toxicity probabilities are unknown, and as a consequence, dose escalation or de-escalation decision is not trivial.

In this paper, two models for multiple agents have been discussed: the copula regression model and the logistic regression model. Both models estimate the joint toxicity probability distribution and define a likelihood function. The dose escalation or de-escalation decision is made after Gibbs sampling, but MTD level is found in different ways in respect of the two methods.

Another incremental problem can be the presence of late-onset toxicities. For a single-agent problem, the use of Time-To-Event Continual Reassessment Method can handle the problem by allowing patients to enter to the trial in a staged fashion. Taking into account multi-agent trials, the Bayesian data augmentation approach can be applied which treats the late-onset toxicity as missing data, and missing values can be handled using data imputation and simulation of posterior samples.

Besides the above discussed Bayesian methods, dose-finding criteria can be calculated using other approaches like toxicity and efficacy odds ratios [41]. In this case, acceptable doses satisfy the following two conditions:

$$\Pr(p_j < \overline{\pi}_T) > p^*, \tag{22}$$

$$\Pr(q_j > \underline{\pi}_E) > q^*, \tag{23}$$

where $\overline{\pi}_T$ is a pre-defined upper toxicity limit, $\underline{\pi}_E$ is pre-defined lower efficacy limit, and p^* and q^* are fixed probability cutoffs.

Dose j has p_j toxicity probability and q_j efficacy probability. Taking into account two-dimensional toxicity and efficacy domain, we expect that p_j and q_j are the closest values to the lower-right corner $(1,0)$. The horizontal and vertical lines which cross point $A(q_j, p_j)$ split the domain into four rectangles.

After that, the odds ratio between the toxicity and efficacy of dose j can be calculated:

$$\omega_j^{(2)} = \frac{p_j/(1-p_j)}{q_j/(1-q_j)} = \frac{p_j(1-q_j)}{(1-p_j)q_j}. \tag{24}$$

Note that $\omega_j^{(2)}$ is exactly the ratio of the lower-right versus the upper-left rectangle's area. In this way, an equivalent odds ratio contour can be defind: along the curve, all the points have the same toxicity-efficacy odds ratio, namely $\omega_j^{(2)}$.

Furthermore, this two-dimensional probability space can be extended by a third scale, where the new axis is the probability of efficacy given no toxicity. Hence in this three-dimensional domain, the toxicity-efficacy odds ratio trade-offs are arranged with an efficacy value given no toxicity. Compared to the two-dimensional domain, there not an equivalent odds ratio contour, but an equivalent odds ratio surface is defined. All the points on this smooth surface have the same odds ratio, $\omega_j^{(3)}$. Based on this, one can find the best dose to treat the patients in the next cohort.

References

[1] A. E. Chang, P. A. Ganz, D. F. Hayes, T. Kinsella, H. I. Pass, J. H. Schiller, R. M. Stone, and V. Strecher. *Oncology: an evidence-based approach*. Springer Science & Business Media, 2007.

[2] M. K. Riviere, Y. Yuan, F. Dubois, and S. Zohar. A Bayesian dose-finding design for drug combination clinical trials based on the logistic model. *Pharmaceutical Statistics*, 13(4):247–257, 2014.

[3] J. Sápi, D. A. Drexler, and L. Kovács. Potential benefits of discrete-time controller-based treatments over protocol-based cancer therapies. *Acta Polytechnica Hungarica*, 14(1):11–23, 2017.

[4] R. Liu, Y. Yuan, S. Sen, X. Yang, Q. Jiang, X. Li, C. Lu, M. Göneng, H. Tian, H. Zhou, R. Lin, and O. Marchenko. Accuracy and safety of novel designs for phase i drug-combination oncology trials. *Statistics in Biopharmaceutical Research*, 14(3):270–282, 2022.

[5] Y. Yuan, J. Lee, and S. G. Hilsenbeck. Model-assisted designs for early-phase clinical trials: Simplicity meets superiority. *JCO Precision Oncology*, 3(PO.19.00032), 2019.

[6] H. Zhou, T. Murray, H. Pan, and Y. Yuan. Comparative review of novel model-assisted designs for phase I clinical trials. *Stat Med.*, 37(14):2208–2222, 2018.

[7] G. Yin. *Clinical trial design: Bayesian and frequentist adaptive methods*, volume 876. John Wiley & Sons, 2012.

[8] A. Hirakawaa, N. A. Wages, H. Sato, and S. Matsui. A comparative study of adaptive dose-finding designs for phase I oncology trials of combination therapies. *Stat Med.*, 34(24):3194–3213, 2015.

[9] X. Chen, R. He, X. Chen, L. Jiang, and F. Wang. Optimizing dose-schedule regimens with bayesian adaptive designs: opportunities and challenges. *Frontiers in Pharmacology*, 14, 2023.

[10] Y. Yuan, R. Lin, and J. J. Lee. *Model-Assisted Bayesian Designs for Dose Finding and Optimization – Methods and Applications*. Chapman and Hall/CRC, 2022.

[11] Y. Yuan, K. R. Hess, S. G. Hilsenbeck, and M. R. Gilbert. Bayesian optimal interval design: A simple and well-performing design for Phase I oncology trials. *Clinical Cancer Research*, 22(17):4291–4301, 2016.

[12] R. Ananthakrishnan, R. Lin, C. He, Y. Chen, D. Li, and M. LaValley. An overview of the BOIN design and its current extensions for novel early-phase oncology trials. *Contemporary Clinical Trials Communications*, 28:100943, 2022.

[13] H. Pan, R. Lin, Y. Zhou, and Y. Yuan. Keyboard design for phase i drug-combination trials. *Contemporary Clinical Trials*, 92:1551–7144, 2020.

[14] E. Garrett-Mayer. The continual reassessment method for dose-finding studies: a tutorial. *Clinical Trials*, 3(1):57–71, 2006.

[15] O. John and C. Mark. Continual reassessment and related dose-finding designs. *Stats (Basel)*, 25(2):202–216, 2010.

[16] G. Wheeler, A. Mander, and A. e. a. Bedding. How to design a dose-finding study using the continual reassessment method. *BMC Medical Research Methodology*, 9(18), 2019.

[17] L. D. Broemeling. *Bayesian biostatistics and diagnostic medicine*. CRC Press, 2007.

[18] Y. Zhang, B. Guo, S. Cao, C. Zhang, and Y. Zang. SCI: A Bayesian adaptive phase I/II dose-finding design accounting for semi-competing risks outcomes for immunotherapy trials. *Pharm Stat.*, 21(5):960–973, 2022.

[19] R. Mu, H. Pan, and G. Xu. SCI: A Bayesian adaptive phase I/II platform trial design for pediatric immunotherapy trials. *Biometric Methodology*, 4(2):382–402, 2021.

[20] K. Messer, L. Natarajan, E. D. Ball, and T. A. Lane. Toxicity-evaluation designs for phase I/II cancer immunotherapy trials. *Statistics is Medicine*, 7-8:712–720, 2010.

[21] M. Adamina, G. Tomlinson, and U. Guller. Bayesian statistics in oncology. *Cancer*, 115(23):5371–5381, 2009.

[22] K. Kelly, S. Halabi, and R. L. Schilsky. *Oncology Clinical Trials: Successful Design, Conduct and Analysis*. Demos Medical Publishing, 2009.

[23] R. Mu, G. Xu, G. Liu, and H. Pan. A two-stage Bayesian adaptive design for minimum effective dose (MinED)-based dosing-finding trials. *Contemporary Clinical Trials*, 108(106504), 2021.

[24] N. Muehlemann, T. Zhou, and R. e. a. Mukherjee. A tutorial on modern bayesian methods in clinical trials. *Ther Innov Regul Sci*, 57:402–416, 2023.

[25] S. Ewings, G. Saunders, and T. e. a. Jaki. Practical recommendations for implementing a bayesian adaptive phase I design during a pandemic. *BMC Medical Research Methodology*, 22(25), 2022.

[26] D. Dejardin. *Statistical Models for the Analysis of Oncology Endpoints*. KU Leuven Biostatistics and Statistical Bioinformatics Centre (L-BioStat), 2013.

[27] P. Thall, R. Millikan, P. Mueller, and S. Lee. Dose-finding with two agents in phase i oncology trials. *Biometrics*, 59(3):487–496, 2003.

[28] S. J. Mandrekar, Y. Cui, and D. J. Sargent. An adaptive phase i design for identifying a biologically optimal dose for dual agent drug combinations. *Stat. Med*, 26(11):2317–2330, 2007.

[29] G. Yin and Y. Yuan. Bayesian dose finding in oncology for drug combinations by copula regression. *Royal Statistical Society: Series C (Applied Statistics)*, 58(Part 2.):211–224, 2009.

[30] K. Wang and A. Ivanova. Two-dimensional dose finding in discrete dose space. *Biometrics*, 61(1):217–222, 2005.

[31] W. Zhao and H. Yang. *Statistical Methods in Drug Combination Studies*. Chapman and Hall/CRC, 2014.

[32] M. R. Conaway, S. Dunbar, and S. D. Peddada. Designs for single- or multiple-agent phase I trials. *Biometrics*, 60(3):661–669, 2004.

[33] M. A. Diniz, S. Kim, and M. Tighiouart. A Bayesian adaptive design in cancer phase I trials using dose combinations with ordinal toxicity grades. *Stats (Basel)*, 3(3):221–238, 2022.

[34] K. Hashizume, J. Tshuchida, and T. Sozu. Flexible use of copula-type model for dose-finding in drug combination clinical trials. *Stats (Basel)*, 78(4):1651–1661, 2022.

[35] Y. Yuan and G. Yin. Bayesian Phase I/II adaptively randomized oncology trials with combined drugs. *Annals of Applied Statistics*, 5(2A):924–942, 2011.

[36] S. Bailey, B. Neuenschwander, G. Laird, and M. Branson. A Bayesian case study in oncology Phase I combination dose-finding using logistic regression with covariates. *Journal of Biopharmaceutical Statistics*, 19:469–484, 2009.

[37] S. Liu and J. Ning. A Bayesian dose-finding design for drug combination trials with delayed toxicities. *Bayesian Analysis*, 8(3):703–722, 2013.

[38] Y. K. Cheung and R. Chappell. Sequential designs for Phase I clinical trials with late-onset toxicities. *Biometrics*, 56(4):1177–1182, 2000.

[39] Y. Zhang, S. Cao, C. Zhang, I. H. Jin, and Y. Zang. A Bayesian adaptive phase I/II clinical trial design with late-onset competing risk outcomes. *Biometric Methodology*, 73(3):796–808, 2021.

[40] I. Lovas. Fixed point, iteration-based, adaptive controller tuning, using a genetic algorithm. *Acta Polytechnica Hungarica*, 19(2):59–77, 2022.

[41] G. Yin, Y. Li, and Y. Ji. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics*, 62(3):777–784, 2006.

A Hybrid Prediction Fault Location Model for Copper Wire Manufacturing Process

Robert Agyare Ofosu[1], Huangqiu Zhu[1]*, Benjamin Odoi[2]

[1]School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China, e-mail: raofosu@umat.edu.gh; zhuhuangqiu@ujs.edu.cn

[2]Faculty of Engineering, University of Mines and Technology, WT-0038-7367, Tarkwa, Ghana, e-mail: bodoi@umat.edu.gh

Abstract: This paper presents a novel prediction of fault location during copper wire manufacturing using a hybrid Nonlinear Autoregression Neural Network (NARNN) and Markov chain model. A four (4) year daily primary data spanning from 2018 to 2022 consisting of 261502 data points obtained from a cable manufacturing company in Ghana was used for the prediction. A comparison between the suggested model and decision tree algorithm was done. To assess the predictive effectiveness of the two models, performance indicators including Mean Absolute Deviation (MAD), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) were used. As determined by their evaluation criteria, the findings revealed that the suggested hybrid model had superior data fitting and accurate prediction capabilities.

Keywords: tension; prediction; wire breaks; fault location; NARNN; decision tree; markov chain model

1 Introduction

Electrical cables are made all over the world due to their widespread use. Basically, electrical cables are current-conducting wires either copper or aluminum that have been twisted, braided, or bonded into a single assembly with or without insulation. They are employed specifically for the transmission of electrical or telecommunication signals. Electrical cables are necessary because they serve as the foundation for the functionality of all electrical equipment. To produce cables that are of good quality, a number of production processes must be used. One of these essential manufacturing processes is the drawing stage, which entails drawing a copper or aluminum rod through a series of progressively smaller synthetic diamond or tungsten carbide dies. Drawing reduces the rod to a wire with the required diameter that has great surface quality and enhanced mechanical qualities like strength and hardness [1]-[3].

A typical copper rod is drawn in four steps. These include unwinding, drawing, annealing and rewinding. The thick copper rod is unwound at the payoff drive to the diesing chamber during the unwinding phase. The ultimate wire diameter is produced in the diesing chamber by forcing the thick copper wire through a sequence of progressively smaller dies. Lubricant, like oil, is circulated in the diesing chamber to lower friction and wire wear. The drawn copper wire is put through a specific heat treatment in the annealing chamber to soften it and increase its malleability. A dancer system is also mounted between the payoff and take-up drives so as to sense and regulate the wire tension to the desired limit. The drawn wire is evenly coiled onto a bobbin at the take-up with the aid of the traverse. A counter installed on the moving web next to the dancer provides data on the length of wire wound onto the bobbin. Figure 1 depicts the process of copper wire drawing [4].

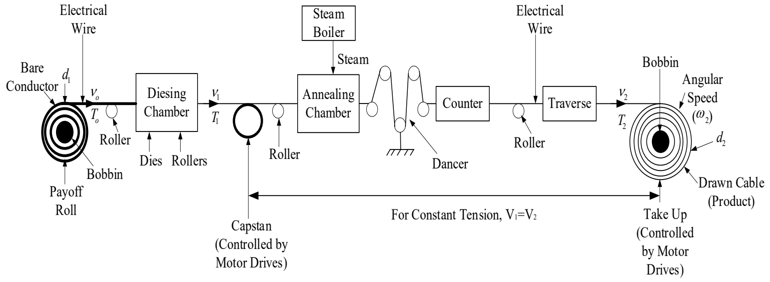


Figure 1

Copper wire drawing process schematic diagram

where, d_1 is the wire's initial diameter before drawing, d_2 is actual wire's diameter upon drawing, $V_0=V_1$ is line speed of pay-off or capstan, V_2 is line speed of take up, $T_0=T_1$ is torque of payoff or capstan, T_2 is torque of take up, ω_2 is angular speed of take-up drive

After the wires are drawn, they are bundled to form a cable, which is then extruded with insulating polymers such as Cross-Linked Polyethylene, Polyethylene, and Polyvinyl Chloride. Extrusion is an important stage in the production line since it inhibits copper losses in cables and protects the conductor from physical harm and environmental hazards. After extrusion, the final product is ready for market consumption after passing quality control tests [3], [5]. Among the most crucial things to take into consideration in the cable manufacturing sector is the final resistance or diameter of the wire, which should not differ substantially from the standard after production. That is, the wire's tension should remain consistent during the drawing stage of the cable production process so that the wire's diameter or cross-sectional area remains unchanged. Failure to ensure this can lead to fire outbreaks that could kill innocent lives and destroy millions of properties as a result of heat generation when substandard wires or cables are used [6]. According to research, tension variation during the wire drawing process is primarily caused by faults in the drawing machines' components. In most cases, this causes machines to abruptly stop, resulting in wire breaks caused by stretching the wire beyond its tensile strength [7], [8].

Similarly, in the event of a fault, determining the exact location of the fault in wire drawing machines is always a bigger challenge for the experts who work on these machines, as it usually takes much longer to locate the faults in order to restore them. Most cable manufacturing industries face significant challenges as a result of this phenomenon, which causes increased downtime, production losses, energy waste, and scrap production [9]. Several factors contribute to wire tension and breakage during the drawing process. Low drawing or annealing solution concentration, entanglement in the basket, bad or copper dust in dies, annealing bearing failure, improper drive and tensioner setting, oxidation of wire due to bad steam flow in the annealer, power outage, torque or speed variation due to changes in roll diameter between the payoff drive and the take-up drive among other factors. These faults are mostly located in the drawing chamber, annealing chamber, capstan, recirculatory system, dancer, payoff, take-up, AC drive and traverse. Therefore, developing a model that can predict the location of faults during the copper wire drawing process is necessary in order to avert these challenges.

Several models for fault prediction have been reported in the literature. These models are categorized as statistical, physical, and artificial intelligence models [10]. Physical models suffer from multiple iterations before achieving the desired results. Furthermore, these models demand a significant amount of reliable data [11]. Seasonal Autoregressive Integrated Moving Average (SARIMA), Auto-Regressive Moving Average (ARMA), Generalized Autoregressive Conditional Heteroscedasticity (GARCH) and Auto-Regressive Integrated Moving Average (ARIMA) are the most frequently employed statistical models [12]. It has been demonstrated that these models can predict if the time series data exhibit a linear relation. Their strength is based on historical data. However, because statistical models have several transitory periods and large variability, they are unable to produce reliable predictions for time series with nonlinearities, such as the wire break location prediction in the drawing machines [13].

Artificial Intelligence (AI) models such as Support Vector Machine (SVM), Logistic Regression (LR), Adaptive Neuro Fuzzy Inference System (ANFIS), Artificial Neural Networks (ANNs), Decision Trees, Fuzzy Logic, and k-Means are well recognized for their capacity to resolve a nonlinear time series with greater precision and have produced favorable outcomes in the creation of extremely precise fault diagnostic systems [14-19]. Among the AI techniques, ANN is one of the most popular and several studies have shown that it outperforms other techniques [20], [21]. The advantages of ANN are enormous because it is fault-tolerant, can learn sophisticated nonlinear relationships, and has powerful classification attributes. Besides, due to the non-parametric nature of ANN prediction, having process knowledge of the production of the time series is not necessary. Furthermore, once trained, ANNs are capable of making accurate predictions [22]. Nonetheless, ANN lack coherence, leading to their inappropriateness for deployment in instances under which it is essential to determine which elements did contribute to a technical fault [21], [22].

There are various subcategories of ANN. These include Radial Basis Function Neural Network (RBFNN), Backpropagation Neural Network (BPNN), Recurrent Neural Network (RNN), Generalized Regression Neural Network (GRNN), and Nonlinear Autoregressive Neural Network (NARNN). Because the prediction of wire break location in the cable manufacturing process involves fluctuating parameters that are highly nonlinear and complex, research has shown that NARNN has the capability to predict the dynamics of this complex system with high precision and quick convergence [23], [24]. NARNN has been utilized successfully in a variety of applications for fault prediction. Some of which include transformer oil dissolved gas concentration [12], fault prediction in software [25], rolling element bearing deterioration prediction [26], prediction of infiltration of underground water by hydraulic fluid leaks [27], engine fault detection and failure prediction in the manufacturing process [28], [29], passenger flow forecasting [30], meteorological time series forecasting [31], [32], Heating Ventilation and Air Conditioning (HVAC) predictions [23], [33] geomagnetic fluctuations prediction [34], prediction of COVID-19 cases [35] among other complex dynamical systems.

To the author's best knowledge, the hybrid NARNN-Markov chain model has never been used to predict fault location in wiring drawing machines and no literature has ever reported on the prediction of wire breaks and fault location in drawing machines during copper wire drawing in the cable manufacturing industries. As a result, this paper proposed a novel approach to accurately classify and locate the probability of wire breaks in the copper wire drawing process by combining NARNN with a Markov chain model. To determine the model's effectiveness in fault location, a comparative analysis was performed using the decision tree algorithm. The findings clearly show the proposed models' effectiveness in predicting fault locations. Therefore, it is anticipated that this research can be the basis for the deployment of corrective maintenance in the cable manufacturing industry. Thus, the operators can anticipate potential fault locations and implement recommended preventive measures.

1.1 Nonlinear Autoregressive Neural Network

The Nonlinear Autoregression Neural Network (NARNN) blends neural network techniques' capacity for matching nonlinear function with that of autoregressive methods for unearthing probable time series sources [12], [32]. The architecture is developed and provided with training in an open loop, with the intended target variables making up the feedback loop to ensure higher training accuracy.

The configuration is changed to a closed loop after learning, and the estimated outputs are utilized as new signals acting as feedback to the network. The NARNN is a nonlinear, discrete, autoregressive model used in forecasting time series data expressed as [26], [27]:

$$y(t) = h(y(t-1),\ y(t-2) + ... + y(t-d)) + \varepsilon(t) \tag{1}$$

where *y(t)* is prediction's outcome at a discontinuous step time *t* of the time series *y*, the series' past data is denoted by *d*, *ε(t)* symbolizes the series y's deviation at time step *t* and h connotes a hypothetical nonlinear quantity that the feedforward part of a neural network can predict while being trained.

The aim for training a neural network is to estimate the functional h(.) by maximizing the bias and weights of the network. As a result, the NARNN model is well-defined by Eq. (2) [26], [27].

$$y(t) = a_0 + \sum_{j=1}^{k} a_j \phi\left(\sum_{i=1}^{a} \beta_{ij} y(t-i) + \beta_{oj} \right) + \varepsilon(t) \tag{2}$$

where *a* depicts the entry number, *k* denotes the hidden layer's quantity containing activation function ϕ, variable β_{ij} determines how strongly the input layer *i* and the hidden layer *j* are connected. The values for the output and hidden layer, respectively, are a_o and β_{oj}, a_j is the connecting weight linking the output and the hidden layer.

The boosting of the NARNN model's architecture necessitates the identification of the quantity of hidden layers, time delays, and activation function, as well as a suitable learning technique. The desirable amount of time delay and hidden layers is determined by experiment. On the basis of Dandy and Maier, the activation function is selected. Finally, due to their accuracy and high convergence speed, Bayesian regularization algorithms and the Levenberg-Marquardt are utilized to train the model [26], [27]. Generally, the input data quality, universality and size, as well as proper model development and assessment, are critical to the successful application of NARNN models [28], [29].

1.2 Decision Tree Method

A kind of supervised learning is the decision tree approach. It is one of the most widely used classification techniques due to its high accuracy and low computational cost. Its flexibility, nonparametric nature, and capacity to deal with nonlinear relationships between features and classes make it suitable for fault classification [36]. The decision-rules model's tree structure formation is based on if/else instructions. In theory, an iterative binary partition approach is used for the desired output to supervise the training sets. To divide the sample space, successive queries with yes/no options are posed. The locations in which the items are examined are referred to as nodes. The test findings are subsequently relayed to a branch.

A decision tree has three different kinds of nodes. These include the internal nodes, leaf nodes, and root nodes. The test's result is based on each node's purity. After achieving the optimum value of purity, the node is terminated. The optimum value is established if a node only produces one kind of output. When classifying new samples, the decision tree and an item quantity will be examined. The chain of attribution from the root node to the leaf node maintains group forecast for the tested samples. The basic procedure in creating a decision tree is to identify the attribute that will be evaluated on a node, and an ancillary node to that node. Splitting refers to the entire process of identifying test and branch.

The process of splitting reduces the dataset's impurity that corresponds to class at a subsequent point. The task necessitates the computation of information gain, which is divided into entropy and the entropy splitting index. An indicator of entropy or impurity *i(t)* at a given node *t*, represents the entropy index as shown in Eq. (3) [37].

$$i(t) = -\sum_{j=1}^{k} p(w_j \mid t) \log p(w_j \mid t) \tag{3}$$

where $p(w_j \mid t)$ denotes the pattern's proportion assigned to a kind at node t.

The optimal splitting basic values x_j^R for the variable x_j are used to segregate a node that has not terminated into right child nodes tR and left child node tL. P_R and P_L make up the equivalent fractions of the new entities. Eq. (4) optimizes the difference by providing the most efficient entropy splitting index.

$$\Delta i(t) = i(t_p) - P_R i(t_L) - P_L i(t_R) \tag{4}$$

1.3 Markov chain Model

One variety of stochastic process is the Markov chain, which is widely used to analyze dynamic systems. The process is random, in which any future data exists in the current state. Furthermore, the probability and state transition matrix are important elements in implementing the Markov chain model. Unlike similar predictive techniques, the Markov chain model is simple to implement and neither does it necessitate a profound comprehension of changes in system dynamics. As a result, it is comparably simple to comprehend the data [38]. The Markov chain approach consists of five stages [38], [39]. The stages are:

Stage 1) Process state definition for the Markov chain.

Stage 2) Develop the state transition probability, P and state transition matrix, N. The Markov chain's state transition matrix, N, denotes the measured number of times of switching between states, as demonstrated in Eq. (5).

$$N = \begin{bmatrix} n_{11} & \cdots & n_{1s} \\ \vdots & \ddots & \vdots \\ n_{s1} & \cdots & n_{ss} \end{bmatrix} \tag{5}$$

n_{ij} indicates the amount of sequential transitions between states i and j

Suppose *P* represents a transition matrix that expresses all of the Markov chain model's transition probabilities for each state. *P* can therefore written as;

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1s} \\ \vdots & \ddots & \vdots \\ p_{s1} & \cdots & p_{ss} \end{bmatrix}, i, j \in I \tag{6}$$

Then,

$$P\{X_{t+1} = j \mid X_t = i\} = p_{ij} \tag{7}$$

The probability of one step is described by Eq. (7). A homogenous Markov chain refers to transition probabilities that change independently with time *t*.

Therefore,

$$P\{X_{t+1} = j \mid X_t = i\} = P\{X_1 = j \mid X_0 = i\} = p_{ij} \tag{8}$$

Non-negative entries in each row summing up to unity are required by the matrix P. Hence,

$$0 \le p_{ij} \le 1 \text{ and } \sum_{j=1}^{t} p_{ij} = 1, \ \sqrt{i} \in I \tag{9}$$

State *i* to state *j* in k-steps probability of state transition is defined by Eq. (10).

$$p_{ij}(k) = P\{X_{n+k} = j \mid X_n = i\}, \ \sqrt{k} > 0, n \ge 0, \ \text{i}, j \in I \tag{10}$$

Eq. (11) describes the transition matrix P.

$$P(n) = P^{n-1} \times P = P^n \tag{11}$$

Stage 3) Ergodic Markov chain validation.

The constrained distribution's occurrence in an ergodic Markov chain must be confirmed by categorizing the P's state. The three parts are the irreducible Markov chain, the periodicity Markov chain, and the recurrent and transitory states.

Step 4) Probability values of Markov process

For probability values of Markov process, it is possible to calculate the mean return time and stationary probability distribution. For an ergodic Markov chain, the maximum allocation for a stationary probability distribution exists and is represented as:

$$\pi_j = \lim_{n \to 0} P(X_n = j \mid X_0 = i) \tag{12}$$

Thus

$$\pi_j = P_j(n) = \sum_k P_k(n-1)P_{kj} \text{ becomes } \pi j = \sum_k \pi_k P_{kj}, \quad \text{as n} \to \infty \text{ for j=0, 1,2,....} \tag{13}$$

Step 5) Model validation and Forecasting

To compute the forecasts, the base probability and state transition probability can be employed through Eq. (14).

$$P(S_j) = \sum_{i=1}^{n} P(S_i)P_{ij} \tag{14}$$

where P_{ij} depicts state transition probability and $P(S_i)$ the base probability.

Based on the assumption of independence, the Markov chain's authenticity is examined using the Chi-square test during the model validation process as shown in Eq. (15) [40].

$$X^2_{calculated} = \sum \frac{(Observed - Expected)}{Expected} \tag{15}$$

If $X^2_{calculated}$ is higher than $X^2_{tabulated}$ on the 0.05 crucial zones, the null hypothesis is refuted.

1.4 Error Metrics

The training and testing errors were used to assess the fitness and performance of prediction of the NARNN and decision tree models. The errors were evaluated using three indices: Mean Absolute Deviation (MAD), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) [41], [42].

1.4.1 Root Mean Square Error

The test of the dispersion of predicted errors over real data sets is the Root Mean Square Error (RMSE). In other words, the RMSE elucidates how close an estimated model's forecasted values are to the actual data points. The formula is given as:

$$RMSE_{forecast} = \sqrt{\sum_{i=1}^{n} \left(\frac{Y_t - \hat{Y}_t}{n} \right)^2} \tag{16}$$

where $\hat{Y}_t$ indicates the prediction, Y_t the real data sets, and n the size of the sample.

1.4.2 Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is a percentage size measurement of a forecast's error. It is used to evaluate forecast accuracy and is expressed as:

$$MAPE_{forecast} = \left(\frac{1}{n} \sum \frac{\left| Y_t - \hat{Y}_t \right|}{\left| Y_t \right|} \right) \times 100\% \tag{17}$$

1.4.3 Mean Absolute Deviation

Mean Absolute Deviation (MAD) is the most fundamental indicator of prediction performance. MAD describes how large an error from the estimation is anticipated on mean as described by Eq. (18).

$$MAD = \left(\frac{1}{n} \sum \left| Y_t - \hat{Y}_t \right| \right) \tag{18}$$

The following is a discussion of the remaining sections: Section 2 explains the research approach, including data collection and preparation, the utilization of hybrid NARNN and Markov Chain model on the data and assessing the model's efficiency using statistical metrics. Explanation of the findings are given in Section 3. Section 4 concludes with some quick observations and suggestions.

2 Methods Used

2.1 Data Collection

A four-year period of daily primary data spanning 2018 to 2022 consisting of wire diameter (mm), length (m), number of spools, total length (km), number of wire breaks, and wire break rate as the independent variables and location of wire breaks as the dependent variable with a total data point of 261502 for each of the variables was used for the study. This data was taken at the wire drawing machines in a cable manufacturing industry in Ghana. The data was merged, cleaned, and organized for the analysis process.

2.2 Hybrid NARNN and Markov Chain Construction

The implementation process for the NARNN model is shown methodically in the flowchart in Figure 2. To start, during the data pre-processing stage, the simulation's data was partitioned into training (80%) and testing (20%) datasets.

The NARNN weights and bias were subsequently initialized with random numbers. Every iteration of their values was adjusted using Levenberg–Marquardt Back-Propagation (LMBP). The goal was to attain a target error with the fewest possible iterations. The last phase involves assessing the NARNN's performance and predicting capacity using the test data.

To predict the fault location on the basis of the output of NARNN, the performance is evaluated based on their probabilities. Figure 3 depicts the prediction stage of the located faults during the wire drawing stage of the cable manufacturing process. The variables *y(t-1), y(t-2), y(t-3) and y(t-4)* represents the input variables for instance the total wire length, number of spools, number of wire breaks and wire break rate. The location of wire breaks such as the drawing chamber, annealing chamber, dancer, and capstan, among others, is represented by the predicted output *y(t)*.

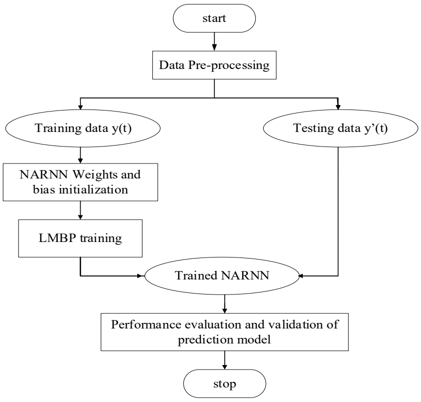

start
Data Pre-processing
Training data y(t)
Testing data y'(t)
NARNN Weights and bias initialization
LMBP training
Trained NARNN
Performance evaluation and validation of prediction model
stop

Figure 2
Flowchart of NARNN Model

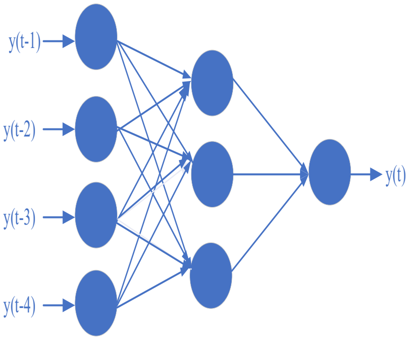


Figure 3
Fault Location Model of the NARNN

After the NARNN had classified the various faults, the Markov chain model was used to predict the probabilities of the faults occurring within a given location and the probability of faults occurring in other locations based on their pivot points. Firstly, the number of faults occurring within a given location was converted into a transitional matrix. Afterwards, a discrete-time Markov chain was created. The results of the Markov chain prediction were then displaced in the form of a chain displaying their transitional probabilities at various locations.

2.3 Construction of the Decision Tree Algorithm

The construction of the decision tree algorithm was achieved using Minitab Statistical Software version 21.1.0 after the dataset was loaded into the software. A decision tree plot and other statistical parameters were then recorded.

2.4 Simulations

The NARNN-Markov chain model and decision tree algorithm were checked for relevance, robustness, and inaccuracies using simulations of the dependent and independent variables. Owing to the NARNN's lack of interpretability of the fault location, the Markov chain was used to predict the probability of fault locations.

3 Results and Discussion

3.1 Exploratory Analysis

A summary of the variables taken into consideration in this research is shown in the Table 1. The average values of the variables considered in the descriptive statistics are given by 13583, 7.6646, 99.32, 0.7834, and 14.02 for length, number of spools, total length, number of wire breaks, and break rate, respectively. The Standard Error of the Mean (SE Mean) was also 192, 0.0819, 2.61, 0.0252, and 0.850 for length, number of spools, total length, number of wire breaks, and break rate. The smaller the standard error, the lower the risk. Likewise, the kurtosis was also reported as 126.66, 0.35, 694.37, 6.43, and 171.88 respectively for length, number of spools, total length, number of wire breaks, and break rate. From a statistical point of view, the variance is not constant, hence the data is heteroskedastic based on the kurtosis of the variables length, total length, number of wire breaks, and break rate being greater than 3.

Table 1
Exploratory Data Analysis

Variable	Mean	Standard Error of Mean	Kurtosis
Length	13583	192	126.66
Number of spools	7.6646	0.0819	0.35
Total length	99.32	2.61	694.37
Number of wire breaks	0.7834	0.0252	6.42
Break rate	14.02	0.850	171.88

3.2 Statistics of Faults at Various Locations

The statistics of wire breaks at various drawing machine locations between 2018 and 2022 are displayed in Figure 4. From the collected data, it can be seen that the drawing chamber had the highest number of wire breaks of 763. This was closely followed by the capstan with 339 wire breaks. The next significant wire break was observed in the annealing chamber, with 276 wire breaks in this location. The number of unspecified wire break locations was reported as 475. Among all this break locations, the traverse was observed to be the list with a value of 1. The data unambiguously demonstrates that, as is typically found in most cable manufacturing companies, the drawing chamber, capstan, annealing chamber, poor basket coiling as a result of entanglement, and the dancer are the major locations where there is frequently occurring wire breakage. The crucial steps in the wire drawing process take place at these locations. The most common faults in the various locations occur due to wire tension as a result of low drawing or annealing solution concentration, entanglement in the basket, inadequate solution flow on dies, bad dies, wire locking in machines, rough surfaces on rollers, annealing bearing failure, improper setting of the drives and tensioner, copper dust in dies, rough surface of contact band, oxidation of wire due to bad steam flow in the annealer, power outages, torque or speed variation due to changes in roll diameter between the payoff drive and the take-up drive, and changing line speed in the drawing stage of the cable, among other factors. Hence, the need to predict the location where faults are likely to occur to help experts easily identify faults when they occur.

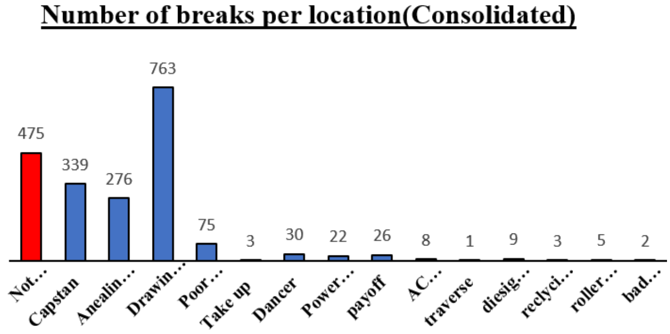


Figure 4

Number of breaks per location

3.3 Trend Analysis

To determine the nature of the data set, a trend analysis was conducted. The linear, exponential and quadratic models was considered. However, the quadratic model was the best among them as shown in Figures 5, 6, 7. Again, a measure of accuracy of the model thus MAPE, MAD and MSD were used as the error metric to determine the model's accuracy. It was discovered that for all the variables considered such as the wire break rate, number of wire breaks and total length the MAD had the least value of 18.47, 0.9311and 47.1 respectively and therefore the best value statistically.

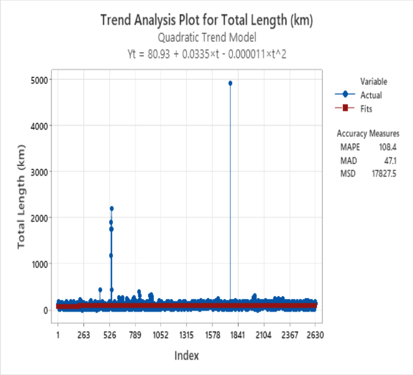


Figure 5

Quadratic trend model for Total Length

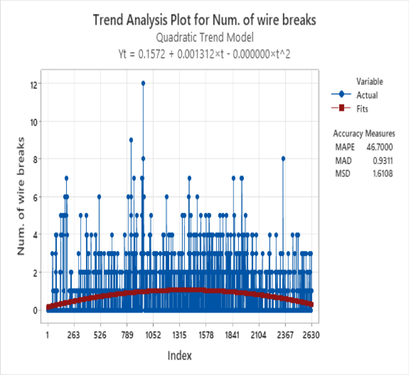


Figure 6

Quadratic trend model for Number of Wire breaks

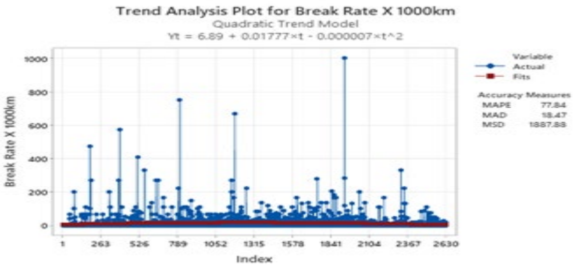


Figure 7

Quadratic trend model for break rate

3.4 Optimal Decision Tree Results

In all, there are four (4) nodes, namely nodes 1 through 4, and four different patterns to achieve an optimal solution as depicted in Figure 8. To achieve an optimal solution based on the mean and standard deviation from the nodes, it could be observed that the system started at node 1 and branched to node 2 and node 4, with a termination at node 4, with a mean and standard deviation of 272.711 and 192.197, respectively. Compared to node 2 and terminal node 4, it can be concluded that node 2 had the optimal values with the least mean and standard deviation of 12.5334 and 36.2605 respectively compared to that of terminal node 4. At terminal node 2, there was another branch to determine an optimal value, namely terminal node 1 and node 3. Finally, at node 3, there was a branch with final termination at terminal nodes 2 and 3. From the analysis, it can be seen that at all the terminal points, terminal node 1 had the best optimal value of 0.546871 and 3.02142, respectively, for the mean and standard deviation. This implies that terminal node 1 is the best model to predict the exact location of the faults during the wire drawing process.

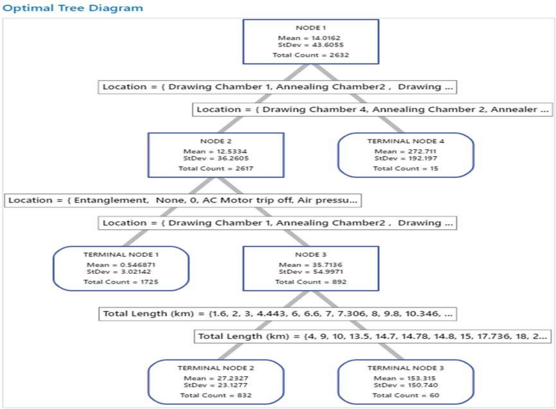


Figure 8
Optimal decision tree diagram

3.5 Hybrid NARNN and Markov Chain Predictive Model

In determining the location of fault during the wire drawing process, the variables that were considered were total length, number of spools, and break rate, and the possible fault locations were the drawing chamber, capstan, payoff, annealing chamber, dancer, bad coiling basket, power outage, winder drive, AC motor trip, and entanglement, with the number of faults recorded in the various locations. Hence, in order to predict the fault occurring within a given location, the NARNN was used to effectively locate the fault condition, as shown in Figure 9. The NARNN diagram was used to determine the pattern of signal in faults per the location in a short time so as to indicate the specific location of faults.

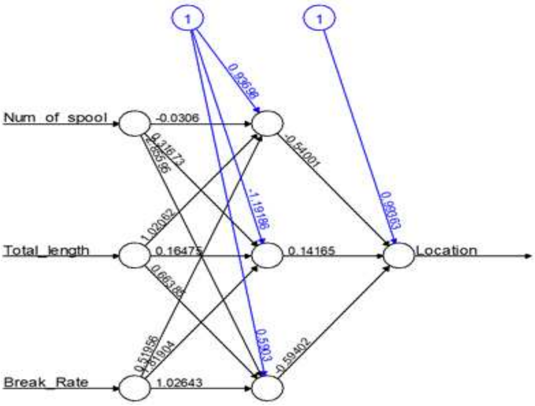


Figure 9
NARNN Predicted Fault Location

To predict the faults from one location to the other based on their transitional probabilities, the Markov chain model was used as shown in Figure 10. This was to determine the probability at which a fault can occur in another location.

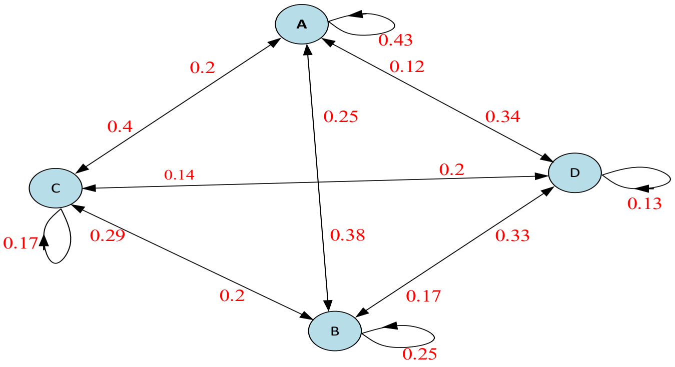


Figure 10
Markov Chain Transitional Probability Diagram

The transitional probability diagram has four pivots, A, B, C, and D, representing the fault locations: drawing chamber, payoff drive, take-up drive, and traverse. The transitional probabilities of fault location occurring at the various pivots are 0.43, 0.25, 0.17, and 0.13, respectively. The transitional probability of a fault from the drawing chamber to the traverse is 0.34, which corresponds to the annealing chamber's transitional probability. Accordingly, there is a 34% likelihood of the fault being located at the annealing chamber from the drawing chamber to the traverse. Once more, the transitional probability of faults from the payoff drive to the take-up drive is 0.29, which matches to the dancer's transitional probability. As a result, there is a 29% possibility that the fault is located at the dancer from the payoff drive to the take-up drive.

Similarly, the transitional probability of faults from the drawing chamber to the payoff drive is 0.38, which represents the transitional probability of the capstan. This indicates that there is a 38% chance that the fault will be found at the capstan linking the drawing chamber to the payoff drive. Additionally, the transitional probability of faults from traverse to payoff drive is 0.17, which is the transitional probability of the recycling system. It translates to a 17% likelihood that the fault will be found at the recycling system between the traverse and the payoff. Furthermore, the diesing chamber's transitional probability, which is given by the transitional probability of faults from the take-up to the traverse, is 0.2. In other words, there is a 20% chance that the fault will be found in the diesing chamber from the take-up to traverse.

These findings are consistent, with distinct transitional probabilities, across different fault locations. Hence, there is a high probability of faults being located at Pivot A, which is the drawing chamber with a high transitional probability. If a fault is detected in the drawing chamber, it can also damage the annealing chamber, capstan, and dancer, which are essential components of the drawing machine. Therefore, extra attention should be paid to the drawing chamber during the production process to minimize faults from affecting other locations of the drawing machine.

3.6 Predictive Performance Indicators of the Models

Table 2 depicts the performance indicators of the decision tree algorithm and the NARNN. To measure the model's accuracy for the training data using the decision tree, it was observed that the MAPE had the lowest value of 1.2191 and was hence the best model. This was followed by MAD with a value of 8.6312. The RMSE had the highest value of 24.4991, which clearly depicts that the RMSE cannot be considered the best indicator when evaluating the model's accuracy. In the same way, the R-square value recorded using the decision tree was 65.66%, indicating that the model is adequate. The same trend was also observed in the testing.

Again, the training results from the NARNN depict that the MAD had the lowest value of 0.01153, which shows that it is the best indicator for measuring the accuracy of the model. This was closely followed by RMSE with a value of 0.4285496 with the worse indicator among the error metrices being MAPE with a value of 1.250. It was also observed that the testing data mimicked the training data, with MAD and RMSE as the best indicators, respectively. In comparison with the decision tree, it can be deduced that the NARNN had the best model based on the error matrices considered in this study.

Table 2

Error Metric Indicators

Error Metrics	Decision Tree		NARNN	
	Training	Testing	Training	Testing
RMSE	24.4991	43.1110	0.4285496	0.5832
MAD	8.6312	10.7418	0.01153	0.7134
MAPE	1.2191	0.9811	1.2500	2.5312
R-squared	65.66%	17.59%		

3.7 Predicted results of Research Data

Figures 11, 12, 13, and 14 show the predicted graphs of the research variables such as total length, number of wire breaks, number of spools, and break rate for the years 2018 to 2024 based on the NARNN model.

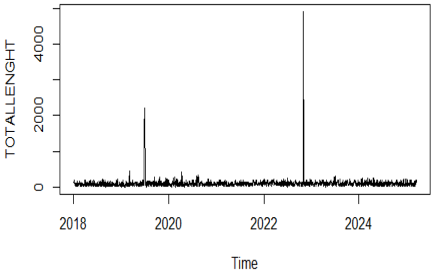


Figure 11
Predicted Results for Total Length

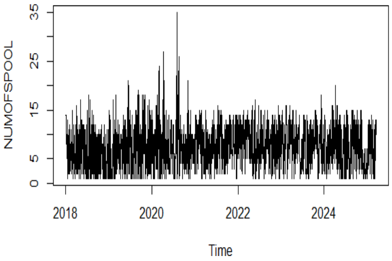


Figure 12
Predicted Results for Number of Spools

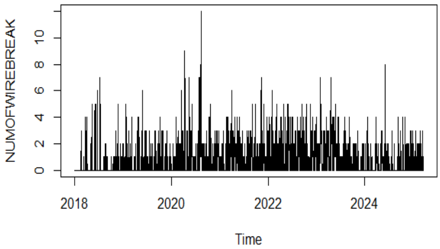


Figure 13
Predicted Results for Number of Wire Breaks

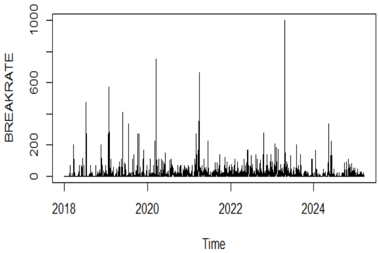


Figure 14
Predicted Results for Wire Break Rate

In summary, the research findings indicate that the hybrid model produced a highly predictive result and smooth transitional probabilities between fault locations when compared to the classical techniques reported in [12], [26], [35], [36].

Conclusions

In summary, the prediction of wire break location during the cable drawing process has been achieved. It was observed that the hybrid NARNN and Markov chain model could effectively predict the break location with higher accuracy when compared to that of the decision tree. One key finding of this research is that no literature has reported on prediction of wire break location in the cable manufacturing industry, and the application of NARNN and the Markov chain model to wire break location prediction further enhances the novelty of this research. It is recommended that cable manufacturers adopt the proposed model for easy fault prediction during the wire drawing process so as to reduce downtime, minimize scrap production, and improve the efficiency and quality of manufactured cables. Subsequent research will center on utilizing machine learning techniques to predict fault locations in copper wire manufacturing.

Acknowledgement

The authors thank Jiangsu University, China for providing the resources needed to finish this research work.

References

[1] Tasevski, G., Petreski, K. A study on the tuner roll impact on the wire drawing process. International journal of industrial engineering and technology, Vol. 6 (2), 2016, pp. 17-22

[2] Verma, S., Sudhakar, R. P. Design and analysis of process parameters on multistage wire drawing process- a review. International journal of mechanical and production engineering research and development, Vol. 9(1), 2019, pp. 403-412

[3] Ofosu, R. A., Normanyo, E., Obeng, L. Temperature control of heaters in cable extrusion machine using PSO-ANFIS controller. In proceedings of IEEE AFRICON Conference, 2019, pp. 1-9

[4] Larsson, J., Jansson, A., Karlsson, P. Monitoring and evaluation of the wire drawing process using thermal imaging. International journal of advanced manufacturing technology, Vol. 3, 2018, pp. 1-14

[5] Mahto, P. K., Murmu, R. Temperature control for plastic extrusion process. International journal of innovative research in science, engineering and technology, Vol. 4(7), 2015, pp. 5748-5758

[6] Jie-Shiou, L., Ming-Yang, C., Ke-Han, S., Mi-Chi, T. Wire tension control of an automatic motor winding machine-an iterative learning sliding mode control approach. Robotics and computer–integrated manufacturing, Vol. 50, 2018, pp. 50-62

[7] Perduková, D., Fedor, P., Fedák, V., Padmanaban, S. Lyapunov based reference model of tension control in a continuous strip processing line with multi-motor drive. Electronics, Vol. 8 (60), 2019, pp. 1-24

[8] Zhewei, G., Sheng, Z., Kaijie, Z., Chenliang, S. Fully-digital tension control system with PID algorithm for winding ultrafine enameled wires. In IOP Conference Series: Materials Science and Engineering, Vol. 892, 2020, p. 012064

[9] Huang, P. Y., Cheng, M. Y., Su, K. H., Kuo, W. L. Control of roll-to-roll manufacturing based on sensorless tension estimation and disturbance compensation. Journal of the chinese institute of engineers, Vol. 44(2), 2021, pp. 89-103

[10] Li, Y., Wang, Z., Zhe, L., Jiang, Z. Research on Fault Prognosis Methods Based on Data-driven : A Survey. In IOP Conference Series: Materials Science and Engineering, Vol. 1043, 2021, p. 042008

[11] Xing, D., Haijian, S., Chunlong, H., Dengbiao, J., Yingtao, J. Wind power forecasting methods based on deep learning: A Survey. Computer Modeling in Engineering and Sciences, Vol. 122(1), 2020, pp. 273-301

[12] Pereira, F. H., Bezerra, F. E., Shigueru, J., Josemir, S., Chabu, I., de Souza, G. F. M., Micerino, F., Nabeta, S. I. Nonlinear autoregressive neural

network models for prediction of transformer oil-dissolved gas concentrations. Energies, Vol. 11(7), 2018, pp. 1-12

[13] Shao, H., Deng, X., Jiang, Y. A novel deep learning approach for short-term wind power forecasting based on infinite feature selection and recurrent neural network. Journal of renewable and sustainable energy, Vol. 10, 2018, pp. 1-12

[14] Fernandes, M., Corchado, J. M., Marreiros, G. Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review. Applied Intelligence, Vol. 52, 2022, pp. 14246-14280

[15] Cheng, L., Yu, T. Dissolved gas analysis principle-based intelligent approaches to fault diagnosis and decision making for large oil-immersed power transformers: A Survey. Energies, Vol. 2018(11), 2018, pp. 1-13

[16] Radu-Emil, P., Claudia-Adina, B. D., Elena-Lorena, H., Raul-Cristian, R., Emil, M. P. Evolving fuzzy models of shape memory alloy wire actuators, Romanian Journal of Information Science And Technology. Vol. 24(4), 2021, 353-365

[17] Radu-Emil, P., Gheorghe, D., Sergey, T., Inga, Z. Processing, neural network-based modeling of biomonitoring studies data and validation on republic of moldova data. In the Romanian Academy, Series A-Mathematics Physics Technical Sciences Information Science. Vol. 23(4), 2022, pp. 403-410

[18] Shahnaz, N. S., Dursun, E. An input-weighted, multi-objective evolutionary fuzzy classifier, for alcohol classification. Acta Polytechnica Hungarica. Vol. 197(10), 2022, 61-81

[19] Ofosu, R. A., Asiedu-Asante, A. B., Adjei, R. B. Fuzzy logic based condition monitoring of a 3-phase induction motor. In proceedings of IEEE AFRICON 2019, pp. 1-8

[20] Blanchard, T., Samanta, B. Wind speed forecasting using neural networks. Wind Engineering, Vol. 44(1), 2020, pp. 33-48

[21] Mihalache, S., Burileanu, D. Speech emotion recognition using deep neural networks, transfer learning and ensemble classification techniques. Romanian Journal of Information, Science and Technology, Vol. 26 (3-4), 2023, pp. 375-387

[22] Li, Z., Wang, Y., Wang, K. S. Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. Advances in Manufacturing, Vol. 5(4), 2017, pp. 377-387

[23] Ruiz, L. G. B., Cuéllar, M. P., Calvo-Flores, M. D., Jiménez, M. D. C. P. An application of non-linear autoregressive neural networks to predict

energy consumption in public buildings. Energies, Vol. 2016,(9), 2016, pp. 1-21

[24] De Giorgi, M. G., Ficarella, A., Quarta, M. Dynamic performance simulation and control of an aeroengine by using NARX models. In Proceedings of MATEC Web conference, Vol. 304, 2019, pp. 1-8

[25] Chatterjee, S., Nigam, S., Singh, J. B., Upadhyaya, L. N. Software fault prediction using Nonlinear Autoregressive with eXogenous Inputs (NARX) network. Applied Intelligence, Vol. 37, 2012, pp. 21-129

[26] Nistane, V. M. Wavelet-based features for prognosis of degradation in rolling element bearing with non-linear autoregressive neural network. Australian Journal of Mechanical Engineering, Vol. 19(4), 2021, pp. 423-437

[27] Umit, K., Esra, S. E., Mumine, K. K. Binary anarchic society optimization for feature selection. Romanian Journal of Information Science and Technology, Vol. 26 (3-4), 2023, pp. 351-364

[28] Suvendu, M., Swarup, P., Soudip, H. Artificial neural network coupled condition monitoring for advanced fault diagnosis of engine. Research square, 2021, pp. 1-40

[29] Zajacko, I., Kuric, I., Gal, T. Application of machine learning for failure prediction in manufacturing process. In Proceedings of ISSAT International Conference on Data Science and Intelligent Systems, 2019, pp. 1-5

[30] Ren, G., Gao, J. Comparison of NARNN and ARIMA models for short-term metro passenger flow forecasting. In 19th COTA International Conference of Transportation Professionals, 2019, pp. 1352-1361

[31] Huang, J. F., Lu, W. C. Forecasting of meteorological time series and pricing of weather index rainbow options: A wavelet-NAR neural network model. Systems Engineering-Theory and Practice, Vol. 36(5), 2016, pp. 1146-1154

[32] Adil, A., Muhammad, K. Multi-step ahead wind forecasting using nonlinear autoregressive neural networks. Energy Procedia, Vol. 134, 2017, pp. 192-204

[33] Islam, M. P., Morimoto, T. Non-linear autoregressive neural network approach for inside air temperature prediction of a pillar cooler. International Journal of Green Energy, Vol. 14, 2017, pp. 41-149

[34] Caswell, J. M. A Nonlinear autoregressive approach to statistical prediction of disturbance storm time geomagnetic fluctuations using solar data. Journal of Signal and Information Processing, Vol. 5, 2014, pp. 42-53

[35] Namasudra, S., Dhamodharavadhani, S., Rathipriya, R. Nonlinear neural network based forecasting model for predicting COVID-19 cases. Neural Processing Letters, Vol. 55, 2023, pp. 171-191

[36] Vanfretti, L., Arava, V. S. N. Decision tree-based classification of multiple operating conditions for power system voltage stability assessment. International Journal of Electrical Power and Energy Systems, Vol. 123, 2020, pp. 1-10

[37] Ioan-Daniel, B., Radu-Emil, P., Alexandra-Bianca, B. Improvement of k-means cluster quality by post processing resulted clusters. Procedia Computer Science, Vol. 199, 2022, pp. 63-70

[38] Arican, E., Aydin, T. An RGB-D descriptor for object classification. Romanian Journal of Information Science and Technology, Vol. 25 (3-4), 2022, pp. 338-349

[39] Zhou, Y., Wang, Y., Zhong, L., Tan, R. A markov chain based demand prediction model for stations in bike sharing systems. Mathematical Problems in Engineering, Vol. 2018, 2018, pp. 1-8

[40] Ong, K., Sugiura, B. T., Zettsu, K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. Neural computing and applications. Vol. 27, 2016, pp. 1553-1566

[41] Ofosu, R. A., Odoi, B., Asamoah, M. Electricity consumption forecast for tarkwa using autoregressive integrated moving average and adaptive neuro fuzzy inference system. Serbian Journal of Electrical Engineering, Vol. 18(1), 2021, pp. 75-94

[42] Twumasi-Ankrah, Sampson., Odoi, B., Adoma, W. P, Gyamfi, H. E. Efficiency of imputation techniques in univariate time series. International Journal of Science, Environment, and Technology, Vol. 8(3) 2019, pp. 430-453

Counting of Shortest Paths in a Cubic Grid

Mousumi Dutt[1], Arindam Biswas[2] and Benedek Nagy[3,4]

[1] Department of Computer Science and Engineering, St. Thomas' College of Engineering and Technology, 4, Diamond Harbour Road, Kidderpore, Kolkata-700023, West Bengal, India, mousumi.dutt@stcet.ac.in

[2] Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711103, West Bengal, India, abiswas@it.iiests.ac.in

[3] Department of Mathematics, Eastern Mediterranean University,
W. Shakespeare Street, 99628 Famagusta, North Cyprus, via Mersin-10, Turkey, benedek.nagy@emu.edu.tr

[4] Institute of Mathematics and Informatics, Eszterházy Károly Catholic University, 3300 Eger, Leányka utca 4, Hungary

Abstract: The enumeration of shortest paths in cubic grid is presented herein, which could have importance in image processing and also in the network sciences. The cubic grid considers three neighborhoods — namely, 6-, 18- *and* 26-*neighborhood related to face connectivity, edge connectivity and vertex connectivity, respectively. The formulation for distance metrics is given.* L_1, D_{18}, *and* L_∞ *are the three metrics for* 6-*neighborhood,* 18-*neighborhood and* 26-*neighborhood. The task is to count the number of minimal paths, based on given neighborhood relations, from any given point to any other, in the three-dimensional cubic grid. Based on the coordinate triplets describing the grid, the formulations for the three neighborhoods are presented in this work. The problem both of theoretical importance and has several practical aspects.*

Keywords: cubic grid; shortest paths; combinatorics; path counting; digital distances

1 Introduction

Shortest path (SP) problems have ample applications in digital geometry, which works on discrete spaces, that is, with points with integer coordinates. Based on the application, the shortest path problem can be formulated. Shortest path problems in various grids are defined based on digital distances. In the square grid, there are two classical neighbor relations defined [35] — cityblock and chessboard. The former contains horizontal and vertical movements; in chessboard motion the diagonal movements are also allowed. Consequently, two kinds of distances are defined in this grid, which are well explained in [23, 34]. In the square grid, every coordinate of a point is independent. In *n* dimensional space, there are *n* independent coordinates to address its elements, i.e., usually either the vertices or the hypercubes of the grid. Working in the *n*-dimensional space, the neighborhood structure of the vertices is isomorphic to the neighborhood structure of the *n*-dimensional hypercubes. The scientific field 'Geometry of Numbers' is about these grids [2, 14, 15, 18, 31]. The terms 'tiling', 'array' and 'lattice' are used approximately in the same meaning as we use 'grid' here. Counting paths as an image analysis tool has already been coined in [35], and solved with the cityblock and chessboard paths/distances in [4].

Considering non-traditional but still regular tilings, the triangular grid and the hexagonal grid have the graph-theoretic dual relation. In digital geometry, they have also been analyzed from various points of view. A connection among them and the cubic grid is established [16, 28, 31] and therefore, symmetric coordinate systems with three coordinates work nicely on these grids. The relation between square grid and hexagonal grid is explained in [38]. The three types of neighbor relation on the triangular tiling are already used in [6]. The three coordinates in this grid depend on each other [28] [30]. The digital distance of any two points, based on a fixed neighborhood criterion, is the length of a minimal-length path between the two points, where in every step along the path one moves to a neighbor point [28] [30].

The general Euclidean Shortest Path (ESP) problem is NP-hard [1] between two points amid polyhedral obstacles in the 3D space, moreover there could be exponentially many minimal path classes in single-source multiple-destination problems. A polynomial time algorithm for Euclidean Shortest Path computations, for cases where all the obstacles are convex and their number is small, is stated in [36]. The Euclidean shortest paths within a given cube-curve with arbitrary accuracy are given in [19]. ESP between two points is stated in [20-22] for 2D and 3D using rubberband algorithms. An algorithm to compute an L1-minimal path from any point to any other that lies on or above a given polyhedral terrain is presented in [24].

The discrete version of the problem is somewhat different. In graphs, a dynamic programming approach, namely the Dijkstra algorithm gives an efficient way of computing a shortest path. Digital grids can be seen as infinite graphs, where the neighbor points (pixels, voxels, etc. depending on the dimension of the used space) are connected by edges. Here, we count the number of shortest (also called minimal) paths (NSP), since there usually exist more than one shortest path depending on the conditions and on the used type of paths as a shortest path is generally not unique (similarly as in graphs). On the square grid, for any two points, a recursive formulation for counting the shortest paths between them, in cityblock, in chessboard and in octagonal approaches, is presented in [4]. It is to be noted here, that the general formulation for chessboard shortest paths, between two points was given by a recursive method based on a generating function. Herein, we also give an alternative, non-recursive formulation, based on enumerative combinatorics in

Sec. 3. In [3], NSP between any pair of points, in a digital image, with respect to a particular neighbor criterion is presented, where the images are considered as matrices and thus matrix operations are used in the computation. Shortest isothetic path (cityblock) is determined between two points inside a digital object for a given grid size, in [8] [9]. Since a shortest isothetic path is usually not uniquely determined, finding the number of them is important [7]. Here, in this paper, we will discuss the path counting problem between two points whose coordinate triplets are given in cubic grid for 6- 18-, and 26-neighborhoods, i.e., L_1, D_{18}, and L_1 metrics respectively. The path counting problems in 3D digital geometry for the three neighborhoods are presented in [11] [12] in a different way. The formulation of path counting problem in 26-neighborhood in [11] is based on the generating function stated in [3] [4], whereas in this paper we propose it in a comprehensive and straight forward way. In [12], the formulation for path counting problem in 18-neighborhood is divided into three cases whereas the first two cases are based on the generating function stated in [3, 4] and the third one is based on induction on the length of minimal paths. The formulation for 18-neighborhood proposed here is much simpler and more definite devoid of any generating function. The computation of number of paths is more directly proposed here compared to the formulae in [11] [12]. All these formulae are proved here using combinatorial techniques.

The number of minimal paths (NSP) is related to various descriptive measures of graphs and networks including graph indices. In networking various packages may be sent in different but same length paths and in this way, NSP could be used to measure the width of the network between the given nodes. Thus, our study has not only theoretical interest, but also practical ones due to applicability both in imaging and in networking.

The paper is written in the following structure. The preliminaries are discussed in Section 2. The formulation of NSP in cubic grid for 6-*neighborhoods*, 18-*neighborhoods*, and 26-*neighborhood*s are given in Sections 3, 4, and 5 respectively. Section 6 presents concluding remarks.

2 Preliminaries

Based on [17], the cubic grid on P^3 will be denoted by Z^3, and defined as $Z^3 = \{(c_1, c_2, c_3) \mid c_1, c_2, c_3 \in Z\}$. Let G be any set of points in P^3. The *Voronoi neighborhood* of $g \in G$ is defined as $N_G(g) = \{v \in P^3 \mid \forall h \in G, \|v - g\| \leq \|v - h\|\}$.

The Voronoi neighborhood of (c_1, c_2, c_3) in Z^3 is a unit cube centered in (c_1, c_2, c_3); in this way, the space is tessellated by unit cubes. When perceived as a set of points in P^3, Z^3 is referred to as a *cubic grid*. The Voronoi neighborhoods in a grid in P^3 are referred to, as *voxels*. Figure 1 represents the directions of the three axes in the cubic grid and the origin is also shown.

There are three widely used neighborhoods in Z^3. Those are 26-, 18- and 6-*neighborhood* called face-edge-vertex neighbors, face-edge neighbors and face neighbors, respectively. Let $r = (x_r, y_r, z_r) \in Z^3$ and $s_i = (x_{s_i}, y_{s_i}, z_{s_i}) \in Z^3$ be all the points fulfilling the condition max $\{|x_r - x_{s_i}|, |y_r - y_{s_i}|, |z_r - z_{s_i}|\} \leq 1$:

$$N_6(r) = \{s_i : |x_r - x_{s_i}| + |y_r - y_{s_i}| + |z_r - z_{s_i}| \leq 1\}$$

$$N_{18}(r) = \{s_i : |x_r - x_{s_i}| + |y_r - y_{s_i}| + |z_r - z_{s_i}| \leq 2\}$$

$$N_{26}(r) = \{s_i : |x_r - x_{s_i}| + |y_r - y_{s_i}| + |z_r - z_{s_i}| \leq 3\}$$

These are shown in Fig. 2. The neighbor voxels $N_6(r)$, $N_{18}(r)$, and $N_{26}(r)$ are shown in red (orange, in the figure on the right), magenta, and yellow colors. The voxels which are in 6-neighborhood of r, are face connected with r. The edge connected and vertex connected voxels are in $N_{18}(r)$, and $N_{26}(r)$ of r respectively.

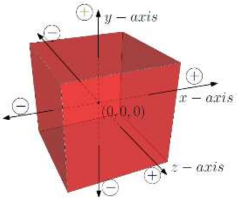
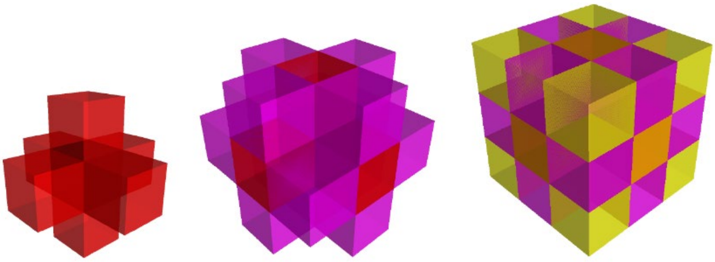


Figure 1
The origin and the directions of the three axes



6-*neighborhood* 18-*neighborhood* 26-*neighborhood*

Figure 2
The three neighborhoods, the central cube with its 6-, 18- and 26-neighbors

Let us consider two points q and p in cubic grid. The problem is to find the NSP between p and q with a given neighbor relation. To formulate the problem, the points have to be translated such that either p or q be in origin (0,0,0). Let the coordinates of the points be $p = (x_p, y_p, z_p)$ and $q = (x_q, y_q, z_q)$. The coordinates of the points after translation will be $p = (x_p - x_q, y_p - y_q, z_p - z_q)$ and $q = (0,0,0)$.

We may also recall the general definition of L_m distances in 3D between two points, which is given below.

$$L_m(p,q) = \left(\left|x_p - x_q\right|^m + \left|y_p - y_q\right|^m + \left|z_p - z_q\right|^m\right)^{\frac{1}{m}} \quad (1)$$

They are usually defined under the condition $m \geq 1$. The digital distances are discussed in [3, 4, 29]. We recall that the length of a minimal path from $p = (x_p, y_p, z_p)$ to $q(0,0,0)$ in cubic grid in 6-neighborhood, 18-neighborhood, and 26-neighborhood are denoted by metrics — L_1, D_{18}, and L_∞ respectively, since, as it is well-known, the 6- and 26-neighborhood based distances coincide with the L_1 distance, and to the L_∞ distance which is obtained in the limit $m \to \infty$.

$$L_1(p, q) = D_6(p, q) = (|x_p| + |y_p| + |z_p||) \quad (2)$$

$$D_{18}(p,q) = \max\left\{\max\{|x_p|, |y_p|, |z_p|\}, \left\lceil\frac{|x_p| + |y_p| + |z_p|}{2}\right\rceil\right\} \quad (3)$$

$$L_\infty(p, q) = D_{26}(p, q) = \max\{|x_p|, |y_p|, |z_p||\} \quad (4)$$

We also note that both the L_2 and D_{18} distances are between the above mentioned "extremal" cases, i.e., both $L_1(p, q) \geq L_2(p, q) \geq L_\infty(p, q)$ and $D_6(p, q) \geq D_{18}(p, q) \geq D_{26}(p, q)$ are satisfied for any pairs of points q, $p \in Z^3$. Furthermore, no digital distance is known that produces L_2 for any pairs of points, thus to approximate the Euclidean distance by digital distances is still a hot topic both in 2D and 3D [5, 13, 25-27, 32, 33, 37].

3 Number of Minimal Paths in 6-Neighborhood

Theorem 1. The number of minimal paths from $q = (0,0,0)$ to any point $p = (i,j,k)$ in 6-neighborhood is

$$f_{6N}(i,j,k) = \frac{(|i|+|j|+|k|)!}{|i|!|j|!|k|!} \quad (5)$$

Proof. Without loss of generality, one may assume that the coordinates of the point p are nonnegative, that is $i, j, k \geq 0$. In 3D, two points, $p'(i_1, j_1, k_1)$ and $q'(i_2, j_2, k_2)$, are in 6-neighborhood if they share a face, i.e., when the following condition holds:

$$|i_1 - i_2| + |j_1 - j_2| + |k_1 - k_2| = 1$$

Thus, since the coordinates are integers, in 6-neighborhood, in each step of a shortest path only one coordinate changes by ±1 and the other two coordinates do not change. The other coordinates of the points coincide respectively. So, the length of a shortest path between $p(i, j, k)$ and $q(0; 0; 0)$ in 6-neighborhood is $i + j + k$ (Eqn. 2), the sum of the movements along three axes. Out of total $i + j + k$ steps i, j, and k steps are taken along the x-, y-, and z-axes respectively. Their order is arbitrary, thus the total number of arrangements for a path length of $|i| + |j| + |k|$ is given by $\frac{(|i|+|j|+|k|)!}{|i|!|j|!|k|!}$, which is the total NSP in 6-neighborhood. ■

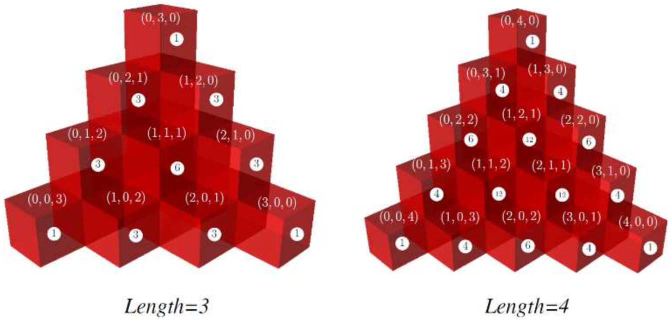


Figure 3

In Fig. 3 the NSP of length three and four are shown for some of the points along with the coordinate triplets. Actually, these numbers are the trinomial coefficients: $\frac{(|i|+|j|+|k|)!}{|i|!|j|!|k|!}$ which could play a role in expansions like $(x + y + z)^n$, see, e.g., [4].

It is to be noted here that NSP in 6-neighborhood in cubic grid is similar for some of the coordinates where $x = 0$ or $y = 0$ or $z = 0$, with the values in 4-neighborhood in 2D, i.e., the cityblock (or L_1) distance in 2D, coinciding with the binomial coefficients. More formally, if the points $p(i_1, j_1, k_1)$ and $q(i_2, j_2, k_2)$ share a coordinate (e.g., $i_1 = i_2$), then their distance and NSP (in 6-neighborhood) between them can be computed in the same way as between the points analogous points in 2D with 4-neighborhood neglecting the common coordinate, i.e., between $p_{yz}(j_1, k_1)$ and $q_{yz}(j_2, k_2)$.

4 Number of Minimal Paths in 18-Neighborhood

Without loss of generality, we assume that $i, j, k \geq 0$. The length of a shortest path in 18-neighborhood is either maximum of i, j, k or $\left\lceil \frac{i+j+k}{2} \right\rceil$ (Eqn. 3). The NSP is discussed in the following two theorems according to two cases. In our theorems q will be the origin, and we are counting the paths to the point $p = (i, j, k)$, i.e., its x coordinate is i, its y coordinate is j and z coordinate is k, the coordinate axes and their directions as shown in Fig. 1.

In the next theorem, without loss of generality, we assume that $i \geq j$ and $i \geq k$, i.e., the first coordinate value of p is (one of) the largest. Further, we use the variables a and b to denote the number of steps in specific directions made in a shortest path: whenever 2 of the coordinates are changed in a step, which is legal in this case, opposite to the previously studied D_6 case, there could be steps where both coordinates are increasing in a step, but also some steps where only the first coordinate is increasing and one of the other is decreasing. The variables a and b denote the possible numbers of such steps when the first coordinate is increasing, but either the third (variable a) or the second coordinate (variable b, resp.) is decreasing.

Theorem 2. Let $q = (0, 0, 0)$ and $p = (i, j, k)$ be two points such that $D_{18} = \max\{i, j, k\} = i$. Then, by using 18-neighborhood, from q to p, the number of minimal paths is

$$f_{18N}(i,j,k) = \sum_{a=0,b=0}^{2(a+b)\leq i-j-k} \frac{i!}{a!b!(k+a)!(j+b)!\big(i-j-k-2(a+b)\big)!}. \tag{6}$$

where a and b are the number of steps in some shortest paths in right-away (positive x and negative z) and right-bottom (positive x and negative y) directions (based on the directions of the axes shown in Fig. 1), respectively.

Proof. By the symmetry of the grid, one may assume that $D_{18} = \max\{i, j, k\} = i$, i.e., there are i-steps from q to p. $D_{18} = i$, implies that $i \geq j$ and $i \geq k$, moreover $i \geq j + k$ by Eqn. 3 (the cases when $D_{18} = j$ or $D_{18} = k$ are similar). In 18-neighborhood, a path can proceed through either a face-shared neighbor (change in only one coordinate value) or an edge-shared neighbor (change in any two coordinate values). In Eqn. 6, a and b refer to the numbers of right-away and right-bottom movements w.r.t. the positive x-axis (Fig. 1), respectively. Let $c = k + a$ and $d = j + b$ be the respective numbers of right and right-top movements. In a right-away movement, the path moves to the edge-shared neighbor where the x-coordinate increases by 1 and the z-coordinate decreases by 1 and in a right movement, both x- and z-coordinates increase by 1. Similarly, in a right-bottom movement x-coordinate increases and y-coordinates decreases while for a right-top movement, both the x and y coordinates increase. The sum of movements cannot be more than i: $a + b + c + d \leq i$, i.e., $2(a + b) \leq i - j - k$. The right-away and the right movements as well as the right-top and the right-bottom movements have some limits. When a number of right-away movements are there, the right movements will be $c = k + a$ such that the decrease of z-coordinate in a right-away moves is compensated by the increase of the z-coordinate in $k + a$ right moves in order that the destination point has z-coordinate as k. Note that in each move the x-coordinate always increases by 1. Similarly, b number of right-bottom movements implies $d = j + b$ number of right-top movements. Otherwise, it will not be possible to reach the destination in i steps. Apart from right-away, right, right-top, and right-bottom moves, there can be movements in x-direction only. For a given a, b, c, and d (i.e., right-away, right-bottom, right, and right-top respectively) steps, there are $i - (a + b + c + d) = i - j - k - 2(a + b)$ number of steps in the positive x-direction to face neighbor. Thus, for a given a, b, c, and d combination, the total number of arrangements for a shortest path of length i is given by $\frac{i!}{a!b!(k+a)!(j+b)!\big(i-j-k-2(a+b)\big)!}$. For different values of a and b, values of c and d are computed satisfying the condition that $a + b + c + d \leq i$. Thus, total NSP is the summation over the different possible combinations of a, b, c, and d values, and is given by Eqn. 6. ■

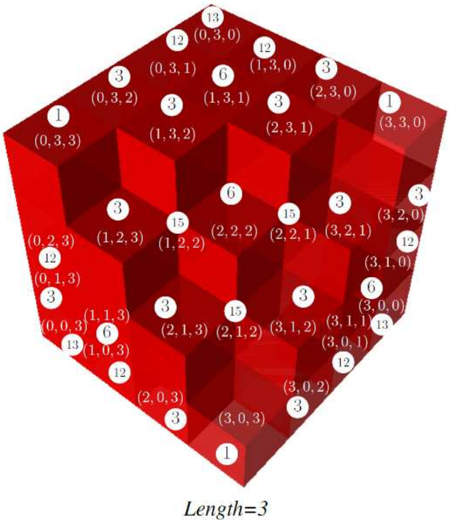


Figure 4

The NSP (given inside the white circle) from origin to other points (coordinates are shown in parentheses) in 18-neighborhoods for path length three

The NSP from $q(0, 0, 0)$ to p where the length of path is three, are given in Fig. 4. It is to be noted that when $D_{18} = \max\ \{i, j, k\} = j$ or k, the above formula (Eqn. 6) will change accordingly. The number of paths from (0, 0, 0) to (0, 3, 0) is 13 and that to (0, 3, 1) is 12.

Theorem 3. The number of minimal paths from q = (0, 0, 0) to any point $p = (i, j, k)$ in 18-neighborhood when $D_{18} = \left\lceil \frac{i+j+k}{2} \right\rceil = \tau$ is

$$f_{18N}(i,j,k) = \begin{cases} \frac{\tau!}{(\tau-i)!(\tau-j)!(\tau-k)!}, & \text{when } (i+j+k) \bmod 2 = 0 \\ \frac{\tau!((\tau-i)(\tau-j)+(\tau-j)(\tau-k)+(\tau-k)(\tau-i))}{(\tau-i)!(\tau-j)!(\tau-k)!}, & \text{when } (i+j+k) \bmod 2 = 1 \end{cases}. \tag{7}$$

Proof. At each step in 18-neighborhood, at most two coordinates can increase by one and the number of steps is τ. When $i + j + k$ is even, as τ divides $i + j + k$ with the quotient 2 which implies that at each step always two (distinct) coordinates will increase. Therefore, in a shortest path from $q(0, 0, 0)$ to $p(i, j, k)$ of length τ, the number of steps in both y- and z-directions is $\tau - i$, in both x- and z-directions is $\tau - j$, and in both x- and y-directions is $\tau - k$. Thus, the number of possible arrangements, i.e., the NSP is $\frac{\tau!}{(\tau-i)!(\tau-j)!(\tau-k)!}$

When $i + j + k$ is odd, τ divides $i + j + k + 1$ and the quotient is 2 as $D_{18} = \left\lceil \frac{i+j+k}{2} \right\rceil = \tau$, it implies that for τ – 1 steps two coordinates will increase and in the rest one step only one of the three coordinates will increase (let it be called a singular step) giving rise to the following cases:

***Singular step in x-direction*:** A shortest path has one singular step in x-direction. Thus, there are rest τ – 1 steps, where at each step there are movements in two directions. The number of steps when there are movements in x- and y-directions in each step is $(\tau - 1) - k$, in x- and z-direction is $(\tau - 1) - j$, and in y- and z-direction is $(\tau - 1) - (i - 1) = \mathrm{L} - i$. So, the number of possible shortest paths with singular x-direction is $\frac{\tau!}{(\tau-i)!((\tau-1)-j)!((\tau-1)-k)!} = \frac{\tau!(\tau-j)(\tau-k)}{(\tau-i)!(\tau-j)!(\tau-k)!}$. When $j > i + k$ and $\tau = j$ or $k > i + j$ and $\tau = k$ the singular step in x-direction will never occur.

***Singular step in y-direction*:** There will be one singular step in y-direction. Here, the number of steps in x- and y-direction is $(\tau - 1) - k$, in x- and z-direction is $(\tau - 1) - (j - 1) = \tau - j$, and in y- and z-direction is $(\tau - 1) - \mathrm{i}$, giving the number of possible shortest path with singular y-direction as $\frac{\tau!(\tau-i)(\tau-k)}{(\tau-i)!(\tau-j)!(\tau-k)!}$. When $i > j + k$ and $\tau = i$ or $k > i + j$ and $\tau = k$ the singular step in y-direction will never occur.

***Singular step in z-direction*:** One of the steps will be in the z-direction. The number of steps in x- and y-direction is $(\tau - 1) - (k - 1) = \tau - k$, in x- and z-direction is $(\tau - 1) - j$, and in y- and z-direction is $(\tau - 1) - i$. Thus, the number of possible shortest paths with singular z-direction is given by $\frac{\tau!(\tau-i)(\tau-j)}{(\tau-i)!(\tau-j)!(\tau-k)!}$. When $j > i + k$ and $\tau = j$ or $i > j + k$ and $\tau = i$ the singular step in z-direction will never occur.

Hence, the total NSP when $i + j + k$ is odd is given by $\frac{\tau!(\tau-j)(\tau-k)}{(\tau-i)!(\tau-j)!(\tau-k)!} + \frac{\tau!(\tau-i)(\tau-k)}{(\tau-i)!(\tau-j)!(\tau-k)!} + \frac{\tau!(\tau-i)(\tau-j)}{(\tau-i)!(\tau-j)!(\tau-k)!} = \frac{\tau!\left((\tau-i)(\tau-j)+(\tau-j)(\tau-k)+(\tau-k)(\tau-i)\right)}{(\tau-i)!(\tau-j)!(\tau-k)!}$. ■

The NSP for path length three is shown in Fig. 4. The NSP from (0, 0, 0) to (2, 2, 2) is 6 where $i + j + k$ is even and that to (1, 2, 2) is 15 where $i + j + k$ is odd. The number of paths from (0, 0, 0) to (0, 3, 2) satisfy both the equations stated in Theorem 2 and 3 and that from (0, 0, 0) to (2, 3, 1) also satisfy both the equations (see Fig. 4). To compute NSP between (0, 0, 0) and (9, 5, 4), the formula stated in Theorem 2 and 3 (here, $i + j + k$ is even) both are applicable and produce same result, 630. Similarly, to find NSP between (0, 0, 0) and (9, 4, 4), the formula stated in Theorem 2 and 3 (here, $i + j + k$ is odd) both yield same result 630. Remember that the distance D_{18} is computed as the maximum of a set. In some cases, it may happen that there are more maximal elements of this set, and thus, both Theorem 2 and 3 can be applied to count NSP. In these cases, they must give the same value, as we state formally in the following.

Corollary 1. The number of minimal paths from $q = (0, 0, 0)$ to any point $p = (i, j, k)$ in 18-neighborhood when $D_{18} = \left\lceil \frac{i+j+k}{2} \right\rceil = \max\{i, j, k\} = \tau = i$, $f_{18\mathrm{N}}(i, j, k)$ is as follows.

$$f_{18N}(i,j,k) = \sum_{a=0,b=0}^{2(a+b)\leq i-j-k} \frac{i!}{a!b!(k+a)!(j+b)!(i-j-k-2(a+b))!} = \begin{cases} \frac{\tau!}{(\tau-i)!(\tau-j)!(\tau-k)!}, & \text{when } (i+j+k) \bmod 2 = 0 \\ \frac{\tau!((\tau-i)(\tau-j)+(\tau-j)(\tau-k)+(\tau-k)(\tau-i))}{(\tau-i)!(\tau-j)!(\tau-k)!}, & \text{when } (i+j+k) \bmod 2 = 1 \end{cases} \quad (8)$$

Proof. The proof can be done mathematically in two parts when $i + j + k$ is even and when it is odd. Let $\tau = \frac{i+j+k}{2} = i$, i.e., $i + j + k$ is even. Thus, $i - j - k = 0$. Putting $i - j - k = 0$, in Eqn. 6 (see Theorem 2) we get, $f_{18N}(i,j,k) = \sum_{a=0,b=0}^{2(a+b)\leq 0} \frac{i!}{a!b!(k+a)!(j+b)!(-2(a+b))!}$, as there is only one possibility for the values of a and b, i.e., $a = b = 0$ since $2(a+b) \leq i - j - k = 0$. By putting these values, we get $\frac{i!}{j!k!}$. Since, $\tau = i$, $i - j = k$ and $i - k = j$. Putting these values in the first expression of Eqn. 7 (see Theorem 3) when $i + j + k$ is even, we get $\frac{i!}{j!k!}$. Hence proved.

For the second part, when $i + j + k$ is odd, $\tau = \frac{i+j+k+1}{2} = i$. Thus, $i - j - k = 1$. Putting $i - j - k = 1$ in Eqn. 6 (see Theorem 2) we get, $f_{18N}(i,j,k) = \sum_{a=0,b=0}^{2(a+b)\leq 1} \frac{i!}{a!b!(k+a)!(j+b)!(1-2(a+b))!}$, as there is one possibility for the values of a and b, i.e., $a = b = 0$ since $2(a+b) \leq i - j - k = 1$. By putting these values, we get $\frac{i!}{j!k!}$. Now, $\tau - i = 0$, $\tau - j = k + 1$, $\tau - k = j + 1$. Thus, $\frac{\tau!((\tau-i)(\tau-j)+(\tau-j)(\tau-k)+(\tau-k)(\tau-i))}{(\tau-i)!(\tau-j)!(\tau-k)!} = \frac{i!(0+(j+1)(k+1)+0)}{0!(j+1)!(k+1)!} = \frac{i!}{j!k!}$. Hence proved. ■

Similarly, the above-mentioned equation (Eqn. 8) can be proved when $\tau = j$ or k.

5 Number of Minimal Paths in 26-Neighborhood

The formulation for NSP in 26-neighborhood in cubic grid depends on NSP in 8-neighborhood in 2D. This is exactly the chessboard distance in 2D [35]. The NSP in 8-neighborhood in 2D had been proposed by Das [3, 4] with recurrence relations, in this paper we show a shorter direct proof with combinatorial tools (Eqn. 9). We count NSP from the origin to a point $p(i, j)$. Without loss of generality, we may assume that the absolute value of the first coordinate of p is not less than the absolute value of its second coordinate, i.e., $|i| \geq |j|$. Similarly, as in Theorem 2, we use a variable, here b, to denote the possible number of steps that in which both coordinates are changed, the first is changed in the direction of p from q, while the second one is changed in opposite way. Based on that we formulate the result in the next theorem, and one may see the formal details in the proof.

Theorem 4. Let $q = (0, 0)$ be the origin, and let point $p = (i, j)$ be such that their distance is i. Then, in 2D with 8-neighborhood, the number of minimal paths from q to the point p is given by

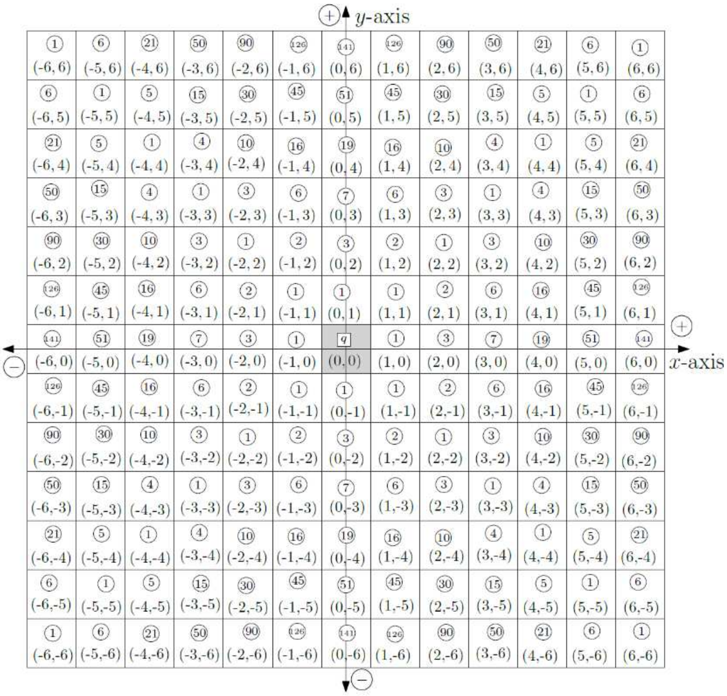
$$f_8(i,j) = \sum_{b=0}^{2b\le|i|-|j|} \frac{|i|!}{b!\,(|j|+b)!\,(|i|-|j|-2b)!} \quad \text{where } |i| \ge |j| \qquad (9)$$

Proof. The length of a shortest path between $p(i, j)$ and $q(0, 0)$ in 8-neighborhood is max $\{|i|, |j|\}$. By the symmetry of the grid, we show the proof for the case $0 \le j \le i$, in this case the distance is i. With respect to the positive x-direction, let b be the number of moves along right-bottom diagonal in the shortest path where x-coordinate increases by 1 and y-coordinate decreases by 1, and $d = |j| + b$ be the number of moves along right-top diagonal in the shortest path where the x- and y-coordinates increases by 1. A shortest path involves i steps, out of which if there are b right-bottom moves then $d = j + b$ moves only in right-top direction, hence, the number of paths is given by $\frac{|i|!}{b!(j+b)!(i-j-2b)!}$. It may be noted here that $b + d \le i$, i.e., $i - j - 2b$ as the total number of moves cannot be more than i. By summing over the different possible combinations of b and d, the total number of shortest paths is given by $f_8(i,j) = \sum_{b=0}^{2b\le i-j} \frac{|i|!}{b!(j+b)!(i-j-2b)!}$. ■

The number of paths from $q(0, 0)$ to other points in 8-neighborhood in 2D are shown in Fig. 5. Figure 6 shows two examples of all possible paths from a source to destination. For a particular path (shown in red) among all possible shortest paths, the l and r values are given for ease of understanding. The formulation for $|j| > |i|$, is just reverse (exchanging i with j) of the above equation (Eq. 9). The values appearing in 8-neighborhood in 2D are also present in 26-neighborhood of cubic grid counting the number of shortest paths from the origin to $p(i, j, k)$ if $i = j \ge k$ or $j = k \ge i$ or $i = k \ge j$ (see Fig. 5 and Fig. 7).

Now we are ready to state our last result. Without loss of generality, we count the NSP from the origin to a point $p = (i, j, k)$ such that their distance is i, i.e., $|i| \ge |j|$ and $|i| \ge |k|$. The variables a and b are used to count those steps where the y and the z coordinates are changing not to the direction of j and k, respectively, i.e., they are decreasing if the appropriate coordinate of p is nonnegative.

Theorem 5. The number of minimal paths from $q = (0, 0, 0)$ to any point $p = (i, j, k)$ in 26-neighborhood is



1 (-6, 6)	6 (-5, 6)	21 (-4, 6)	50 (-3, 6)	90 (-2, 6)	126 (-1, 6)	141 (0, 6)	126 (1, 6)	90 (2, 6)	50 (3, 6)	21 (4, 6)	6 (5, 6)	1 (6, 6)
6 (-6, 5)	1 (-5, 5)	5 (-4, 5)	15 (-3, 5)	30 (-2, 5)	45 (-1, 5)	51 (0, 5)	45 (1, 5)	30 (2, 5)	15 (3, 5)	5 (4, 5)	1 (5, 5)	6 (6, 5)
21 (-6, 4)	5 (-5, 4)	1 (-4, 4)	4 (-3, 4)	10 (-2, 4)	16 (-1, 4)	19 (0, 4)	16 (1, 4)	10 (2, 4)	4 (3, 4)	1 (4, 4)	5 (5, 4)	21 (6, 4)
50 (-6, 3)	15 (-5, 3)	4 (-4, 3)	1 (-3, 3)	3 (-2, 3)	6 (-1, 3)	7 (0, 3)	6 (1, 3)	3 (2, 3)	1 (3, 3)	4 (4, 3)	15 (5, 3)	50 (6, 3)
90 (-6, 2)	30 (-5, 2)	10 (-4, 2)	3 (-3, 2)	1 (-2, 2)	2 (-1, 2)	3 (0, 2)	2 (1, 2)	1 (2, 2)	3 (3, 2)	10 (4, 2)	30 (5, 2)	90 (6, 2)
126 (-6, 1)	45 (-5, 1)	16 (-4, 1)	6 (-3, 1)	2 (-2, 1)	1 (-1, 1)	1 (0, 1)	1 (1, 1)	2 (2, 1)	6 (3, 1)	16 (4, 1)	45 (5, 1)	126 (6, 1)
141 (-6, 0)	51 (-5, 0)	19 (-4, 0)	7 (-3, 0)	3 (-2, 0)	1 (-1, 0)	*q* (0, 0)	1 (1, 0)	3 (2, 0)	7 (3, 0)	19 (4, 0)	51 (5, 0)	141 (6, 0)
126 (-6,-1)	45 (-5,-1)	16 (-4,-1)	6 (-3,-1)	2 (-2,-1)	1 (-1,-1)	1 (0,-1)	1 (1,-1)	2 (2,-1)	6 (3,-1)	16 (4,-1)	45 (5,-1)	126 (6,-1)
90 (-6,-2)	30 (-5,-2)	10 (-4,-2)	3 (-3,-2)	1 (-2,-2)	2 (-1,-2)	3 (0,-2)	2 (1,-2)	1 (2,-2)	3 (3,-2)	10 (4,-2)	30 (5,-2)	90 (6,-2)
50 (-6,-3)	15 (-5,-3)	4 (-4,-3)	1 (-3,-3)	3 (-2,-3)	6 (-1,-3)	7 (0,-3)	6 (1,-3)	3 (2,-3)	1 (3,-3)	4 (4,-3)	15 (5,-3)	50 (6,-3)
21 (-6,-4)	5 (-5,-4)	1 (-4,-4)	4 (-3,-4)	10 (-2,-4)	16 (-1,-4)	19 (0,-4)	16 (1,-4)	10 (2,-4)	4 (3,-4)	1 (4,-4)	5 (5,-4)	21 (6,-4)
6 (-6,-5)	1 (-5,-5)	5 (-4,-5)	15 (-3,-5)	30 (-2,-5)	45 (-1,-5)	51 (0,-5)	45 (1,-5)	30 (2,-5)	15 (3,-5)	5 (4,-5)	1 (5,-5)	6 (6,-5)
1 (-6,-6)	6 (-5,-6)	21 (-4,-6)	50 (-3,-6)	90 (-2,-6)	126 (-1,-6)	141 (0,-6)	126 (1,-6)	90 (2,-6)	50 (3,-6)	21 (4,-6)	6 (5,-6)	1 (6,-6)

Figure 5

The NSP from origin to other points in 8-neighborhood in 2D. (The coordinate pairs are written in parentheses and the corresponding NSP values are also mentioned.)

$$f_{26N}(i,j,k) = \sum_{b=0}^{2b\le|i|-|j|} \frac{|i|!}{b!\,(|j|+b)!\,(|i|-|j|-2b)!} \times \sum_{a=0}^{2a\le|i|-|k|} \frac{|i|!}{a!\,(|k|+a)!\,(|i|-|k|-2a)!} \tag{10}$$

Proof. The length of a minimal path between $p(i, j, k)$ and $q(0, 0, 0)$ in 26-neighborhood is max $\{|i|, |j|, |k|\} = i$ (given in Eqn. 4) (say). In each step of a shortest path in 26-neighborhood, at most three coordinates can change. A shortest path in 26-neighborhood is a combination of a shortest path in *xy*-plane, from $q(0, 0, 0)$ to $p_{xy}(i, j, 0)$ and a shortest path in *xz*-plane, from $q(0, 0, 0)$ to $p_{xz}(i, 0, k)$. With each shortest path in *xy*-plane, each shortest path in *xz*-plane is combined to get the total NSP in 3D. Thus, the NSP in 26N is given by $f(i, j) \times f(i, k)$ where $f(i, j)$ and $f(i, k)$ are NSPs in *xy*- and *xz*-planes respectively (Theorem 4). Thus,

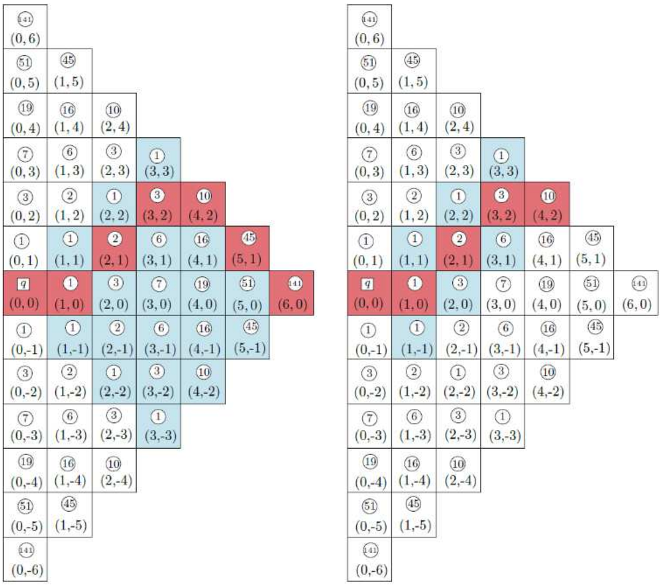


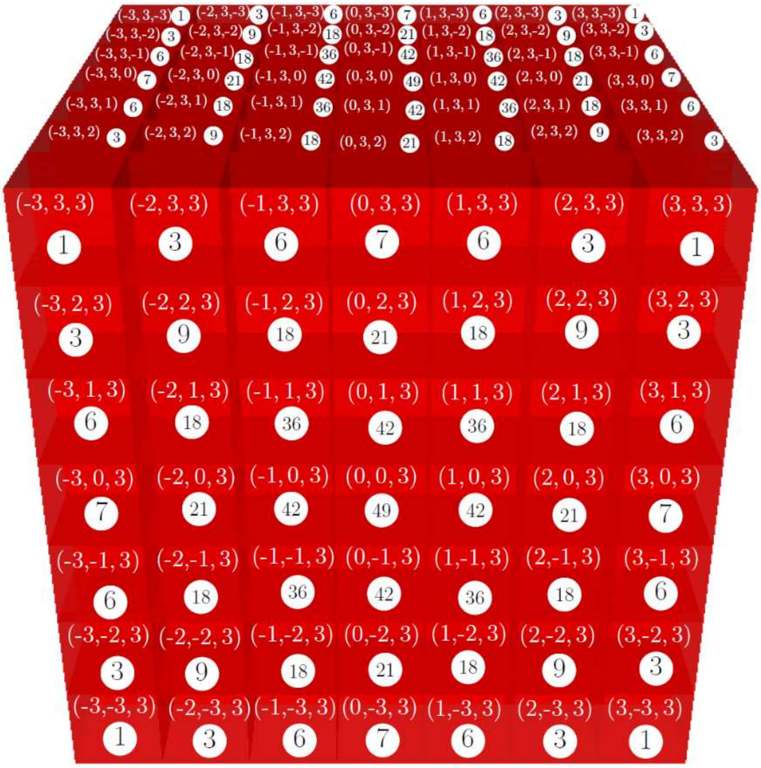
Figure 6

The NSP between the origin and other points in 8-neighborhood in 2D. The shaded portion shows the cells covered by all possible paths between two points out of which one path is shown by red color where |*j*| ≥ |*i*|. The path in left figure has *b* = 3 and *d* = |*j*| + *b* = 3 and that of right figure is *b* = 0 and *d* = |*j*| + *b* = 2.

$$f_{26N}(i,j,k) = \sum_{b=0}^{2b\leq|i|-|j|} \frac{|i|!}{b!(|j|+b)!(|i|-|j|-2b)!} \times \sum_{a=0}^{2a\leq|i|-|k|} \frac{|i|!}{a!(|k|+a)!(|i|-|k|-2a)!}$$

Where, *a* and *b* indicate the numbers of steps in right-bottom projected directions (moves that simultaneously increasing the first and decreasing the second coordinates, i.e., in positive *x*- and in negative *y*-directions, the *z*-directions might be arbitrary, i.e., ±1 or +0 for these moves) and right-away projected directions (moves in positive *x*- and negative *z*-directions, by increasing the first and decreasing the third coordinate, while the second coordinate might change by ±1 or not in these moves), respectively, if *j* and *k* are nonnegative. ■

From the Equation 10, the formulation for NSP between two points for |*j*| ≥ |*k*|, |*i*| and |*k*| ≥ |*i*|, |*j*| can be derived similarly. The NSP of length three between *q*(0, 0, 0) and some other points are shown in Fig. 7. Figure 8 shows an example path from *q*(0, 0, 0) to *p*(7, 4, 2) which has 2 right-away movements with corresponding 4 right movements and 1 right-bottom with corresponding 5 right-top movements.



Length=3

Figure 7

The NSP from origin to other points in 26-neighborhoods for *Length* = 3. (The coordinate triplets are written in parentheses and the corresponding NSP are also mentioned.) Observe that the results are not only symmetric, but they are according to a multiplication table, by Eqn. 10, where the elements of the border rows and columns are specified by the formula for the 2D L_∞ distance, i.e., the chessboard distance (Eqn. 9). Obviously, the diagonals contain the squares of the numbers shown at the borders.

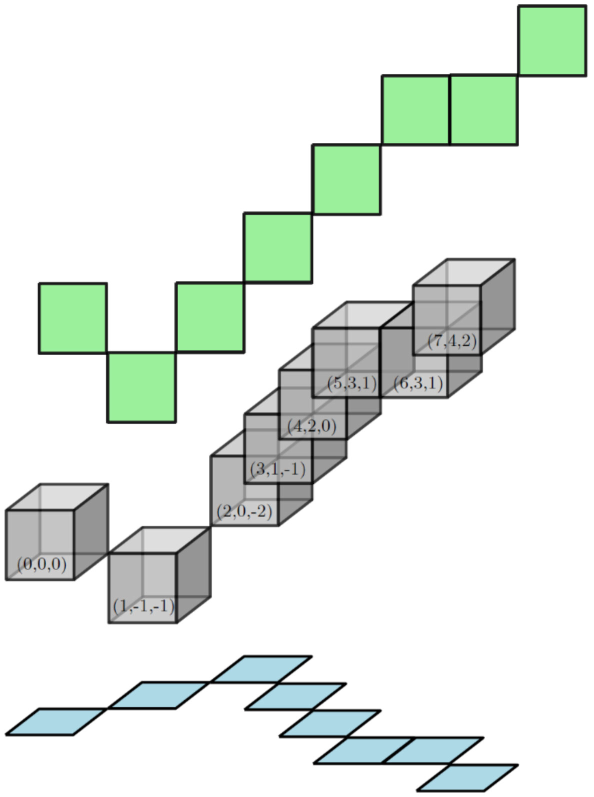


Figure 8
One of the shortest paths of length 7 from (0, 0, 0) to (7, 4, 2) is shown and correspondingly the values of *l*, *r*, *b*, and *t* are 2, 4, 1, and 5 respectively. The projection of the paths in *xy*-plane is shown at back in green color and that in *xz*-plane in blue color at the bottom.

Conclusions

The shortest path problem has various applications in several fields, especially in image processing and image analysis. Digital distances are some of the important features in this regard and many studies have already been presented. In this paper, extending the results of Das [4] from 2D to 3D, using L_1, D_{18} and L_∞ distances, the number of minimal paths (NSP) between any point pair in the cubic grid are presented for 6-, 18- and 26-neighborhood where the coordinate triplets of the two points are provided. It is also to be noted that the formulation for the NSP in 8-neighborhood in 2D is stated in this paper using combinatorial tool. In future, NSP problem in cubic grid can be extended for general orthogonal polyhedron. A 3D object can be represented by 3D orthogonal polyhedron. The critical points at different parts of 3D orthogonal polyhedrons need to be identified and the numbers of paths, between all such pairs of points, are important features for the shape analysis of 3D objects. Similar to the methods of path counting were applied in 2D images in [35].

Acknowledgements

The paper has an earlier version archived and can be found as [10].

References

[1] J. Canny and J. H. Reif (1987) New lower bound techniques for robot motion planning problems. In Proceedings of the IEEE Conference Foundations Computer Science, pp. 49-60, IEEE

[2] P. Crawley and R. P. Dilworth (1973) Algebraic theory of lattices. Prentice-Hall Inc., Englewood Cliffs, NY, USA

[3] P. P. Das (1989) An algorithm for computing the number of the minimal paths in digital images. Pattern Recognition Letters, 9(2), 107-116, DOI: 10.1016/0167-8655(89)90043-3

[4] P. P. Das (1991) Counting minimal paths in digital geometry. Pattern Recognition Letters, 12(10), 595-603, DOI: 10.1016/0167-8655(91)90013-c

[5] P. P. Das (1992) Best simple octagonal distances in digital geometry, J. Approx. Theory 68: 155-174

[6] E. S. Deutsch (1972) Thinning algorithms on rectangular, hexagonal and triangular arrays. Communications of the ACM, 15(3):827-837, 1972

[7] M. Dutt, A. Biswas, and B. B. Bhattacharya. Enumeration of shortest isothetic paths inside a digital object. In 6th International Conference on Pattern Recognition and Machine Intelligence, Vol. 9124 of LNCS, pp. 105-115, Warsaw, Poland, July 2015. Springer-Verlag

[8] M. Dutt, A. Biswas, P. Bhowmick, and B. B. Bhattacharya. On finding shortest isothetic path inside a digital object. In IWCIA'12, Vol. 7655 of LNCS, pp. 1-15, Austin, Texas, November 2012. Springer-Verlag

[9] M. Dutt, A. Biswas, P. Bhowmick, and B. B. Bhattacharya. On finding a shortest isothetic path and its monotonicity inside a digital object. Annals of Mathematics and Artificial Intelligence, 75(1-2):27-51, 2015

[10] M. Dutt, A. Biswas, and B. Nagy. Counting of Shortest Paths in Cubic Grid. CoRR abs/1803.04190, 2018

[11] Seung-Cheol Goh and Chung-Nim Lee. Counting minimal paths in 3D digital geometry. Pattern Recognition Letters, 13(11):765-771, 1992

[12] Seung-Cheol Goh and Chung-Nim Lee. Counting minimal 18-paths in 3D digital space. Pattern Recognition Letters, 14(1):39-52, 1993

[13] J. Farkas, Sz. Baják and B. Nagy. Notes on approximating the Euclidean circle in square grids. Pure Mathematics and Applications - PU.M.A. 17: 309-322, 2006

[14] P. M. Gruber. Geometry of numbers. In Handbook of convex geometry, Vol. A, B, pp. 739-763, Elsevier, North-Holland, Amsterdam, 1993

[15] P. M. Gruber and C. G. Lekkerkerker. Geometry of numbers. Second edition. North-Holland Mathematical Library, 37, North-Holland Publishing Co., Amsterdam, 1987

[16] I. Her. Geometric transformations on the hexagonal grid. IEEE Transaction on Image Processing, 4(9):1213-1221, 1995

[17] G. T. Herman. Geometry of Digital Spaces. Birkhäuser Basel, Boston, 1998

[18] C. G. Lekkerkerker. Geometry of numbers. Bibliotheca Mathematica, Vol. VIII. Wolters-Noordhoff Publishing, Groningen, North-Holland Publishing Co., Amsterdam-London, 1969

[19] F. Li and R. Klette. Euclidean shortest paths in simple cube curves at a glance. In Proceedings of the 12th International Conference on Computer Analysis of Images and Patterns, CAIP’07, pp. 661-668, Berlin, Heidelberg, 2007, Springer-Verlag

[20] F. Li and R. Klette. Rubberband algorithms for solving various 2d or 3d shortest path problems. In Proceedings of the Platinum Jubilee Conference Computing: Theory and Applications, Plenary Talk, pp. 9-18, IEEE, 2007

[21] F. Li and R. Klette. Euclidean Shortest Paths Exact or Approximate Algorithms. Springer, London, 348 2011

[22] F. Li, R. Klette, and J. B. T. M. Roerdink. Approximate ESPs in simple cubic polytopes using a rubberband algorithm. In Proceeding of Pacific Rim Symposium on Image and Video Technology, PSIVT’07, pp. 236-247, 2007

[23] R. A. Melter. A survey of digital metrics. Contemporary Mathematics, 119:95-106, 1991

[24] J. S. B. Mitchell and M. Sharir. New results on shortest paths in three dimensions. In Proceedings of the Twentieth Annual Symposium on Computational Geometry, SCG’04, pp. 124-133, New York, NY, USA, 2004, ACM

[25] J. Mukherjee. Error analysis of octagonal distances defined by periodic neighborhood sequences for approximating Euclidean metrics in arbitrary dimension. Pattern Recognit. Lett. 75: 16-23, 2016

[26] J. Mukherjee, P. P. Das, M. A. Kumar, and B. N. Chatterji. On approximating Euclidean metrics by digital distances in 2D and 3D, Pattern Recognit. Lett. 21: 573-582, 2000

[27] Jayanta Mukhopadhyay. Approximation of Euclidean Metric by Digital Distances. Springer, 2020

[28] B. Nagy. Calculating distance with neighborhood sequences in the hexagonal grid. In R. Klette and J. Zunic, editors, Proceedings of the 10th International Workshop on Combinatorial Image Analysis, IWCIA 2004, Vol. 3322 of LNCS, pp. 98-109, Auckland, New Zealand, 2004. Springer-Verlag

[29] B. Nagy. Metric and non-metric distances on Z^n by generalized neighbourhood sequences. In Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, ISPA'05, pp. 215-220, Zagreb, Croatia, 2005, IEEE

[30] B. Nagy. Digital geometry of various grids based on neighbourhood structures. In Proceedings of the 6th Conference of Hungarian Association for Image Processing and Pattern Recognition, KEPAF 2007, pp. 46-53, Debrecen, Hungary, 2007

[31] B. Nagy and R. Strand. A connection between $\mathbf{Z}^n$ and generalized triangular grids. In Proceedings of the 4th International Symposium on Advances in Visual Computing (Part II), ISVC'08, pp. 1157-1166, Las Vegas, NV, USA, 2008, Springer

[32] B. Nagy and R. Strand. Approximating Euclidean circles by neighbourhood sequences in a hexagonal grid Theoretical Computer Science 412: 1364-1377, 2011

[33] B. Nagy, R. Strand and N. Normand. A Weight Sequence Distance Function, ISSM 2013, 11th International Symposium on Mathematical Morphology, LNCS 7883: 292-301, 2013

[34] A. Rosenfeld and R. A. Melter. Digital geometry. The mathematical intelligencer, 11(3):69-72, 1989

[35] A. Rosenfeld, and J. Pfaltz. (1968) Distance functions on digital pictures. Pattern Recognition, 1(1), 33-61, DOI: 10.1016/0031-3203(68)90013-7

[36] M. Sharir. On shortest paths amidst convex polyhedra. SIAM J. Comput., 16(3):561-572, 1987

[37] R. Strand and B. Nagy. Weighted Neighbourhood Sequences in Non-Standard Three-Dimensional Grids – Parameter Optimization. Combinatorial Image Analysis, IWCIA 2008, Lecture Notes in Computer Science, LNCS 4958: 51-62, 2008

[38] C. A. Wüthrich and P. Stucki. An algorithmic comparison between square- and hexagonal-based grids. CVGIP: Graphical Models and Image Processing, 53:324-339, 1991

Hierarchy of Roadmap Items: Prioritization Strategy Development in Aircraft MRO Industry to Enhance Profit and Sustainability

Sally Ichou[1,2], Árpád Veress[1]

[1]Department of Aeronautics and Naval Architecture, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary

[2]Aeroplex of Central Europe Aircraft Technology Center Ltd., 1185 Budapest, Liszt Ferenc International Airport, Hungary, e-mail: sichou@edu.bme.hu, veress.arpad@kjk.bme.hu

Abstract: Aviation's global aftermarket is expected to grow 22% in 2023, topping $94 billion, and will reach $125 billion by 2033 with a 2.9% compound annual growth rate. Besides that, the aircraft maintenance industry operates in a highly dynamic and competitive environment where ceaseless development is a requirement to ensure continuity, profitability, and sustainability. Hence, and based on the actual technological level and so striving for higher level advancement and introduction of new technologies and processes in this area, a significant number of research and development projects are under way or in the pipeline. With so many new ideas and innovations, it is hard for the upper management to make informed decisions and be sure that those decisions are what the company needs. Usually, these decisions are mainly made from one or a set of managers' points of view. However, the decision might not be suitable from a scientific point of view due to the numerous factors, data, and concerns that are very hard to spot without a numerical perspective. Hence, it is important to create a scientific strategy to prioritize these new ideas or items based on accurate factors and indications to give a vision of what is the most needed idea to adopt by the company. This is what the present paper is focused on. A case study investigation was made herein, where after collecting all the proposed ideas and development areas from the assigned managers and decision-makers, a prioritization was made. The ideas were ranked based on a novel hierarchy model which was developed to govern the development process of the maintenance repair and overhaul activity in the optimum and effective way to reach the set expectations. It means that the companies can have a scientific tool to point at the development direction that they should follow, and their focus will be shed on the most essential activities at hand and aim to get the most value towards the fulfilment of the upper management goals and stakeholders' vision. The proposed framework has assessed three different groups of aircraft maintenance development concepts based on a set of picked criteria and concludes which out of the three should be pursued next for more development. The results taken from the created framework showed that the "Development and Digitalization of the Operational Process for MRO Applications" topic is the most interesting one with 34 scores, which is higher than the second one by 21%. The proposed solution not *only showed the reliability of this framework to give good decision support but also showed that it could build a suitable, unified structure and procedure to follow while determining the company's future development direction.*

Keywords: Aircraft maintenance repair and overhaul; Management strategy; Roadmap Item Prioritization; Decision Making

1 Introduction

The aviation industry in the shape it is now is the outcome of years of human engineering and development, with aircraft serving as the lifeline of modern transportation and connectivity. Maintaining and assuring the safety and dependability of these aircraft with their complex components and systems is a constant challenge. This is where aircraft Maintenance, Repair, and Overhaul (MRO) comes in, playing a huge and critical role in ensuring that aircraft stay in optimum condition throughout their operating lifespan [1]. However, in the ever-changing aviation market and aircraft technology [2], MRO companies need to be flexible enough to adapt to new ideas and methodologies to keep services of high quality, efficiency, and accuracy.

MRO in the context of aviation refers to all activities and actions taken to keep aircraft airworthiness, such as inspections, repairs, component changes, and system updates [3]. It is the backbone of aviation safety and efficiency, ensuring that aircraft function consistently while meeting severe regulatory requirements from both the manufacturers and authorities [4]. Achieving one single flight requires many man-hours of maintenance labour, which often goes unseen to passengers but is extremely indispensable.

Aircraft design and manufacturing science have developed quite a lot in the past few decades, the proof of that is the countless research articles in different fields, which include many developments in electrification [5], [6], sustainability [7] [8] [9], noise reduction [10], the propulsion system [11] and its components [12] for example.

Based on [13] MRO demand expanded 18% in 2022. The compound annual growth of global maintenance, repair, and overhaul services is expected to grow 22% in 2023, topping $ 94 billion, which is a mere 2% below its 2019 peak. It will reach $ 125 billion by 2033 expectedly at a compound annual growth rate of 2.9%.

The developments of the aeronautical sciences and technologies together with the increasing need/business for the MRO services, force the concerned companies to reach for cutting-edge technologies and innovations. And since maintenance goes hand in hand with aircraft design, it needs to keep pace with the aircraft evolution. It is no longer enough to use basic and standardized approaches and procedures, MRO is growing into a highly complex sector that employs modern data analytics [14], predictive maintenance [15], robotics [16], digitalization [17], and automation [18] to enhance efficiency and reduce downtime.

However, in a sea full of new ideas, picking the correct development direction or idea is not as easy as it looks. The management needs to be up to date with the actual situation in the scientific and industrial world at once, plus, they need to be aware of all the pros and cons that one idea could have. This is why, it is important to pick the highest-ranking idea and correctly manage the allocation of human and financial resources that will bring additional profit and success to the company rather than wasting those resources on a direction that the work environment is not ready for. Hence, upper management must make well-established decisions regarding which concepts or projects to invest in.

This paper focuses on concept evaluation, highlighting the importance of prioritization in the aircraft MRO field, and it sheds light on a specific case study conducted in a modern MRO enterprise. The goal of this study is to create a framework on which companies can rely while picking a direction that is most aligned and suitable for their own internal development that will lead to further international strength and presence in global MRO markets.

The next chapter of this study will begin by introducing how will a company gather ideas relying on the industry's future innovation, then it will establish the vision, explaining in detail the concept generation and how items are selected for the roadmap. Afterward, the criteria, on which the items or concepts will be ranked off are defined, finally, the framework for the ranking is created, and the result of the assessment is announced. The created framework although used on a specific MRO study case and MRO enterprise can be a promising tool that can be developed further to be used as a universal decision-support tool for upper management and innovation teams.

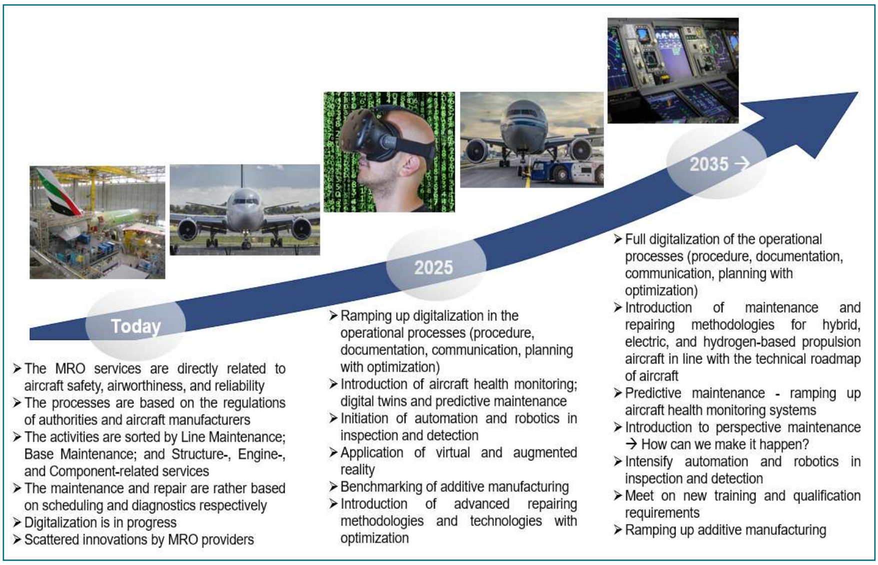
2 MRO Roadmap

When aviation first began, the maintenance culture was quite limited, if not "non-existent," because there were few sensors aboard the aircraft and therefore a minimal amount of data and information available. Maintenance was ignored as efforts were focused primarily on aircraft development. As time and electrification/digitalization progressed, more sensors were placed on the aircraft structure and airframe, resulting in a greater amount of data that may be used to create Knowledge-Based Maintenance (KBM) [19].

Due to the need for cleaner and more ecologically friendly for the next decade, innovations are shifting to low-emission and hybrid-powered aircraft. The maintenance costs will be reduced when conventional and electric-powered propulsion systems are integrated since the operational time over a given period will be reduced, and therefore the intervals between inspections may be extended. MRO-related culture is predicted to receive more attention [20].

Since the MRO technology needs to follow the aircraft technology, it is crucial to keep in mind that all concepts and idea proposals should align with the MRO roadmap.

Figure 1 depicts the technology roadmap for MRO activities in the aeronautical sector. This roadmap essentially specifies the elements of the direction, from which numerous projects may be initiated. The roadmap contains various groups and product items, which can be updated based on the actual economic situation from time to time in one hand, and then year by year, and according to business values, items are picked from this roadmap to be evaluated and ranked, and only when the Business Cases are available and Return on Invest (ROI) is sufficiently high to be approved, can these items be started to be developed.



- The MRO services are directly related to aircraft safety, airworthiness, and reliability
- The processes are based on the regulations of authorities and aircraft manufacturers
- The activities are sorted by Line Maintenance; Base Maintenance; and Structure-, Engine-, and Component-related services
- The maintenance and repair are rather based on scheduling and diagnostics respectively
- Digitalization is in progress
- Scattered innovations by MRO providers

- Ramping up digitalization in the operational processes (procedure, documentation, communication, planning with optimization)
- Introduction of aircraft health monitoring; digital twins and predictive maintenance
- Initiation of automation and robotics in inspection and detection
- Application of virtual and augmented reality
- Benchmarking of additive manufacturing
- Introduction of advanced repairing methodologies and technologies with optimization

- Full digitalization of the operational processes (procedure, documentation, communication, planning with optimization)
- Introduction of maintenance and repairing methodologies for hybrid, electric, and hydrogen-based propulsion aircraft in line with the technical roadmap of aircraft
- Predictive maintenance - ramping up aircraft health monitoring systems
- Introduction to perspective maintenance → How can we make it happen?
- Intensify automation and robotics in inspection and detection
- Meet on new training and qualification requirements
- Ramping up additive manufacturing

Figure 1

Technical roadmap of MRO activities in the aeronautical sector [20]

3 Prioritization Strategy

Concept or project evaluation means the critical examination of collected and proposed maintenance ideas and strategies. Although innovative ideas are generated long before any design or engineering work is done, they can come from either the management or the engineers and technicians, companies need to take into consideration many aspects to make the best decision about what projects will be started. These aspects are the outputs of the market study, the customer experiences and specifications, the product features, the product-market fit, the cost, ROI, the profit, the sustainability, among many others, which are needed to support the decision about the concepts to be initiated, and this is why developing a prioritization strategy becomes essential [21].

Prioritization has existed in our daily life routine since we started having more responsibility. Effective prioritizing is essential for development since one can only discover what works and what doesn't, what is worth doing and what is not, by putting things to the test in real life after detailed examination. Similarly, organizations and bigger enterprises perform some sort of prioritization process [22].

Additionally, there are many improvements, developments, new products, and services available in the market, so it can be a bit overwhelming to pick and choose from the numerous ideas. Therefore, there is an emphasis on the importance of finding a system to create this hierarchy of importance. Finding this kind of system, will not only maximize the profit and the sustainability, but it makes the job easier and will also increase transparency so that every employee understands what the plan is to follow, and what types of ideas should be supported and sought after in the first place.

In this case study, Aeroplex of Central Europe (ACE) [23] was approached in cooperation with the Budapest University of Technology and Economics (BME) [24] in order to find the most important project to develop under the framework of a Ph.D. research.

3.1 Generation of Concept from an Industrial Point of View

The five-step process was used as a reference when building the prioritization framework as indicated in Figure 2.

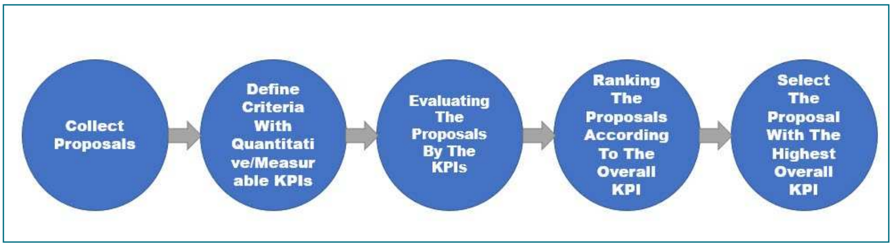

Collect Proposals
Define Criteria With Quantitative/Measurable KPIs
Evaluating The Proposals By The KPIs
Ranking The Proposals According To The Overall KPI
Select The Proposal With The Highest Overall KPI

Figure 2
The five-step process prioritization framework

The first step to generate a selective concept system for prioritization is to complete deep literature research in order to collect all available innovations in the field of MRO including the vision of aircraft manufacturers. The ideas ranged from implementing new technology for maintenance to improve current procedures or implementing cost-cutting efforts, to create totally new maintenance methodologies by integrating robotics.

It was challenging trying to prioritize every proposal as it was a wide spectrum of ideas and solutions, which is why it was necessary to do a grouping for the ideas under three main categories as indicated in Figure 3.

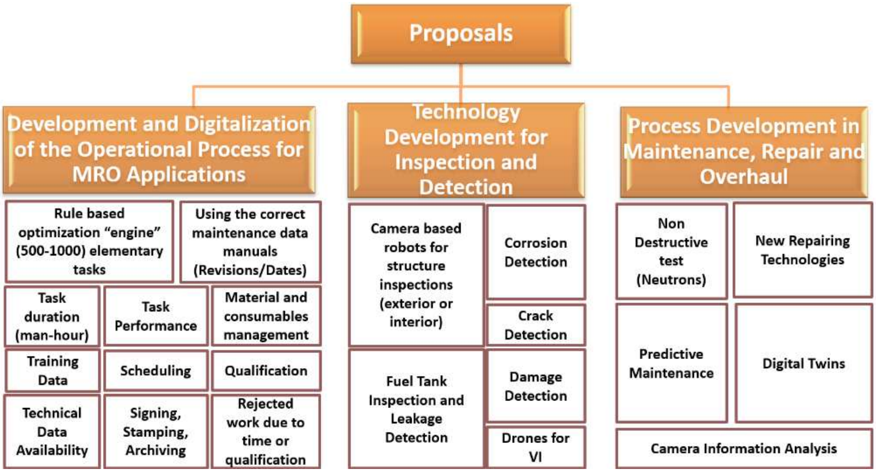

Proposals
Development and Digitalization of the Operational Process for MRO Applications
Technology Development for Inspection and Detection
Process Development in Maintenance, Repair and Overhaul
Rule based optimization "engine" (500-1000) elementary tasks
Using the correct maintenance data manuals (Revisions/Dates)
Task duration (man-hour)
Task Performance
Material and consumables management
Training Data
Scheduling
Qualification
Technical Data Availability
Signing, Stamping, Archiving
Rejected work due to time or qualification
Camera based robots for structure inspections (exterior or interior)
Corrosion Detection
Crack Detection
Fuel Tank Inspection and Leakage Detection
Damage Detection
Drones for VI
Non Destructive test (Neutrons)
New Repairing Technologies
Predictive Maintenance
Digital Twins
Camera Information Analysis

Figure 3
Grouping of MRO innovation project proposals as an output of the scientific literature research

Therefore, the topics of the research work have been categorized into the following three main groups, and these concepts will be evaluated instead of each individual idea:

- Development and Digitalization of the Operational Process for MRO Applications
- Technology Development for Inspection and Detection,
- Technical Developments in Maintenance, Repair, and Overhaul.

To make sure that these ideas were relevant in the scientific world and aligned with the MRO roadmap, a cross-check was made from the scientific literature studies and investigations. And indeed, every study that was found fits perfectly under one of these three groups. There are several studies that investigate planning [25] and scheduling problems [26] that can be included in topic 1 Development and Digitalization of the Operational Process for MRO Applications. Even research that concerns digitalization [24], Augmented Reality (AR) new maintenance manuals

[28], Virtual Reality (VR) new training material [29], speech recognition to aid technicians [30], and even applications [31] and software development [32]. As for the second concept, the literature is full of novel ideas that help with the detection and inspection of the aircraft structure. Some researchers proposed using robots for visual inspections [33] [34] [35], or to perform fuel tank inspections instead of humans due to the health hazards [36], and other studies that incorporated drones for structure inspections [37] and many more. The third concept includes various repair innovations such as using machine learning for predictive maintenance [38], new ways to do aircraft structural health monitoring [39], digital twins [40], intelligent troubleshooting [41], integrating new eddy current sensors into the repair patches [42], and many more studies like these.

So, it seems like these three concepts apply even to the studies found in the literature, and based on that it will be enough if the evaluation would only focus on them.

Each concept must be rigorously evaluated, considering criteria definitions that match the company's set goals and future vision such as practicality, cost-effectiveness, safety concerns, and the potential to improve the maintenance process's reliability and dependability for example.

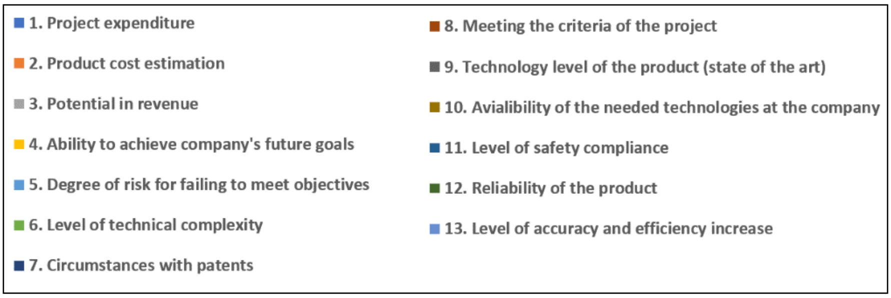
3.2 Criteria Definition

The project prioritization is a complex problem. There are three main methods for the project prioritization in the scientific literature as 1. Scoring Model, 2. Project Prioritization Matrix and the 3. Payback Period [43]. A technique for project prioritization based on the opinions of subject matter experts is the scoring model. All that is involved in the scoring process is assessing various project components and then giving each one a numerical value scale. This model is the most general, it considers all aspects needed for decision making with special care for the R&D sector. The mentioned list of criteria in [43] is limited. Hence, an MRO industry-related specification list is developed in the framework of the present research.

The ideal strategy to define the prioritizing criteria is to begin by identifying the company's key development drivers and determining with them which concept out of the three to focus on.

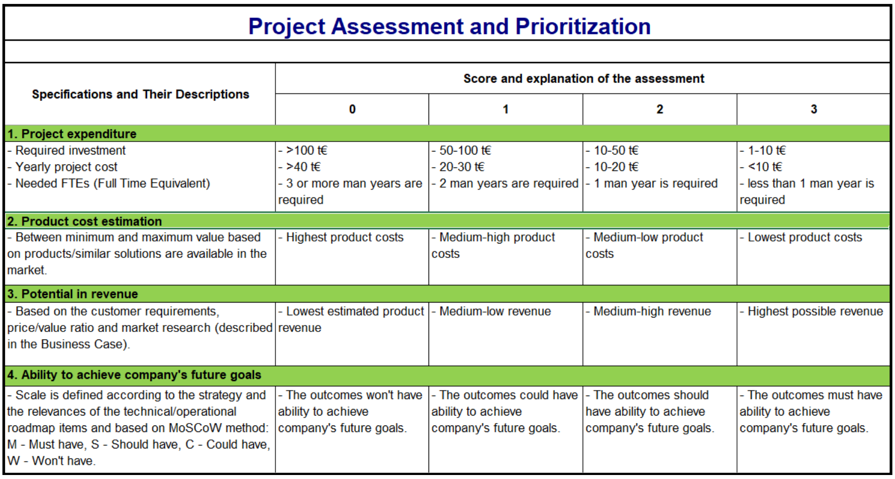
Having a straightforward and structured set of criteria for ranking the project and research proposal allows one to make more consistent and better selections than merely depending on perceptions and individual opinions. Hence, the criteria were chosen that describe the greatest ability to have impact for the future of the enterprises, as shown in Figure 4.

The next step was to create an evaluate spectrum or scale for the 13 criteria, this enabled to have enough variation between the results. An example of the evaluation spectrum can be identified in Figure 5.



- 🟦 1. Project expenditure
- 🟧 2. Product cost estimation
- ■ 3. Potential in revenue
- 🟨 4. Ability to achieve company's future goals
- 🟦 5. Degree of risk for failing to meet objectives
- 🟩 6. Level of technical complexity
- 🟦 7. Circumstances with patents
- 🟫 8. Meeting the criteria of the project
- ■ 9. Technology level of the product (state of the art)
- 🟨 10. Avialibility of the needed technologies at the company
- 🟦 11. Level of safety compliance
- 🟩 12. Reliability of the product
- 🟦 13. Level of accuracy and efficiency increase

Figure 4
Prioritization criteria for the concepts



Project Assessment and Prioritization

Specifications and Their Descriptions	Score and explanation of the assessment			
	0	1	2	3
1. Project expenditure				
- Required investment - Yearly project cost - Needed FTEs (Full Time Equivalent)	- >100 t€ - >40 t€ - 3 or more man years are required	- 50-100 t€ - 20-30 t€ - 2 man years are required	- 10-50 t€ - 10-20 t€ - 1 man year is required	- 1-10 t€ - <10 t€ - less than 1 man year is required
2. Product cost estimation				
- Between minimum and maximum value based on products/similar solutions are available in the market.	- Highest product costs	- Medium-high product costs	- Medium-low product costs	- Lowest product costs
3. Potential in revenue				
- Based on the customer requirements, price/value ratio and market research (described in the Business Case).	- Lowest estimated product revenue	- Medium-low revenue	- Medium-high revenue	- Highest possible revenue
4. Ability to achieve company's future goals				
- Scale is defined according to the strategy and the relevances of the technical/operational roadmap items and based on MoSCoW method: M - Must have, S - Should have, C - Could have, W - Won't have.	- The outcomes won't have ability to achieve company's future goals.	- The outcomes could have ability to achieve company's future goals.	- The outcomes should have ability to achieve company's future goals.	- The outcomes must have ability to achieve company's future goals.

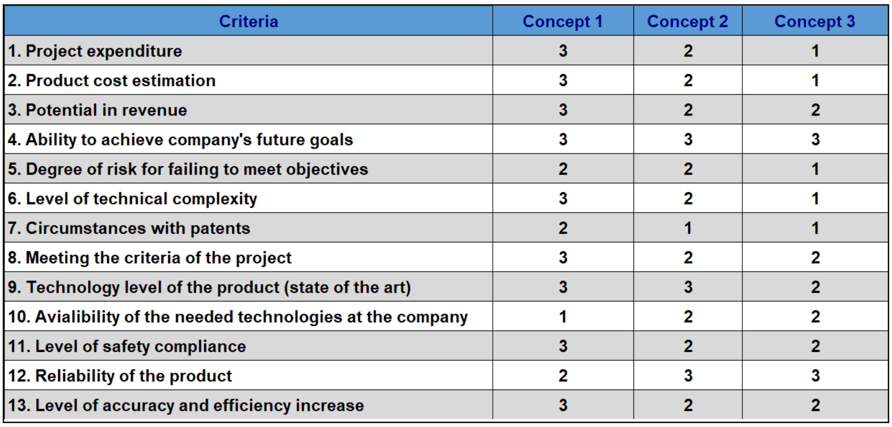
Figure 5
Evaluation spectrum

4 Assessment of the Results

The evaluation itself is the next step of the priority process. The votes were made by all partners involved in the project. Following the ratings and summarising the results, the prioritization is made. The ranking of the concepts is done based on their strategic value and expected impact. Concepts that, besides the others, promise the most advantages, whether in terms of increased safety, operational reliability, profit, sustainability, or cost savings, were prioritized. It is worth mentioning that given the frequently restricted resources available, not every concept could be pursued concurrently, so the accepted budget determines the number of projects to be started.

Based on the detailed assessment using the mentioned criteria, the results are found in Figure 6.

As a summary of the assessment, concept one, entitled “Development and Digitalization of the Operational Process for MRO Applications” has received the highest mark 34 and so the priority level amongst the others (see Figure 7), so it has been selected to be the most important from the industry point of view today. Thus, it has been nominated to be the topic of the development project to work out.



Criteria	Concept 1	Concept 2	Concept 3
1. Project expenditure	3	2	1
2. Product cost estimation	3	2	1
3. Potential in revenue	3	2	2
4. Ability to achieve company's future goals	3	3	3
5. Degree of risk for failing to meet objectives	2	2	1
6. Level of technical complexity	3	2	1
7. Circumstances with patents	2	1	1
8. Meeting the criteria of the project	3	2	2
9. Technology level of the product (state of the art)	3	3	2
10. Avialibility of the needed technologies at the company	1	2	2
11. Level of safety compliance	3	2	2
12. Reliability of the product	2	3	3
13. Level of accuracy and efficiency increase	3	2	2

Figure 6

Concepts assessment results

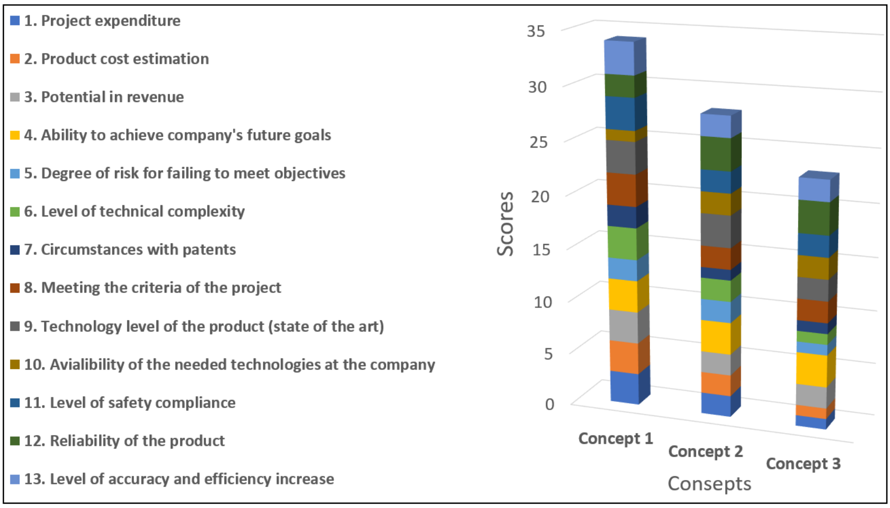


Figure 7

Assessment result chart

The expected benefit from selecting this concept and so to promote sustainable development are saving cost, time, and capacity due to the paperless documentation/administration based on

- reduction in the material and energy consumption,
- decrease in human error,
- less employees required to manage and deliver work packages,
- smaller size of the space for storing documents,
- better project/task/cost transparency,
- improved communication, and
- higher flexibility of the workplaces (e.g.: open office, home office).

Additionally, the capacity, the cost, the investment, the inventory, the tools, the workplace, the asset use, and the incoming task planning can be optimized more quickly, accurately, and frequently.

Whereas concept "Technology Development for Inspection and Detection" was in the middle with 28, and concept number three "Technology Developments in Maintenance, Repair, and Overhaul" ranked the lowest with 23 points, meaning these two concepts didn't bring the highest value from the company's perspective.

After the assessment procedure comes to the realization of the concept and execution of the project in the most efficient way in accordance with the company's set goals and visions. Following that is the plausibility check, verification, and validation of the outcomes and the realisation of the above-mentioned benefits are going to be compared with the actual situation using the baseline version of the process.

As these results are unique, company-dependent, and are under a confidentiality agreement in most of the cases, there are no similar topics published in the scientific literature according to the best knowledge of the authors.

Conclusions

The need for services in the field of aircraft maintenance, repair, and overhaul is increasing continuously. Although there is quite a development in the MRO industry today, there is still a potential for more technological and procedural advancement. Also, there is a gap in utilizing the latest innovations as much as it would be. This can also be proven by the fact that processes are rather based on human intervention instead of digital solutions.

Aircraft MRO is a vital and active part of the aviation sector. It guarantees that aircraft are safe, airworthy, and reliable throughout their operating lifespan. However, as aviation technology advances, the necessity for effective concept appraisal and prioritizing becomes even more critical. Stakeholders may strategically manage resources by carefully reviewing and prioritizing maintenance ideas, keeping aircraft at the forefront of safety and performance while navigating the ever-changing skies of the aviation business.

Based on the detailed literature research in the field of MRO innovation, and on the higher number of the project proposals, grouping was made for the given developments and concept ideas. Three groups have been identified as follows: “Development and Digitalization of the Operational Process for MRO Applications”, “Technology Development for Inspection and Detection”, and “Technical Developments in Maintenance, Repair, and Overhaul” for distinguishing the innovations. Then, it was important to identify specifications that can be used to prioritize these roadmap items. The specifications were made based on the identified criteria items which were picked to be in accordance with the goals and vision of the upper management. As a practical implication, the present work offers a direction also for managers of the MRO sector in structuring and applying this hierarchy for prioritizing processes aimed at organizational efficiency.

It was concluded that the concept titled “Development and Digitalization of the Operational Process for MRO Applications” ranked noticeably the highest compared to the other two. It means for the time being the improvements of the operational processes are the direction to be realised while considering the company’s future steps. Hence, for future developments, the company needs to establish a project plan, which includes a project description with clear goals, members and roles of the steering committee, achievable outcomes, and a timeline with milestones according to the official process description.

According to the prioritization framework and the criteria that it was based on and taking into consideration the high score of 34 points that the “Development and Digitalization of the Operational Process for MRO Applications” concept reached, it is safe to say that if the company follows that road, a major decrease in cost, time, and capacity will be realized. Hence, following the detailed investigation and analyses of the recent process, a new digital framework is going to be developed in the future in the form of a development project and Ph.D. research to improve the operational characteristics that are aligned with the results achieved.

Digitalization in this context means not only data transferring, handling, and storing but using them to find the best solutions for planning using the theories of the Internet of things, big data, machine learning, business analytics, blockchain, data processing, and artificial intelligence for example. The outcomes of this project development are expected to be in line with the regulation of quality insurance, which includes the expectations of aviation authorities and manufacturers. And of course, the provided solutions are going to be verified and validated based on different examples and test scenarios.

The benefit of the developed digital solution must be proven with the recently used, rather manual-based process management and planning activities by using KPIs (Key Performance Indexes).

References

[1] E. Karakilic, E. Gunaltili, S. Ekici, A. Dalkiran, O. Balli, and T. H. Karakoc, “A comparative study between paper and paperless aircraft maintenance: A case study,” *Sustainability*, Vol. 15, No. 20, p. 15150, 2023

[2] F. Ekici, G. Orhan, Ö. Gümüş, and A. B. Bahce, “A policy on the externality problem and solution suggestions in air transportation: The environment and sustainability,” *Energy*, Vol. 258, p. 124827, 2022

[3] EASA, “Study on the need of a common worksheet/work card system.” EASA, Dec. 10, 2007, Accessed: Dec. 17, 2023 [Online] Available: https://www.easa.europa.eu/sites/default/files/dfu/Study%20for%20task%20145-020%20-%20work%20card%20system.pdf

[4] M. Quinlan, I. Hampson, and S. Gregson, “Outsourcing and offshoring aircraft maintenance in the US: Implications for safety,” *Safety science*, Vol. 57, pp. 283-292, 2013

[5] D. Sziroczak, I. Jankovics, I. Gal, and D. Rohacs, “Conceptual design of small aircraft with hybrid-electric propulsion systems,” *Energy*, Vol. 204, p. 117937, 2020, doi: https://doi.org/10.1016/j.energy.2020.117937

[6] J. Rohacs and D. Rohacs, “Energy coefficients for comparison of aircraft supported by different propulsion systems,” *Energy*, Vol. 191, p. 116391, 2020, doi: https://doi.org/10.1016/j.energy.2019.116391

[7] S. Ekici, A. Dalkiran, and T. H. Karakoc, “A short review on sustainable aviation and public promises on future prospects,” in *Research Developments in Sustainable Aviation: Proceedings of International Symposium on Sustainable Aviation 2021*, Springer Nature, 2023, p. 1

[8] M. S. Abdalla, O. Balli, O. H. Adali, P. Korba, and U. Kale, “Thermodynamic, sustainability, environmental and damage cost analyses of jet fuel starter gas turbine engine,” *Energy*, Vol. 267, p. 126487, 2023

[9] P. Korba, O. Balli, H. Caliskan, S. Al-Rabeei, and U. Kale, “Energy, exergy, economic, environmental, and sustainability assessments of the CFM56-3 series turbofan engine used in the aviation sector,” *Energy*, Vol. 269, p. 126765, 2023

[10] J. Bera and L. Pokorádi, “Monte-Carlo simulation of helicopter noise,” *Acta Polytechnica Hungarica*, Vol. 12, No. 2, pp. 21-32, 2015

[11] K. Beneda, “Investigation of novel thrust parameters to variable geometry turbojet engines,” in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, IEEE, 2021, pp. 000339-000342

[12] M. Spodniak, L. Főző, R. Andoga, K. Semrád, and K. Beneda, “Methodology for the water injection system design based on numerical models,” *Acta Polytechnica Hungarica*, Vol. 18, No. 4, 2021

[13] OliverWyman, “Global fleet and MRO market forecast 2023-2033.” Accessed: Dec. 21, 2023 [Online] Available: https://www.oliverwyman.com/our-expertise/insights/2023/feb/global-fleet-and-mro-market-forecast-2023-2033.html

[14] M. Pelt, K. Stamoulis, and A. Apostolidis, “Data analytics case studies in the maintenance, repair and overhaul (MRO) industry,” *MATEC Web Conf.*, Vol. 304, p. 04005, 2019, doi: 10.1051/matecconf/201930404005

[15] T. Tyncherov and L. Rozkova, “Predictive maintenance model of refined aircraft tires replacement,” in *International Conference on Reliability and Statistics in Transportation and Communication*, Springer, 2020, pp. 164-173

[16] G. Niu, J. Wang, and K. Xu, “Model analysis for a continuum aircraft fuel tank inspection robot based on the Rzeppa universal joint,” *Advances in Mechanical Engineering*, Vol. 10, No. 5, p. 1687814018778229, 2018

[17] J. Ordieres-Meré, T. Prieto Remon, and J. Rubio, “Digitalization: An opportunity for contributing to sustainability from knowledge creation,” *Sustainability*, Vol. 12, No. 4, p. 1460, 2020

[18] S. Bouarfa, A. Doğru, R. Arizar, R. Aydoğan, and J. Serafico, “Towards automated aircraft maintenance inspection. A use case of detecting aircraft dents using Mask R-CNN,” in *AIAA Scitech 2020 forum*, 2020, p. 0389

[19] F. Ansari, R. Glawar, and W. Sihn, “Prescriptive maintenance of CPPS by integrating multimodal data with dynamic bayesian networks,” in *Machine Learning for Cyber Physical Systems*, Springer, 2020, pp. 1-8

[20] S. Ichou and Á. Veress, “Technology roadmap for aircraft maintenance, repair and overhaul,” *Aeronautical Science Bulletins*, Vol. 34, No. 3, pp. 19-30, 2022

[21] K. Hayat, M. I. Ali, F. Karaaslan, B.-Y. Cao, and M. H. Shah, “Design concept evaluation using soft sets based on acceptable and satisfactory levels: an integrated TOPSIS and Shannon entropy,” *Soft Computing*, Vol. 24, pp. 2229-2263, 2020

[22] J. R. Marques, “Project process improvement and other points – How to do it and apply it?,” Portal. Accessed: Sep. 20, 2023 [Online] Available: https://www.ibccoaching.com.br/portal/rh-gestao-pessoas/projeto-melhoria-processos-outros-pontos-como-fazer-aplicar/

[23] Aeroplex of Central Europe Aircraft Technology Center, “Company Profile.” Accessed: Sep. 20, 2023 [Online] Available: https://www.aeroplex.com/content/company-profile.html

[24] Budapest University of Technology and Economics, “Budapest University of Technology and Economics,” Budapest University of Technology and

Economics. Accessed: Sep. 20, 2023 [Online] Available: https://www.bme.hu/?language=en

[25] D. Dinis, A. Barbosa-Póvoa, and Â. P. Teixeira, "A supporting framework for maintenance capacity planning and scheduling: Development and application in the aircraft MRO industry," *International Journal of Production Economics*, Vol. 218, pp. 1-15, 2019

[26] S. Albakkoush, E. Pagone, and K. Salonitis, "An approach to airline MRO operators planning and scheduling during aircraft line maintenance checks using discrete event simulation," *Procedia Manufacturing*, Vol. 54, pp. 160-165, 2021

[27] M. Esposito, M. Lazoi, A. Margarito, and L. Quarta, "Innovating the maintenance repair and overhaul phase through digitalization," *Aerospace*, Vol. 6, No. 5, p. 53, 2019

[28] S. Hongli, W. Qingmiao, Y. Weixuan, L. Yuan, C. Yihui, and W. Hongchao, "Application of AR technology in aircraft maintenance manual," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 012133

[29] A. Siyaev and G.-S. Jo, "Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality," *Sensors*, Vol. 21, No. 6, p. 2066, 2021

[30] A. Siyaev and G.-S. Jo, "Neuro-Symbolic speech understanding in aircraft maintenance metaverse," *IEEE Access*, Vol. 9, pp. 154484-154499, 2021

[31] H. M. Shakir and B. Iqbal, "Application of Lean principles and software solutions for maintenance records in continuing airworthiness management organisations," *The Aeronautical Journal*, Vol. 122, No. 1254, pp. 1263-1274, 2018

[32] C.-C. Yuan, C.-H. Li, and C.-C. Peng, "Development of mobile interactive courses based on an artificial intelligence chatbot on the communication software LINE," *Interactive Learning Environments*, Vol. 31, No. 6, pp. 3562-3576, 2023

[33] Y. Sun, L. Zhang, and O. Ma, "Robotics-assisted 3D scanning of aircraft," in *AIAA AVIATION 2020 FORUM*, 2020, p. 3224

[34] A. Doğru, S. Bouarfa, R. Arizar, and R. Aydoğan, "Using convolutional neural networks to automate aircraft maintenance visual inspection," *Aerospace*, Vol. 7, No. 12, p. 171, 2020

[35] B. Ramalingam *et al.*, "Visual inspection of the aircraft surface using a teleoperated reconfigurable climbing robot and enhanced deep learning technique," *International Journal of Aerospace Engineering*, Vol. 2019, 2019

[36] F. Heilemann, A. Dadashi, and K. Wicke, “Eeloscope—Towards a novel endoscopic system enabling digital aircraft fuel tank maintenance,” *Aerospace*, Vol. 8, No. 5, p. 136, 2021

[37] M. Hrúz, M. Bugaj, A. Novák, B. Kandera, and B. Badánik, “The use of UAV with infrared camera and RFID for airframe condition monitoring,” *Applied Sciences*, Vol. 11, No. 9, p. 3737, 2021

[38] Z. M. Çınar, A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, “Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0,” *Sustainability*, Vol. 12, No. 19, p. 8211, 2020

[39] C. Boller, “Ways and options for aircraft structural health management,” *Smart materials and structures*, Vol. 10, No. 3, p. 432, 2001

[40] T. Tyncherov and L. Rozkova, “Aircraft lifecycle digital twin for defects prediction accuracy improvement,” in *International Conference on Reliability and Statistics in Transportation and Communication*, Springer, 2019, pp. 54-63

[41] A. Y. Yurin, Y. V. Kotlov, and V. M. Popov, “The conception of an intelligent system for troubleshooting an aircraft,” 2021

[42] S. Schmid, U. Martens, W. K. Schomburg, and K.-U. Schröder, “Integration of eddy current sensors into repair patches for fatigue reinforcement at rivet holes,” *Strain*, Vol. 57, No. 5, p. e12387, 2021

[43] D. Wakeman, “Project prioritization: How to prioritize projects & strategy,” ProjectManager. Accessed: Dec. 21, 2023 [Online] Available: https://www.projectmanager.com/blog/how-to-prioritize-projects-and-strategy

Spatial Arrangement of a Counter-Rotating Dual Rotor Wind Turbine

Csaba Hetyei, Ferenc Szlivka, Ildikó Molnár

Óbuda University, Doctoral School on Safety and Security Sciences and Bánki Donát Faculty of Mechanical and Safety Engineering, Népszínház utca 8, 1081 Budapest, Hungary; hetyei.csaba@bgk.uni-obuda.hu, szlikva.ferenc@bgk.uni-obuda.hu, molnar.ildiko@bgk.uni-obuda-hu

Abstract: Nowadays increasing energy demand and the current energy crisis in Europe highlighted the need for independent and cheap energy sources which can be produced at the place of use. A good example of this energy sources are the renewables, from which wind energy is one. Humanity is using wind energy since the beginning of the history, but electrical energy generation from the wind started at the end of the 19th Century. During the evolution of wind energy utilization, wind turbines are becoming more and more efficient. A special kind of these turbines are the non-conventional wind turbines which are aiming to be efficient in a special condition. One of these new turbine designs is the CO-DRWT (Counter-Rotating Dual Rotor Wind Turbine), where there are two rotors in one tower. During our research, we examined some layouts for a CO-DRWT. In these spatial arrangements, we were able to change the second rotor's axial and radial positions. Within two in axial and one diameter in the radial region, we were running CFD (Computational Fluid Dynamics) simulation to determine the interaction of the two turbines and for calculating the overall power coefficient (c_P) for the two rotors. Meanwhile, in our analysis, we defined some spatial arrangements where the CO-DRWT's overall c_P is less than the c_P of a Single Rotor Wind Turbine (SRWT) from the same geometry. We also defined regions where the CO-DRWT's c_P is higher than the SRWT's. With our geometry and with our simulation's boundary conditions we find the optimal place for operating a CO-DRWT is the R = 0D radial distance with A = 2.1D axial distance (where the D is the rotor's diameter) where the c_P is 0.514, also the worst arrangement is the R = 0.35D with A = 1.25D where the c_P = 0.354, while an SRWT's c_P from the same geometry is 0.377. According to our simulations, the energy density and the power coefficient of an optimized CO-DRWT are 1.363 times higher than an SRWT has.

Keywords: CFD, CO-DRWT; Design of Experiments; DOE; Dual Rotor Wind Turbine; Optimisation

1 Introduction

The utilization of wind energy has a long history. It started with the sailing and the wind-powered organ of the Hero of Alexandrina. The first known windmill was built in Nastifan in the 9th Century for grinding. In Europe, windmills started to spread in the 12th Century. Until the 19th Century, windmills were used for grinding or lifting water [1]. In 1887, James Blyth built the first Vertical Axis Wind Turbine (VAWT) to generate energy in his rear garden in Marykirk for energy generation. In 1888, Charles Brush built the first Horizontal Axis Wind Turbine (HAWT) to generate energy in Cleveland [3]. Charles Brush's wind turbine has "only" 12 kW capacity [3], while nowadays typical turbine capacities are in the MW scale, thanks to research and developments.

An indicator of wind energy utilisation is the total installed turbine capacity, which was 24 GW in 2001, 238 GW in 2011, 488 GW in 2016, and 837 GW in 2021. The mostly installed turbine capacity is onshore but offshore installation is also possible. Currently, the total installed turbine's capacity is 780 GW onshore and 57 GW offshore [5].

Generally, and during energy crises [6], researchers and energy providers try to find solutions to meet the energy demand. To solve the necessary electricity supply for users, wind turbine developers optimise their turbines for different environments [7], e.g., there are diffusers to catch more wind or to increase the wind's kinetic energy [8] [9], or there are airfoil [10] and blade designs [11] for specific environments, or there are new places where the wind turbines can produce electricity like a solar chimney [12] or like a turbine installed on buildings in cities [13].

Besides the core turbine design and optimisation, new kinds of wind turbines also appear in the energy generation marketplace, which are unconventional wind turbines. These types of new wind turbines are modified in some respects. Two examples in this category are the Dual Rotor Wind Turbines [14] [15] which are multiple rotor turbines made from traditional and the second is the modified rotors, e.g., the Archimedes Screw Turbine [16], which are highly modified turbines for special environments and their needs.

Next to the research and development of wind turbines, there were also developments in turbine-related industries and products, such as operation, performance, and diagnostic monitoring [17] [18].

The previous examples mainly focused on the horizontal axis wind turbines, but there are energy-generating systems containing more renewable and non-renewable sources, which can be installed in the urban and the non-urban zones. In our energy needy system, the smallest energy-consuming unit can be a single house [19] which can produce energy with solar, geothermal or heat pumps [20] for families.

2 Different Impeller Layouts for Higher Extractable Power

In our research, we analysed a Counter-Rotating Dual Rotor Wind Turbine (CO-DRWT), which is an unconventional wind turbine. We chose this wind turbine type because it has a bigger performance than a single or a Co-Rotating Dual Rotor Wind Turbine [14] [22]. We used a CO-DRWT for our simulations, which we used in our previous studies. Firstly, we created a wind turbine geometry which we mirrored. The original part is our first rotor and the second is the mirrored one. These rotors are rotate in opposite directions. Therefore if the radial gap is 0.5 in diameter (100 mm) or more, the second blade is not covered by the first. This turbine is shown in the following figure.

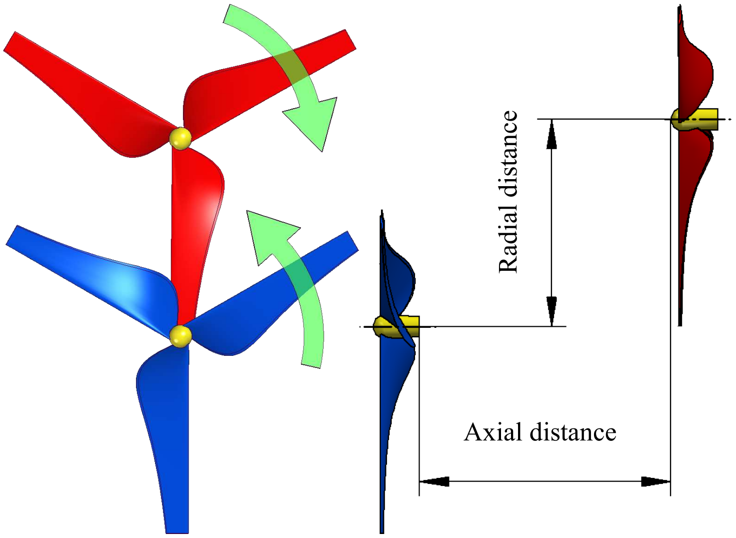


Figure 1
CO-DRWT with the indication of the rotors' rotating direction and the variable axial and radial distances

In Figure 1, the rotational directions are shown with green arrows. For our research, we were able to change the axial and radial distances between the CO-DRWT's rotors, these distances are indicated. The distances were de-dimensioned with the rotor's diameter, which was 200 mm, therefore, the R=1D distance is 1·200 mm=200 mm in the radial direction. The minimal distance was 0.005D and the maximum was 2D for the axial distance and 0D (no offset, single axis) was the minimal and 1D for the radial distance.

To measure the wind turbines efficiency, we used the power coefficient (c_p) for our CO-DRWT during our tests which can be calculated from the torque on the blades and from the incoming flow with the following equation:

$$c_p = \frac{P_{turbines}}{P_{wind}} = \frac{T_1 \cdot \omega_1 + T_2 \cdot \omega_2}{\frac{1}{2} \cdot \rho \cdot A \cdot v^3} = \frac{P_{turbine\ 1} + P_{turbine\ 2}}{P_{wind}} = \frac{P_{turbine\ 1}}{P_{wind}} + \frac{P_{turbine\ 2}}{P_{wind}} = c_{p1} + c_{p2} \quad (1)$$

In the previous equation c_p is the CO-DRWT overall power coefficient, c_{p1} and c_{p2} are the power coefficient of the first and the second turbine, $P_{turbines}$ is the CO-DRWT overall performance, $P_{turbine1}$ and $P_{turbine2}$ are the first and the second turbine's performance and P_{wind} is the wind performance. T_1 and T_2 are the torque on the first and on the second turbine, ω_1 and ω_2 are the angular velocity of the first and the second turbine. ρ, is the density of the air, A is the swept area of the wind turbine's blade and v is the wind's velocity in the freestream area.

The swept area in the previous equation depends on the radial shift of the two turbines, and its value is between $d^2 \cdot \pi/4$ and $2 \cdot d^2 \cdot \pi/4$, where d is the diameter of the turbine. This area is shown in the next figure.

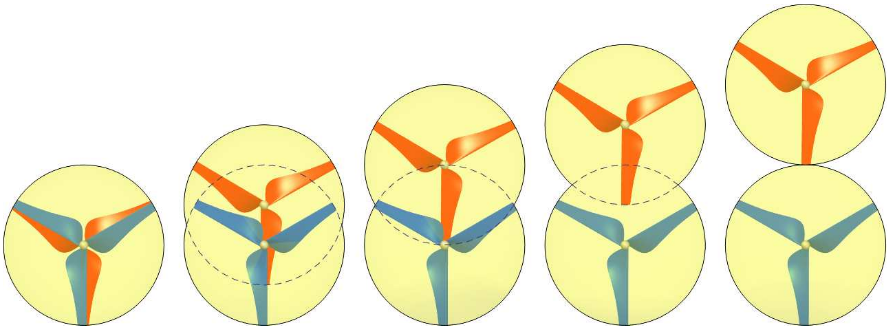


Figure 2

Swept area for different CO-DRWT layouts (from 0 to 1 diameter radial shift)

In our previous research, we measured the torque on the turbines, then we calculated the power coefficient [23]. After the physical testing, we started to use CFD software to simulate the different axial distances [24]. For our simulation, we used Reynolds Averaged Navies Stokes (RANS) equations-based CFD software. This numerical simulation method is based on the continuity, momentum, and energy equations which were solved iteratively with the SIMPLE algorithm.

A single rotor's maximum power coefficient (c_p) by the Betz law is c_p=16/27≈59.259%. The Betz law is a theoretical limit for a wind turbine with an ideal flow and boundary conditions, where the turbine has an infinite number of blades. The Betz law was created at the beginning of the 20th Century. 8 decades after Betz, Gorban *et al.* created their model, known as the GGS model. By the GGS model, a wind turbine's maximum power coefficient (c_p) is c_p=30.113 % [25]. Using CFD simulations and measurement, the wind turbines' c_p is between these two values provided by the Betz law and the GGS model.

3 Simulation Parameters

For our simulations [24] we used Simcenter FLOEFD with the same boundary conditions as we used for the measurements [23]. The wind velocity in the freestream region (v_∞) was 3.79 m/s, the ambient pressure was 1 atm, the fluid was "air" from the CFD software's database, and the temperature was 20 C. The tip speed ratio (λ), which is the ratio of the wind turbine's angular velocity and the free stream velocity, was 4. The tip speed ratio can be calculated with the following equation:

$$\lambda = \frac{\omega \cdot R}{v_\infty} \tag{2}$$

In the previous equation, the λ is the tip speed ratio, ω is the angular velocity, R is the radius of the blade, and v_∞ is the freestream velocity.

For the simulations, we used a rectangular domain, within which we used Cartesian mesh with polyhedral elements on the surfaces of the wind turbines. The mesh contains 2.8-3 million elements depending on the CO-DRWT's position.

The k-ε turbulence model was used for the turbulence modelling. For validation, we ran simulations in steady and unsteady states, but for the optimization of the spatial arrangement, we used only the steady-state results. In the steady-state, to model the turbine's rotation we used the Mixing Plane method.

During our simulations, we monitored the torque and the static pressure on the turbines' blades, and the averaged and maximum velocity, static, and total pressure in the whole computational domain and in the rotating regions. We used these parameters as finishing conditions. If all the parameters converged and the simulation ran for at least 10,000 iterations, the calculation ended.

4 CFD Results

After running our simulations, in the turbines' region, we had a similar flow field. Near the wind turbines' region in the wake region, the velocity was generally lower than in the freestream region. The first turbine slowed down the incoming air by taking out the air's kinetic energy to the rotational motion. The flow which reached the second turbine was turbulent and slower than the wind which arrived at the first turbine. Depending on the configuration the wake region's shape changed. Typical velocity distribution of the CO-DRWT is shown in the next figure.

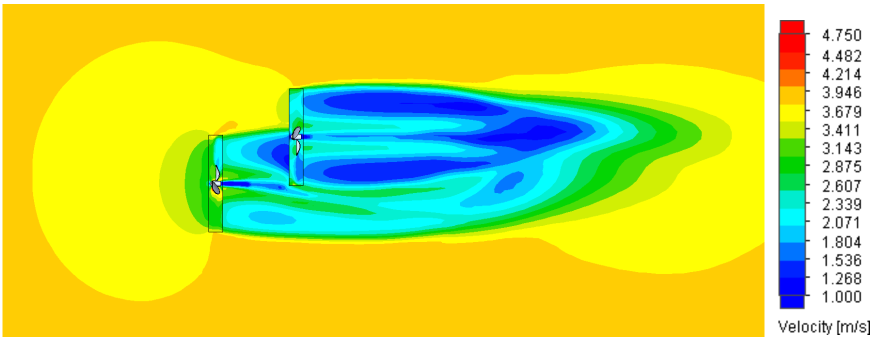


Figure 3
Flow field in the turbines' region (velocity distribution, steady state, A = 0.5D, R = 0.75D axial and radial distance)

Using the (1) equation we were able to calculate the power coefficient for each rotor (c_{p1} and c_{p2}) and the overall power coefficient (c_p) of the CO-DRWT. In the following figures the c_{p1}, c_{p2}, and c_p are shown for the CO-DRWT different axial and radial shifts.

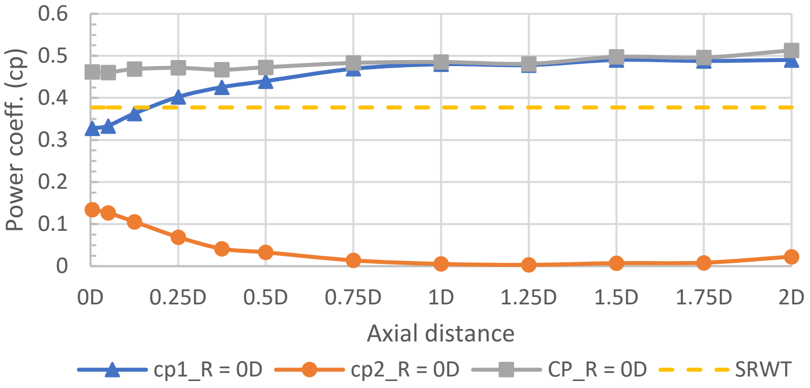


Figure 4
Power coefficient for the first (c_{p1}), second (c_{p2}) rotors and the overall c_p for the CO-DRWT with R=0D radial distance (0 mm)

In Figure 4 the CO-DRWT's radial shift was 0D, therefore, the rotors were coaxial. The axial gaps were between 0.005D and 2D. The power coefficient of the first turbine increased, while the power coefficient of the second rotor (c_{p2}) decreased with the axial distance. The yellow dashed line is the power coefficient of a single rotor turbine (SRWT), which was simulated with the same geometry. The power coefficient of the second rotor (c_{p2}) was in each case lower than the power coefficient of the SRWT (c_{p_SRWT}), while the power coefficient (c_{p1}) of the first rotor was higher than the c_p of the SWRT after A≈0.15D axial distance. The overall c_p of the CO-DRWT was higher than the SRWT's in each configuration.

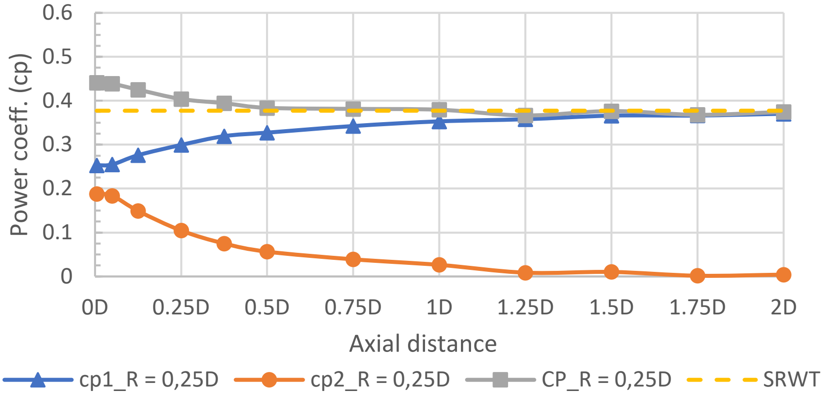


Figure 5

Power coefficient for the first (c_{p1}), second (c_{p2}) rotors and the overall c_p for the CO-DRWT with R=0.25D radial distance (50 mm)

In the previous figure (Figure 5) the radial distance was R=0.25D (50 mm). In this case, the c_{p1} increased and the c_{p2} decreased with the growth of the axial distance, like in the case of R=0D case. The overall c_p of the CO-DRWT was higher than the power coefficient of the SRWT (c_{p_SRWT}), with a relatively small axial gap. While the axial distance rise, the c_p of the CO-DRWT decreased. Near A=1.25D axial distance, the power coefficient of the CO-DRWT decreased lower than the c_p. of the SRWT. Between the A=0.75D and the A=2D axial distances, the power coefficient of the CO-DRWT was similar to the SRWT.

Comparing this case to the R=0D, the c_{p1} was lower in each axial distance than in the R=0D and the c_{p2} too.

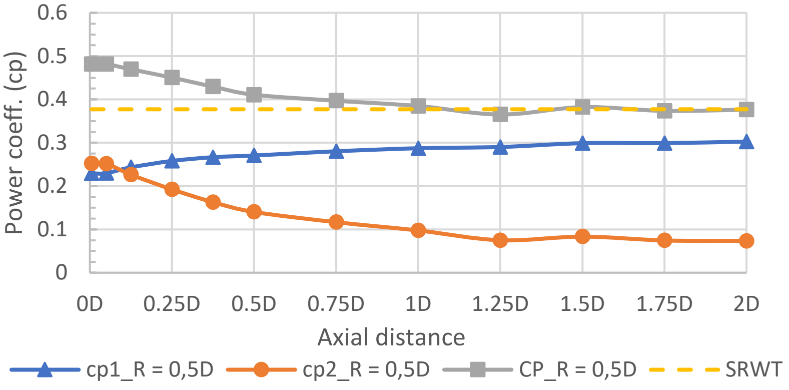


Figure 6

Power coefficient for the first (c_{p1}), second (c_{p2}) rotors and the overall c_p for the CO-DRWT with R=0.5D radial distance (100 mm)

In Figure 6 the c_{p1} increased with the axial distance like in the R=0D and R=0.25D cases, but the difference between the rise between the starting and the end distance was smaller. The slope of the c_{p2}, compared to the previous cases (R=0D, R=0.25D) was also smaller. In comparison with the two previous cases, the starting values of the c_{p1} was lower and the c_{p2} was higher.

The overall c_p for the CO-DRWT was similar to the R=0.25D but its values were different. It started with a high value which decreased with the growth of the axial distance. In the region R=1D and R=2D, the CO-DRWT's overall power coefficient was similar to the SRWT's c_p.

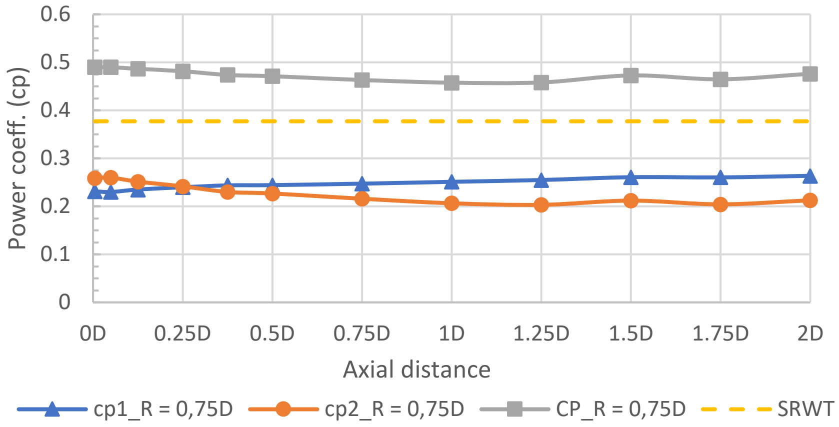


Figure 7

Power coefficient for the first (c_{p1}), second (c_{p2}) rotors and the overall c_p for the CO-DRWT with R=0.75D radial distance (150 mm)

In Figure 7, where the CO-DRWT has R=0.75 radial distance, the c_{p1} and the c_{p2} show a rise and a fall, but the differences between the two ends are smaller than they were in the R=0D, R=0.25D, and R=0.5D cases. The c_{p1} and the c_{p2} values are similar. They are close to each other.

The overall c_p of the CO-DRWT for each configuration is higher than the power coefficient of the SRWT (c_{p_SRWT}). The curve of the c_p of the CO-DRWT also decreased.

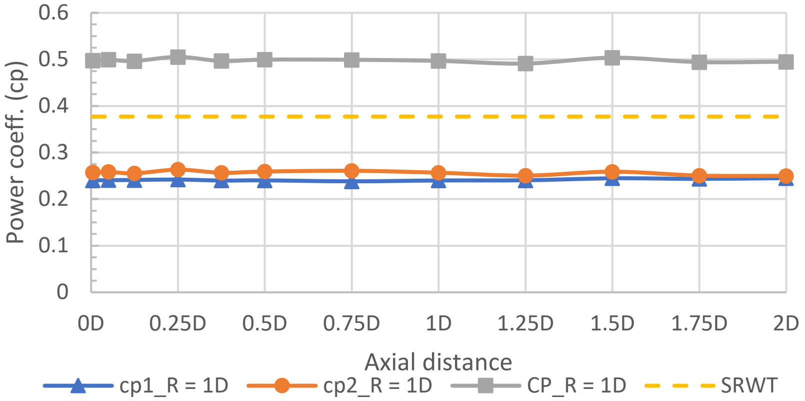


Figure 8

Power coefficient for the first (c_{p1}), second (c_{p2}) rotors and the overall c_p for the CO-DRWT with R=1D radial distance (200 mm)

In Figure 8 the radial distance was 1 diameter. In this case, the swept area of the first rotor does not cover the second rotor and its swept area. The power coefficient of the two rotors was similar and with the growth of the axial distance, they do not change much. In this case, the two rotors have some effects on each other, because the power coefficients were not the same as the c_p of the SRWT (c_{p_SRWT}). The overall c_p of the CO-DRWT was higher in each case than the SRWT's.

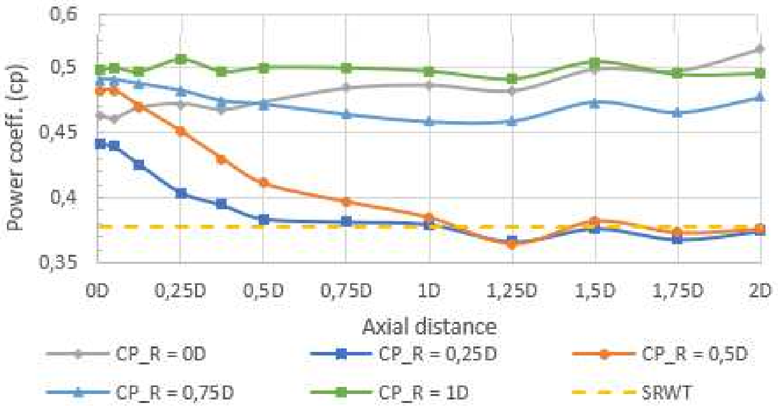


Figure 9

Overall power coefficients for the different CO-DRWT configurations

In Figure 9 the overall c_p of the CO-DRWT are summarized and compared with the c_p of the SRWT (c_{p_SRWT}). We can observe the power coefficients,

- in the R=0D case (when the two rotors have the same axis) the overall c_p shows a rise. We assume that the reason for this increase in the c_p is because of the influence of the second rotor on the first rotor. This can be seen in Fig. 4, where the c_{p1} increased more than the power coefficient of the SRWT, while the c_{p2} decreased (almost to zero).

- in the R=0.25D and in the R=0.5D we can see a similar slope for the power coefficients. The c_p starting values are higher than they decreased and in some cases, its values are lower than the SRWT's.
- in the R=0.75D we can see a slope in the first half of the examined axial distance, then after the A=1D axial distance, the overall c_p rose.
- in the R=1D case (when the two rotors' swept areas do not cover each other), the overall c_p was almost the same for each axial distance. We assume the reason for this was the flow which reached the second turbine. This flow (in the second turbine's region) was not turbulent and did not disturb as in the R=0.25D, R=0.5D and the R=0.75D cases.

5 Surface Fittings on the CFD's Results

Using our CFD results shown in Figure 9 [24], we created surfaces for layout optimisation. In the next figures, we used a CAD system (Solid Edge) with a self-made coordinate system for easier representation. In Fig. 10 a surface is shown with a cubic interpolation, based on the CFD results.

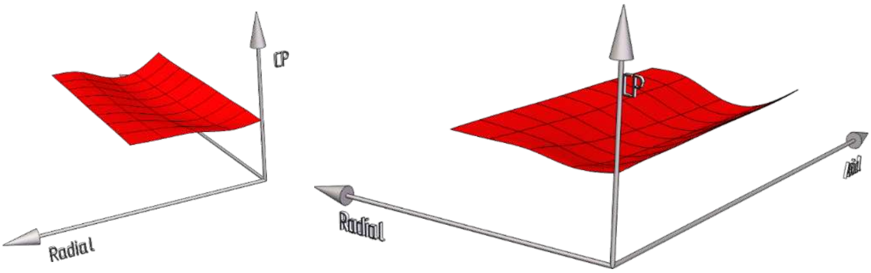


Figure 10

The fitted surface of the overall power coefficients for the different CO-DRWT configurations

The surface from Figure 10 was not appropriate for optimisation because the highest c_p was the highest simulated c_p at the R=0D and A=2D position. For the previous reason, we increased our surface for the optimisation process with the following considerations:

- In the negative direction of the "Axial" axis, we mirrored the power coefficients with negative values (thereby the plane for the mirror was the Axial-Radial plane).
- In the positive direction of the "Axial" axis, we copied the power coefficients' values until the A=2.5D distance without changing its values.

- In the negative direction of the "Radial" axis, we mirrored the power coefficients without changing their values (thereby the plane for the mirror was the c_p-Axial plane).
- In the positive direction of the "Radial" axis, we copied the power coefficients' values until R=1.5D distance without changing their values.

The surface created with our assumption is shown in the following figure.

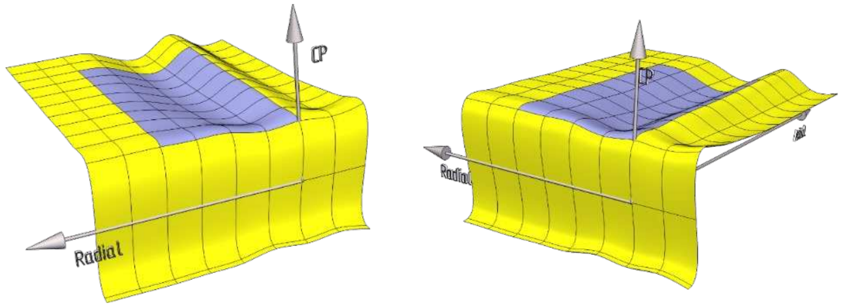


Figure 11
CO-DRWT's power coefficient in the region augmented by assumptions

In Figure 11 the original zone of our simulation is coloured in purple, meanwhile, the region of the assumption is yellow. The original surface from Figure 10 and the enlarged surface from the Figure 11 are different due to the different boundaries. The differences are shown in Figure 12.

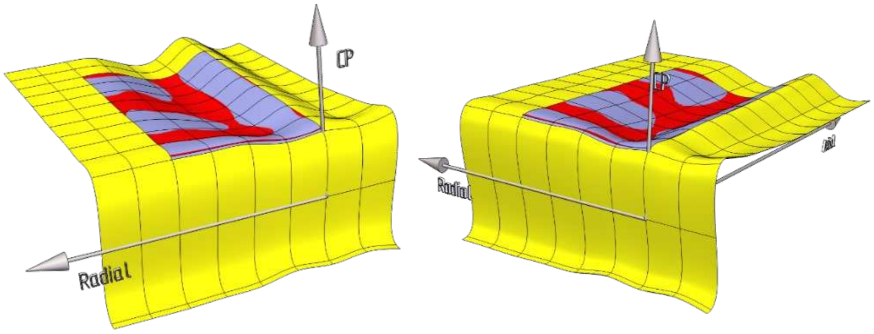


Figure 12
Original vs. augmented surface

In Figure 12 we could observe that the red (original) surface was higher in some regions than the purple one (augmented surface with the assumptions). In those regions where the purple surface covers the red, the augmented surface has higher c_p-s due to the surface fitting methodology.

For comparison, we created a surface for a Single Rotor Wind Turbine too. As it was expected (based on the results from Fig. 4 to Fig. 9), in some regions, this surface is higher than the surface, which is increased by our assumption. In Figure 13 the SRWT's surface is coloured green while the surface with the augmented values has the same colour as before.

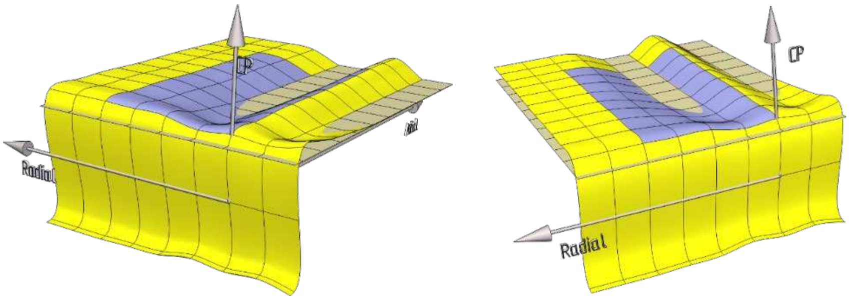


Figure 13
Augmented surface vs. a SRWT's c_p

For the optimisation process, we only used the overall power coefficient of the CO-DRWT (c_p), therefore, the c_{p1} and c_{p2} values were ignored.

For our optimisation process, we created an optimisation script in MATLABA 2022a, where we used the cubicinterp, poly33 and poly55 methods. Using these surface fitting methods, the highest power coefficients and their positions are shown in the following table.

Table 1
Highest power coefficients with their positions

Interpolation type	Highest power coefficient (c_p)	Radial distance for the highest c_p	Axial distance for the highest c_p
poly33	0.634	1.5D	0.9D
poly55	0.649	1.4D	0.35D
cubicinterp	0.514	0D	2.1D

The poly33 and poly55 methods create polynomial surfaces which are based on the input c_p-s, but the surfaces do not lie on the entered points. While the cubicinterp method creates a surface with cubic spline interpolation, where the surface fits with the input data. From the previous interpolations, the surface lies on the input data only with the cubicinterp method, which we chose for our optimisation.

The surfaces with are created with poly33 and poly55 algorithms are shown in Figure 14.

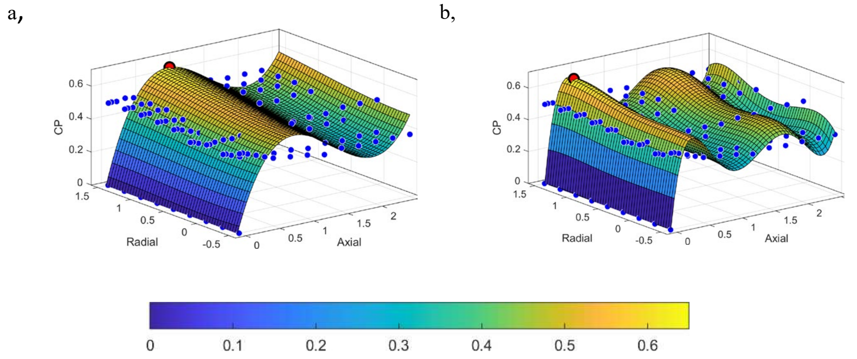


Figure 14

CO-DRWT's maximum power coefficient on a surface which was created by polynomial interpolation. a, surface with poly33; b, surface with poly55

The surface, which is created with a cubic spline interpolation is shown in Figs. 15-16. The surface is coloured by its c_p value, the maximum value is marked with a black-edged red dot, while the input data points are represented with a blue dot.

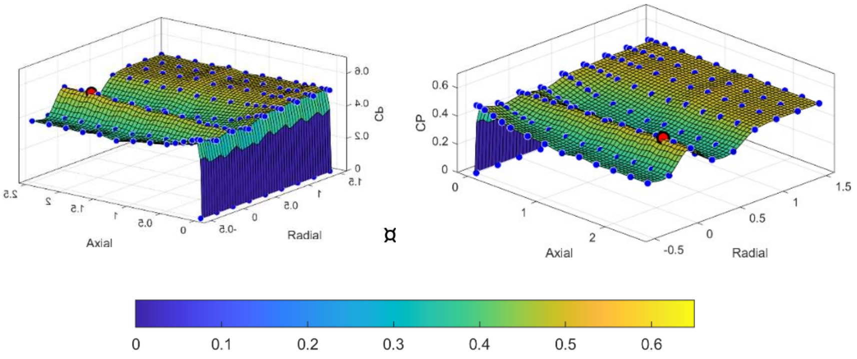


Figure 15

CO-DRWT's maximum power coefficient on a surface which was created with a cubic spline interpolation

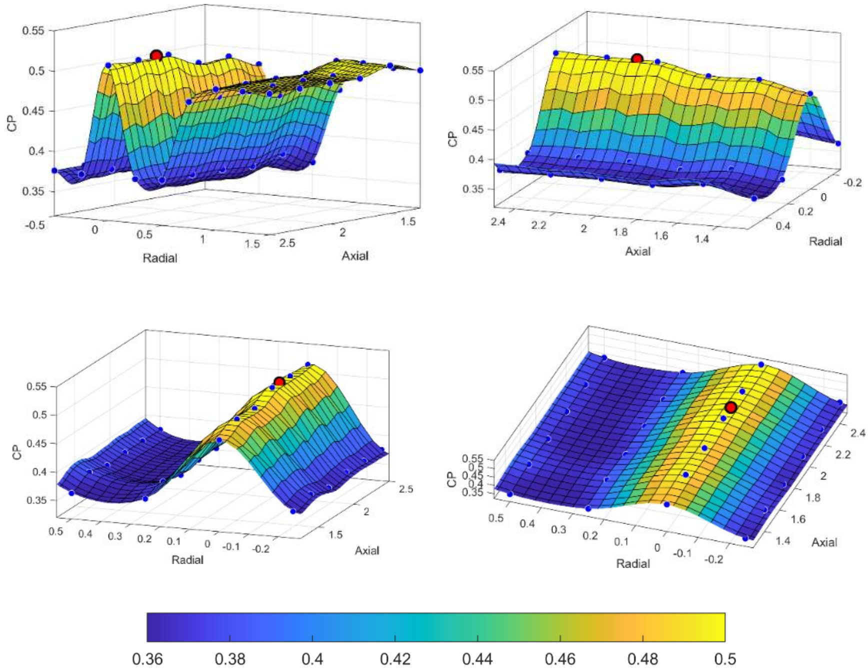


Figure 16

CO-DRWT's maximum power coefficient and its region

As in the previous figures (Figs. 4-9 and Fig. 13), the c_p is lower in some regions than the SRWT's power coefficient (c_{p_SRWT}). We changed our script to determine the worst-case layout. The minimum search algorithm looked for the minimum value in the region of the original simulations (from A=0.005D to A=2D and from R=0D to R=1D). We used this limitation because when we enlarge our surfaces in the negative Axial direction, we mirrored our results with a negative value, therefore this region had the lowest overall power coefficients on the surface.

The lowest overall power coefficient of the CO-DRWT in the region of the simulations was c_p=0.354 at the R=0.35D radial with A =1.25D axial distance. In Figs. 17 and 18 the minimum value is marked on the surface with a black dot with a red border.

The minimum and the maximum values are shown on the surface with their previous marks (the maximum is a red dot with a black corner, and the minimum is a black dot with a red border).

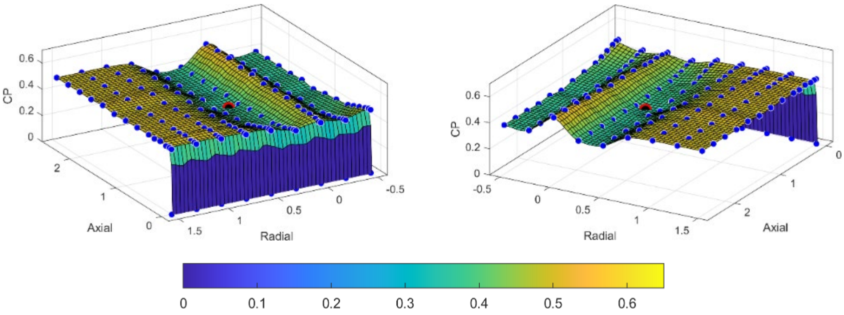


Figure 17

CO-DRWT's minimum power coefficient on a surface which was created with a cubic spline interpolation

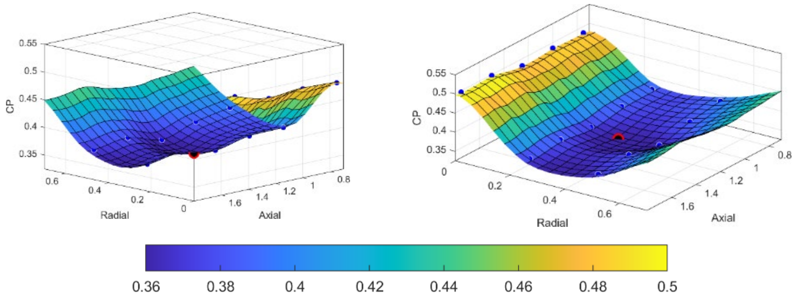


Figure 18

CO-DRWT's minimum power coefficient and its region

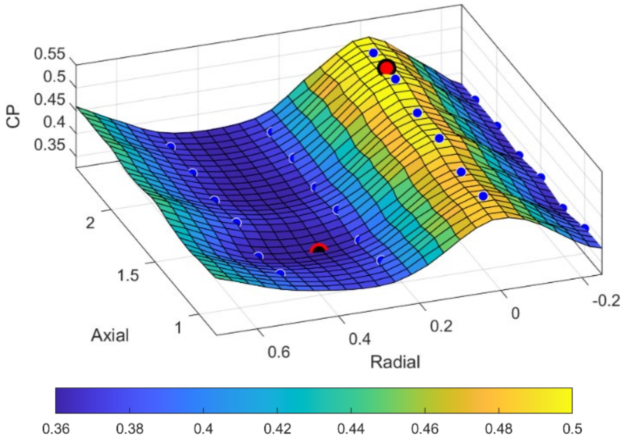


Figure 19

CO-DRWT's maximum and minimum power coefficient and their region

To find the regions which are more efficient than the SRWT, we recoloured the previous figure with the limit of the power coefficient of the SRWT's (c_{p_SRWT}=0.37727) and then subtracted the regions which are lower than the SRWT's c_p. The regions which have a higher power coefficient than an SRWT are shown in the following figures.

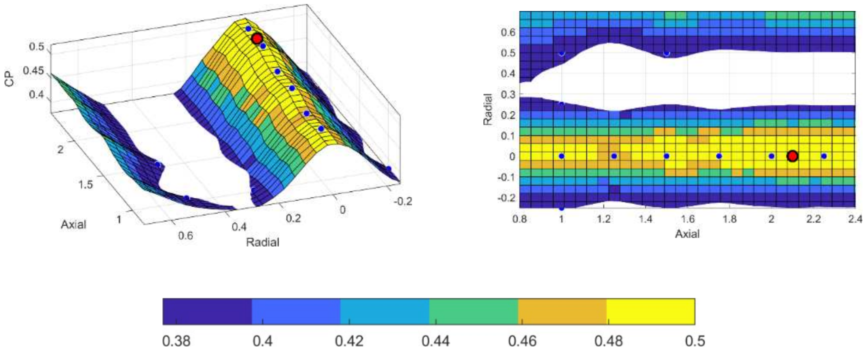


Figure 20

Power coefficients and their regions which are higher than an SRWT's c_p (near the surface's maximum)

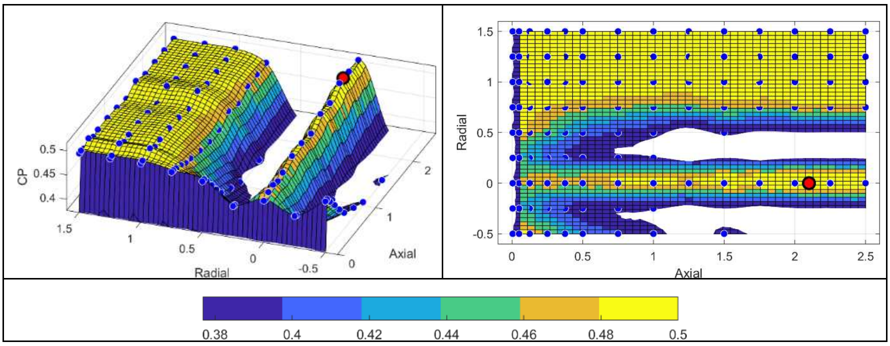


Figure 21

Power coefficients and their regions which are higher than an SRWT's c_p

If we limit the radial and axial axes to the simulation's original region (A=0.005D to A=2D and R=0D to R=2D) we have the power coefficient distribution, which is shown in the following figure.

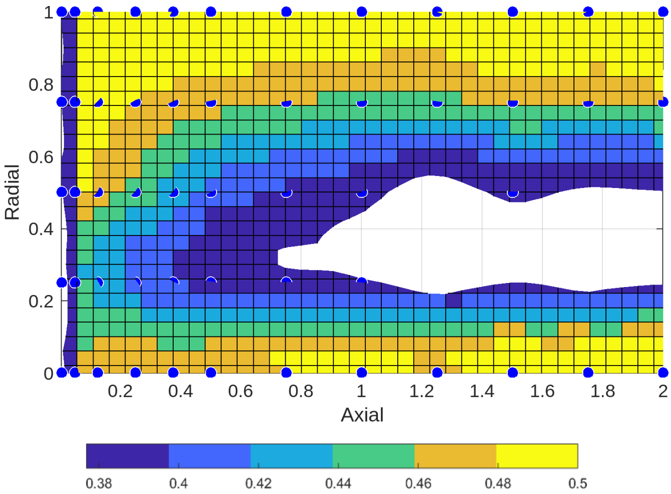


Figure 22

Power coefficients and its regions which are higher than an SRWT's c_p

With the previous figures (Fig. 21 and Fig. 22) we can establish the following:

- The CO-DRWTs in most layouts produce more electricity (based on their overall power coefficient) than a Single Rotor Wind Turbine.
- Between the approx. from R=0.2D and R=0.6D radial distances there is a region where the CO-DRWT's power coefficient is less than an SRTW's.
- Small radial distances (approx. from R=0D to 0.1D) and high radial distances (approx. from R= 0.7D) have a good effect on the CO-DRWT's c_p.

Conclusions

In our paper, we presented an optimisation for a CO-DRWT's (Counter-Rotating Dual Rotor Wind Turbine) spatial arrangement. During our research, we created several layouts for a CO-DRWT which we used within CFD studies. Using the results of the numerical simulation we created surfaces with different interpolation techniques, where we chose a cubic spline interpolation.

For the optimisation method, we used a script for the best and the worst cases. With this script, based on our simulations with our geometry and our boundary conditions, we find the R=0D radial distance with the A=2.1D axial distance has the highest overall power coefficient (c_p=0.514) for the CO-DRWT, while the R=0.35D with A=1.25D distance has the lowest overall power coefficient (c_p=0.354).

By comparison, the c_p of an SRWT (Single Rotor Wind Turbine) which was made with the same geometry and with the same simulation parameters is 0.377. We find some regions where the overall power coefficient of the CO-DRWT is less than the SRWT's, but in most regions, the CO-DRWT's c_p is higher.

Using our results (Figs. 20 and 21) we determined regions where a CO-DRWT has a higher power coefficient than a Single Rotor Wind Turbine. Using this "heat map" we are able to design a small dual-rotor wind turbine which requires less space than two SRWTs have. Therefore, in an urbanized region, it could generate more energy than an SRWT, or if it is used in wind farms the farm could have a higher energy density due to the CO-DRWT's lower space requirement than using traditional wind turbines.

References

[1] F. Szlivka, I. Molnár, "Víz- és szélenergia hasznosítás (Hydro and wind energy utilization),“ Edutus Főiskola Kiadó, 2012

[2] T. J. Price, "Blyth, James (1839-1906)," Oxford University Press, https://doi.org/10.1093/ref:odnb/100957 (Access Date: 08. 09. 2022)

[3] R. W., Righter, "Wind Energy in America: A History," University of Oklahoma Press, ISBN: 9780806128122, 1996, https://books.google.hu/books?id=kGnGw7AEkAEC

[4] Wind Turbines: the Bigger, the Better, Office of Energy Efficiency & Renewable Energy (online) https://www.energy.gov/eere/articles/wind-turbines-bigger-better (Access Date: 26. 09. 2022)

[5] Global Wind Energy Council, 'Global Wind Report 2022," Brussels (Belgium), p. 111, 2022, https://gwec.net/global-wind-report-2022/

[6] Patrick Jackson, "Ukraine war: EU moves to cut peak electricity use by 5%," BBC News (online), https://www.bbc.com/news/world-europe-62899940 (Access Date: 26.09.2022)

[7] K. R. Kumar, M. Selvaraj, "Review on Energy Enhancement Techniques of Wind Turbine System," Advances in Science and Technology, 106, pp 121-130, 2021, https://doi.org/10.4028/www.scientific.net/AST.106.121

[8] A. Alonso-Estébanez, P. Pascual-Muñoz, F. P. Alvarez Rabanal, D. Castro-Fresno and J. J. Del Coz Díaz, "New System for the Acceleration of the Airflow in Wind Turbines," Recent Patents on Mechanical Engineering, 12(2), pp. 158-167, 2019, http://dx.doi.org/10.2174/2212797612666190311154747

[9] M. Anbarsooz, M. Amiri, I. Rashidi, "A novel curtain design to enhance the aerodynamic performance of Invelox: A steady-RANS numerical simulation", Energy, 168, pp. 207-221, 2019, https://doi.org/10.1016/j.energy.2018.11.122

[10] T. Lutz, "Airfoil Design and Optimisation," ZAMM Journal of applied mathematics and mechanics: Zeitschrift für angewandte Mathematik und Mechanik, 81(S3), pp. 787-788, 2001, http://dx.doi.org/10.1002/zamm.200108115166

[11] V. M. Kumar, B N. Rao, Sk. Farooq, "Modeling and analysis of wind turbine blade with advanced materials by simulation", International Journal of Applied Engineering Research, 11(6), pp. 4491-4499, 2016, https://www.ripublication.com/ijaer16/ijaerv11n6_128.pdf

[12] W. M. A-Elmagid, I. Keppler, I. Molnár, "Efficient Axial Flow Turbine for Solar Chimney," Journal of Thermal Science and Engineering Applications, 12(3), p. 031012, 2020, https://doi.org/10.1115/1.4044903

[13] I. Molnár, F. Szlivka, G. Sándor, "Advantages and disadvantages of different types of wind turbines their usage in the city," WinerCost'17: International Conference on Wind Energy Harvesting, Coimbra, Portugal, April 20-21, 2017, pp. 269-271, http://www.winercost.com/cost_files/WINERCOST17_Proceedings_Book.pdf

[14] A. Ozbay, W. Tian, H. Hu, "Experimental Investigation on the Wake Characteristics and Aeromechanics of Dual-Rotor Wind Turbines," Journal of Engineering for Gas Turbines and Power, 138(4), pp. 1-15, 2016, https://doi.org/10.1115/1.4031476

[15] E. Erturk, S. Sivrioglu és F. C. Bolat, „Analysis Model of a Small Scale Counter-Rotating Dual Rotor Wind Turbine with Double Rotational Generator Armature," International Journal of Renewable Energy Research, 8(4), pp. 1849-1858, 2018, https://www.ijrer.org/ijrer/index.php/ijrer/article/view/8235

[16] H. Jang, D. Kim, Y. Hwang, I. Paek, S. Kim, J. Baek, "Analysis of Archimedes Spiral Wind Turbine Performance by Simulation and Field Test," Energies, 12(24), 4624, 2019, https://doi.org/10.3390/en12244624

[17] Nagy A., Jahn I., „Advanced Data Acquisition System for Wind Energy Applications," Periodica Polytechnica Transportation Engineering, 47(2), pp. 124-130, 2019, https://doi.org/10.3311/PPtr.11515

[18] S. Butler, J. Ringwood, F. O'Connor, "Exploiting SCADA system data for wind turbine performance monitoring," 2013 Conference on Control and Fault-Tolerant Systems (SysTol), pp. 389-394, 2013, http://dx.doi.org/10.1109/SysTol.2013.6693951

[19] R. Dziugaite, V. Jankauskas, V. Motuziene, "Energy Balance of a Low Energy House," Journal of Civil Engineering and Management, 18(3), 2012, pp. 369-377, http://dx.doi.org/10.3846/13923730.2012.691107

[20] H. F. Ummah, R. Setiati, Y. B. V. Dadi, M. N. Ariq, M. T. Malinda, "Solar energy as natural resource utilization in urban areas: Solar energy efficiency literature review," IOP Conference Series: Earth and Environmental Science, 780, 2021, 012007, http://dx.doi.org/10.1088/1755-1315/780/1/012007

[21] Sánta, R. "Comparative Analysis of Heat Pump System with IHX Using R1234yf and R134a," Periodica Polytechnica Mechanical Engineering, 65(4), pp. 363-373, 2021, https://doi.org/10.3311/PPme.18390

[22] A. M. Labib, A. A. Gawad és M. M. Nasseif, „Effect of Aspect Ratio on Aerodynamic Performance of Archimedes Spiral Wind Turbine,” EIJEST, 32, pp. 66-72, 2020, https://dx.doi.org/10.21608/eijest.2020.45256.1017

[23] F. Szlivka, I. Molnár, P. Kajtár, G. Telekes, “CFX Simulations by Twin Wind Turbine,” 2011 International Conference on Electrical and Control Engineering, 2011, Yichang, China, 16-18 Sept. 2011, pp. 5780-5783, https://doi.org/10.1109/ICECENG.2011.6057550

[24] Cs.Hetyei, F. Szlikva, “Counter-Rotating Dual Rotor Wind Turbine Layout Optimisation”, Acta Polytechnica, 61(2), pp. 342-349, 2021, https://doi.org/10.14311/AP.2021.61.0342

[25] A. N. Gorban, A. M. Gorlov, V. M. Silantyev “Limits of the Turbine Efficiency for Free Fluid Flow,” Journal of Energy Resources Technology, 123(4), pp. 311-317, 2001, https://doi.org/10.1115/1.1414137

[26] “List of Library Models for Curve and Surface Fitting,” MathWorks (online), https://www.mathworks.com/help/curvefit/list-of-library-models-for-curve-and-surface-fitting.html (Access date: 13.07.2022)

Feed-Forward and Long Short-Term Neural Network Models for Power System State Estimation

Tuan-Ho Le

Faculty of Engineering and Technology, Quy Nhon University
170 An Duong Vuong, 55100, Quy Nhon City, Binh Dinh Province, Vietnam
tuanhole@qnu.edu.vn

Abstract: The primary objective of this paper is to propose the two new combined approaches based on Feed-Forward and Long Short-Term Memory Neural Network models for Power System State Estimation. First, the Weighted Least Square method and the Generalized Maximum-Likelihood Estimator using the Projection statistics method are used to estimate the voltage magnitude and phase angle. Secondly, the Feed-Forward Neural Network model is proposed to combine the obtained voltages and angles. The optimal structure of the proposed Feed-Forward Neural Network model is defined based on the Akaike Information Criterion. Thirdly, the Long Short-Term Neural Network model is proposed as an alternative hybrid power system state estimation approach. Finally, the different case studies including IEEE 9-bus system and IEEE 14-bus system are used to validate the effectiveness of the proposed approaches. The final results imply that the proposed approaches can provide more effective solutions than the existing approaches according to Mean Absolute Percentage Error and Weighted Average Percentage Error criteria.

Keywords: Power System State Estimation; Weighted Least Square; Feed-Forward Neural Network; Long Short-Term Neural Network

1 Introduction

State Estimation (SE) of power systems or Power System State Estimation (PSSE) is an important tool of the Energy Management System to provide a reliable estimation of the states of the electrical power systems. The phasor voltages (voltage magnitude and angle) for all system buses are estimated based on a set of available measurements. In real-time power systems monitoring, the Supervisory Control and Data Acquisition (SCADA) system is responsible for gathering and preprocessing information such as the values of active and reactive power flows, power injections, bus voltage magnitude, and status of the circuit breaker switches. However, the SCADA system is not capable of performing a convenient treatment of inconsistent information because of gross errors in the measurements, telemetered values, communication noise, etc. [1]. In addition, the collected measurement data may not directly extract the parameters of interest. Besides that telemetering all the data of interest may require large numbers of sensors which are not feasible economically or practically [2]. To handle these issues, PSSE is used to detect and eliminate unreliable data from SCADA to identify the optimal estimate of the operating states consisting of voltage magnitude and angle. SE was first introduced by Gauss and Legendre (around 1800). The main purpose of the SE issue is to identify the voltage magnitude and angle from the various input factors. In engineering modeling, various approaches are proposed for modular robotics and human reasoning [3], cognition processes [4], photovoltaic model [5], biomonitoring studies data [6], tower crane systems modeling [7], and opportunities model [8]. Then, the SE was first applied to electric power systems by Fred Schweppe [9]-[11]. Traditionally, the Weighted Least Square (WLS) method is used to solve the SE. However, several works prove that the conventional WLS method has many disadvantages. The WLS technique is less robust as a single outlier can severely distort the estimation results. Further, the technique has a problem of sluggishness and the likelihood of convergence to local optima [12], [13]. To improve the WLS method, several methods are proposed as Iteratively ReWLS SE through givens rotations [14], robust WLS estimator using Reweighting techniques [15], robust linear-WLS method [16], and linear WLS based Singular Value Decomposition approach [17]. Other methods are also provided to improve the estimate of operating states in comparison with the WLS method, such as Least Absolute Value (LAV) [18], weighted LAV [19], and weighted LAV using Interior Point methods [20]. Recently, the Kalman filter (KF) technique is used in several works to provide efficient results of SE [2]. To increase the efficiency of this method, several improved KF variants are introduced as linear KF [21], the extended KF [22]-[24], the unscented KF [25], [26], the cubature KF [27], the Correntropy KF [28], and the ensemble KF [29]. Other works seek to provide the estimator robustness such as M-estimators [30], the Generalized Maximum-Likelihood (GM) Estimation [31], the H-infinite [32], the Robust Cubature KF [33], and the GM-estimator using Projection statistics [34]-[36]. Another research direction in PSSE is to apply Artificial Intelligence methods, such as the Neural Network (NN) models [37]-[39], Deep Learning [40], [41], Fuzzy Logic [42], [43], and Support Vector Machine [44], [45]. All of these works in the literature show the efficiency of the applied method in a case study. In addition, the optimal estimates of the operating states are normally identified using a single method. Unfortunately, none of the existing works try to combine the advantages of these methods to provide better PSSE solutions.

Therefore, the main contribution of this paper is to combine these single methods using Feed-Forward NN (FFNN) and Long Short-Term Memory NN (LSTM) models to provide better PSSE results. The obtained PSSE results (voltage magnitude and phase angle) using the WLS method and the GM-estimator using Projection statistics are considered from the inputs of FFNN and LSTM models.

These NN models are trained to achieve the optimal solutions. Case studies are conducted to verify the proposed combined approaches in PSSE. To the best of our knowledge, this is the first attempt to provide the combined approaches in this area. The proposed combined approaches for PSSE are represented in Figure 1.

The remainder of this paper is organized as follows. In Section 2, the proposed approaches including the WLS method, the GM-estimator using Projection statistics, the FFNN and LSTM models are presented. In Section 3, the case studies of the IEEE 9-bus system and IEEE 14-bus system are conducted. The conclusions are given in Section 4.

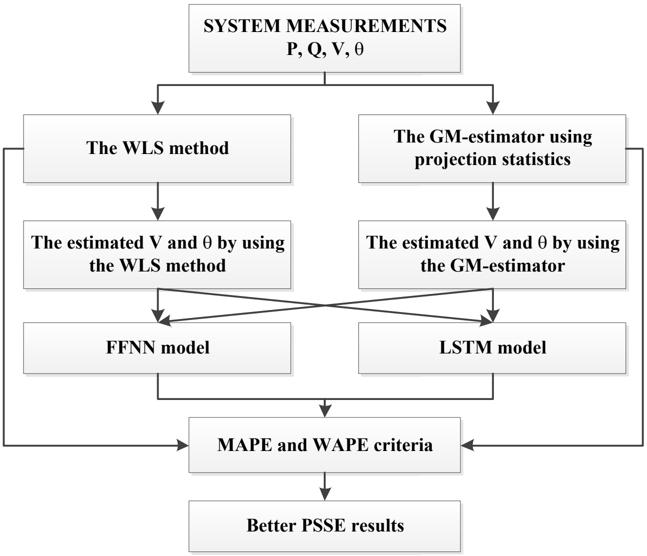

SYSTEM MEASUREMENTS
P, Q, V, θ
The WLS method
The GM-estimator using projection statistics
The estimated V and θ by using the WLS method
The estimated V and θ by using the GM-estimator
FFNN model
LSTM model
MAPE and WAPE criteria
Better PSSE results

Figure 1

Proposed combined approaches for PSSE

2 Proposed Combined Approaches

2.1 WLS Method

The WLS method in [46] can be described as follows:

The measurement model of the SE is represented as a set of non-linear equations $\mathbf{h}$ relating measurements $\mathbf{z}$ to state variables $\mathbf{x}$:

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{e} \tag{1}$$

where vector $\mathbf{x}$ comprises all nodal voltage magnitudes and angles. Vector $\mathbf{z}$ includes active and reactive power injections, active and reactive power flows, and voltage magnitudes. Vector $\mathbf{e}$ is the measurement error that is usually assumed to be independent identically distributed Gaussian with zero mean and diagonal covariance matrix $\mathbf{R} = diag\{\sigma_1^2,...,\sigma_m^2\}$.

The WLS method minimizes the objective function to determine the optimal estimate of x:

$$\mathbf{J}(\mathbf{x}) = \left[\mathbf{z} - \mathbf{h}(\mathbf{x})\right]^T \mathbf{R}^{-1}\left[\mathbf{z} - \mathbf{h}(\mathbf{x})\right]. \tag{2}$$

At the minimum, these can be expressed as follows:

$$\mathbf{g}(\mathbf{x}) = \frac{\partial \mathbf{J}(\mathbf{x})}{\partial \mathbf{x}} = -\mathbf{H}(\mathbf{x})^T \mathbf{R}^{-1}\left[\mathbf{z} - \mathbf{h}(\mathbf{x})\right] = 0 \tag{3}$$

where $\mathbf{H}(\mathbf{x}) = \partial \mathbf{h}(\mathbf{x}) / \partial \mathbf{x}$.

Using the Gauss-Newton method, the estimated values of $\hat{\mathbf{x}}$ can be calculated by the following iterative solution

$$\mathbf{G}(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k) = \mathbf{H}(\mathbf{x}^k)^T \mathbf{W}\left[\mathbf{z} - \mathbf{h}(\mathbf{x}^k)\right] \tag{4}$$

where k is the iteration index. The measurement weight matrix $\mathbf{W} = \mathbf{R}^{-1}$ is the inverse of the measurement covariance matrix.

2.2 GM-estimator using Projection Statistics

This method is represented in [34], [35]. Instead of using Equation (2), the objective function in this method is:

$$\mathbf{J}(\mathbf{x}) = \sum_{i=1}^{m} \omega_i^2 \rho(r_{S_i}) \tag{5}$$

where ω_i is the weight used to bound the influence of the leverage point. The Huber function can be defined as:

$$\rho(r_{S_i}) = \begin{cases} r_{S_i}^2 / 2 & \text{for } \left|r_{S_i}\right| \le \beta \\ \beta\left|r_{S_i}\right| - \beta^2 / 2 & \text{for } \left|r_{S_i}\right| > \beta \end{cases} \tag{6}$$

where $r_{S_i} = r_i / s\omega_i$ is the standardized residual; r_i is the normalized residual; the parameter β is a fixed value; s is the robust scale estimation.

$$s = 1.4826 b_m median_i \left| r_i - median_j (r_j) \right| \tag{7}$$

where b_m is a correction factor for unbiasedness at the Gaussian distribution.

To solve Equation (5), one takes its partial derivative and sets it equal to zero, yielding

$$\frac{\partial \mathbf{J}}{\partial \mathbf{x}} = \sum_{i=1}^{m} -\frac{\omega_i \mathbf{a}_i}{s_i^2} \psi(r_{S_i}) = 0 \tag{8}$$

where $\psi(r_{S_i}) = \partial \rho(r_{S_i}) / \partial r_{S_i}$; $\mathbf{a}_i$ is the *i*th row of the Jacobian matrix $\mathbf{H} = \partial \mathbf{h} / \partial \mathbf{x}$.

This set of equations can be solved by using iterated re-WLS algorithm [47].

2.3 Proposed FFNN-based Approach

In recent decades, NNs have become a hot topic of research; NNs are now widely used in various fields, including speech recognition, multi-objective optimization, function estimation, and classification. NNs can model linear and nonlinear relationships between inputs and outputs without any assumptions based on the activation function's generalization capacity [48]. A NN comprises one input layer, one output layer, and one or more hidden layers. Among the NNs, the FFNN models are the most popular type for function approximation and multi-objective optimization [49]. In this paper, the proposed FFNN-based approach with one hidden layer is illustrated in Figure 2.

The two different FFNN models, i.e., one for voltage magnitude and one for voltage angle, are used to provide the optimal estimate of the power states. The transfer functions for the hidden layer in the FFNN model are Hyperbolic Tangent Sigmoid (i.e., tansig) and Log-Sigmoid (i.e., logsig). The number of neurons in the hidden layer must be identified carefully since the improper number may lead to overfitting or underfitting. A review that discusses how to fix the number of hidden neurons in NNs was presented in [50]. In this paper, the Akaike Information Criterion (AIC) is used to determine the number of hidden neurons. This criterion is defined as

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2k \tag{9}$$

where n is the number of data points (observations), p is the number of estimated parameters, and SSE is the Residual Sum of Squares.

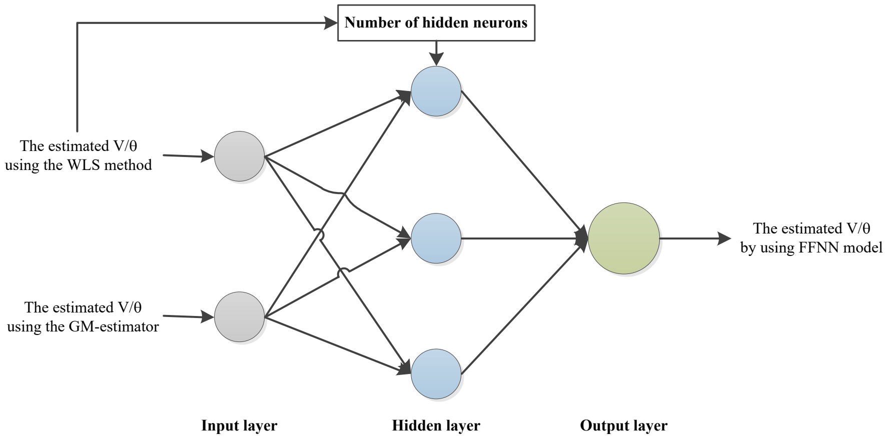

Number of hidden neurons
The estimated V/θ using the WLS method
The estimated V/θ using the GM-estimator
The estimated V/θ by using FFNN model
Input layer
Hidden layer
Output layer

Figure 2
Proposed FFNN-based approach for PSSE

Among several learning algorithms such as Error Correction, Perception Learning, Boltzmann Learning, Hebbian rules, or Back-Propagation (BP), BP is one of the most popular network training algorithms since it is both simple and generally applicable [51]. The difference between actual output and the desired value of the FFNN model is minimized as much as possible by finding the optimal learning rate. The optimization methods for finding the local minimum are Conjugate Gradient such as Gradient Descent with Adaptive Learning Rate (i.e., traingda) and Gradient Descent with Momentum and Adaptive Learning Rate (i.e., traingdx), Steepest Descent such as Resilient BP (i.e., trainrp), and Newton's method such as Levenberg-Marquardt (i.e., trainlm).

Another problem in training NN models is the choice of the number of epochs. The number of epochs determines the number of times that the learning algorithm will work through the entire training dataset. Too many epochs or too few epochs may lead to overfitting or underfitting of the training dataset, respectively. In this case, the early stopping method is often used to solve the generalization issue.

2.4 Proposed LSTM-based Approach

LSTM is a deep learning method proposed in [52]. LSTM can be used as a complex nonlinear unit to construct a larger deep NN, which can reflect the effect of long-term memory and has the ability of deep learning [53]. The LSTM model consists of an input layer, an output layer, and several hidden layers. The basic principle of LSTM is shown in Figure 3.

In Figure 3, x_t t. h_{t-1} and C_{t-1} $t-1$, respectively. f,

i , g , and o are the forgetting gate, input gate, memory cell, and output gate at time t , respectively. C_t and h_t are the updated historical information and the output of the hidden layer at time t , respectively.

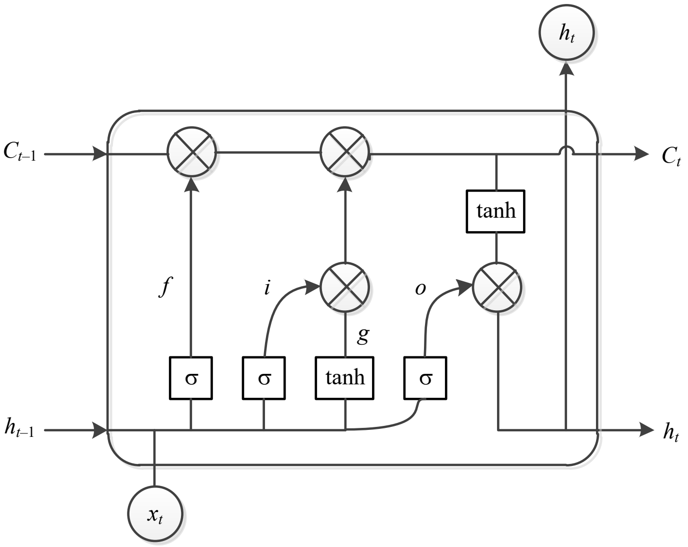


Figure 3
The basic principle of LSTM [52]

The weights and biases to the input gate, forget gate, and output gate control the extent in the cell to compute the output activation of the LSTM block, respectively. Their calculation methods are shown in Equations (10)-(13) as follows:

$$f = \sigma\left(\mathbf{W}_f x_t + \mathbf{U}_f h_{t-1} + \mathbf{b}_f\right), \tag{10}$$

$$i = \sigma\left(\mathbf{W}_i x_t + \mathbf{U}_i h_{t-1} + \mathbf{b}_i\right), \tag{11}$$

$$o = \sigma\left(\mathbf{W}_o x_t + \mathbf{U}_o h_{t-1} + \mathbf{b}_o\right), \tag{12}$$

$$g = \sigma\left(\mathbf{W}_g x_t + \mathbf{U}_g h_{t-1} + \mathbf{b}_g\right) \tag{13}$$

where $\mathbf{W}$, $\mathbf{U}$, $\mathbf{b}$, and σ are the parameter matrix from the input layer to the hidden layer, self-recurrent parameter matrix from the hidden layer to the hidden layer, bias parameter vector, and sigmoid function, respectively.

Then, the internal memory cell state C_t is updated. Finally, the output information of the memory cell h_t is obtained. The calculation methods are shown as follows:

$$C_f = f_t \otimes C_{t-1} \otimes i * e, \tag{14}$$

$$h_t = o \otimes \tanh(C_t) \tag{15}$$

where e is the saved new information.

Similar to the FFNN models, two different LSTM models, i.e., one for voltage magnitude and one for voltage angle, are used to provide the optimal estimate of the power states. The estimated voltage magnitudes and angles using the WLS method and GM-estimator are the inputs of the LSTM models.

2.5 Evaluation Criteria

To evaluate the performance of the estimation methods, several criteria were proposed. In this paper, two primary evaluation criteria are used as:

Mean Absolute Percentage Error (MAPE) is used to measure the percentage error of the estimate in relation to the actual values. It is represented as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A-\hat{A}}{A}\right|100 \tag{16}$$

where n is the number of observations. A and $\hat{A}$ are the actual and estimated values, respectively.

Weighted Average Percentage Error (WAPE) is similar to MAPE. However, it weighs the estimated error.

$$WAPE = \frac{\sum_{i=1}^{n}\left|A-\hat{A}\right|}{\sum_{i=1}^{n}\left|A\right|}100. \tag{17}$$

3 Case Studies

3.1 IEEE 9-bus System

The proposed approaches along with the conventional WLS method and GM-estimator using Projection statistics have been applied on IEEE 9-bus system. The simulation results of voltage magnitudes and angles are tabulated in Tables 1 and 2, respectively. The proposed FFNN and LSTM models are coded and trained in Matlab. Information about the trained FFNN model for estimating the voltage magnitudes including the transfer function, training function, number of hidden neurons, and number of epochs, is tansig, trainlm, 87, and 4, respectively. Information about the trained FFNN model for estimating the voltage angles including the transfer function, training function, number of hidden neurons, and number of epochs, is logsig, trainlm, 30, and 5, respectively.

Information about the trained LSTM model for estimating the voltage magnitudes including the solver, maximum epochs, gradient threshold, initial learn rate, number of hidden neurons, and dropoutlayer is adam optimizer, 10000, 0.01, 0.001, 6, and 0.1, respectively. Information about the trained LSTM model for estimating the voltage angles including the solver, maximum epochs, gradient threshold, initial learn rate, number of hidden neurons, and dropoutlayer is adam optimizer, 10000, 0.01, 0.001, 35, and 0.1, respectively.

Table 1

Comparison of true and estimated voltage magnitudes for IEEE 9-bus system

Bus no	True voltage magnitude (p.u)	Estimated voltage magnitude			
		WLS	GM	FFNN	LSTM
1	1.04000	0.98871	0.99927	1.03863	1.03529
2	1.02500	1.00340	1.01731	1.02508	1.03016
3	1.02500	1.02414	1.03592	1.02500	1.02393
4	1.02579	0.97384	0.98454	1.02579	1.01423
5	0.99563	0.95003	0.96177	0.99563	1.00710
6	1.01265	1.04889	1.06516	1.01265	1.01544
7	1.02577	1.00357	1.01808	1.02577	1.01884
8	1.01588	1.00153	1.01468	1.01588	1.02322
9	1.03235	1.03242	1.04327	1.03235	1.03157

Table 2

Comparison of true and estimated voltage angles for IEEE 9-bus system

Bus no	True voltage angle (degree)	Estimated voltage angle			
		WLS	GM	FFNN	LSTM
1	0.00000	0.00000	0.00000	-0.00752	0.00894
2	9.28001	11.17367	10.97289	9.28386	9.28298
3	4.66475	7.14438	6.99829	4.62074	4.50812
4	-2.21679	-2.45629	-2.40388	-1.72290	-2.26689
5	-3.98881	-3.93075	-3.88208	-3.98727	-4.00079
6	-3.68740	0.11674	0.22492	-3.68426	-3.70708
7	3.71970	5.34103	5.32414	3.72129	3.66028
8	0.72754	2.58574	2.58591	0.74213	0.66242
9	1.96672	4.47395	4.36141	1.96860	1.82292

The true and estimated values of voltage magnitudes and angles in this case study are shown in Figures 4 and 5, respectively. In these figures, the true values, estimated values by using WLS, estimated values by using GM, estimated values by using FFNN, and estimated values by using LSTM are demonstrated by the solid line (black color), dash line with * marker (red color), dash-dot line with square marker (yellow color), dash-dot line with + marker (blue color), and dash-dot line with pentagram marker (green color), respectively.

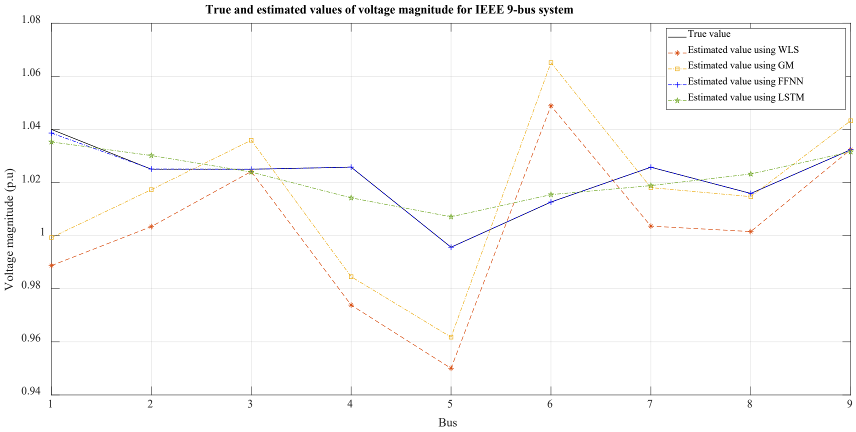


Figure 4
True and estimated values of voltage magnitudes for IEEE 9-bus system

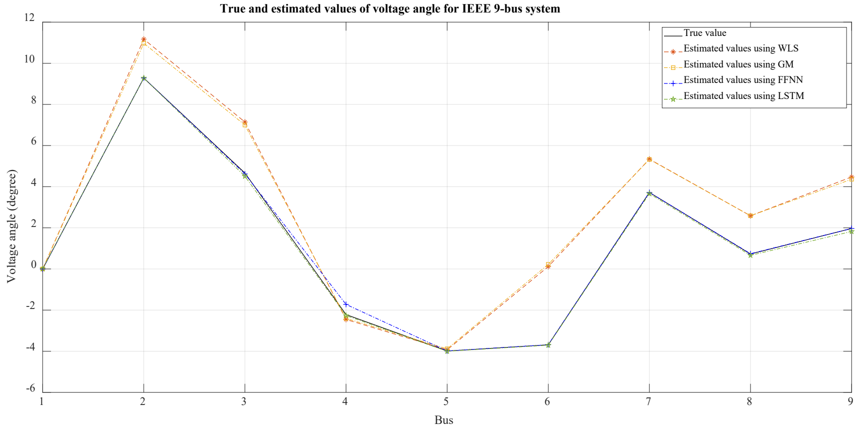


Figure 5
True and estimated values of voltage angles for IEEE 9-bus system

The MAPE and WAPE criteria of the WLS, GM, FFNN, and LSTM methods for estimating voltage magnitudes and angles in this case study are represented in Tables 3 and 4, respectively.

Table 3
Evaluation criteria of the proposed approaches in voltage magnitude estimation for IEEE 9-bus system

Evaluation criteria	WLS	GM	FFNN	LSTM
MAPE (%)	2.65879	2.25172	0.01566	0.56536
WAPE (%)	2.65441	2.24802	0.01592	0.56320

Table 4
Evaluation criteria of the proposed approaches in voltage angle estimation for IEEE 9-bus system

Evaluation criteria	WLS	GM	FFNN	LSTM
MAPE (%)	68.38519	67.31203	2.83688	2.70489
WAPE (%)	47.80473	46.57606	1.89083	1.71443

As shown in Table 3, the GM-estimator (MAPE = 2.25172%, WAPE = 2.24802%) can provide better-estimated results compared to the conventional WLS method (MAPE = 2.65879%, WAPE = 2.65441%). Moreover, the proposed FFNN (MAPE = 0.01566%, WAPE = 0.01592%) and LSTM (MAPE = 0.56536%, WAPE = 0.56320%) approaches can provide more accurate estimated voltage magnitudes compared to the two above methods. The proposed FFNN approach is the best estimation method in this case study.

As shown in Table 4, the GM-estimator (MAPE = 67.31203%, WAPE = 46.57606%) can provide better-estimated results compared to the conventional WLS method (MAPE = 68.38519%, WAPE = 47.80473%). Moreover, the proposed FFNN (MAPE = 2.83688%, WAPE = 1.89083%) and LSTM (MAPE = 2.70489%, WAPE = 1.71443%) approaches can provide more accurate estimated voltage angles compared to the two above methods. The proposed FFNN approach is the best estimation method in this case study.

3.2 IEEE 14-bus System

Similarly, the estimated results of voltage magnitudes and angles in this case study are tabulated in Tables 5 and 6, respectively. Information about the trained FFNN model for estimating the voltage magnitudes including the transfer function, training function, number of hidden neurons, and number of epochs, is logsig, trainlm, 32, and 6, respectively. Information about the trained FFNN model for estimating the voltage angles including the transfer function, training function, number of hidden neurons, and number of epochs, is tansig, trainlm, 34, and 14, respectively. Information about the trained LSTM model for estimating the voltage magnitudes including the solver, maximum epochs, gradient threshold, initial learn rate, number of hidden neurons, and dropoutlayer is adam optimizer, 10000, 0.01, 0.001, 5, and 0.1, respectively. Information about the trained LSTM model for estimating the voltage angles including the solver, maximum epochs, gradient threshold, initial learn rate, number of hidden neurons, and dropoutlayer is adam optimizer, 10000, 0.01, 0.001, 19, and 0.1, respectively.

Table 5

Comparison of true and estimated voltage magnitudes for IEEE 14-bus system

Bus no	True voltage magnitude (p.u)	Estimated voltage magnitude			
		WLS	GM	FFNN	LSTM
1	1.06000	1.04531	1.06873	1.06071	1.06275
2	1.04500	1.02799	1.04954	1.04529	1.03994
3	1.01000	0.98940	1.01203	1.00970	1.01431
4	1.01767	0.99053	1.01540	1.01780	1.00976
5	1.01951	0.99448	1.01947	1.01975	1.02871
6	1.07000	1.02631	1.05667	1.07012	1.05692
7	1.06152	1.01274	1.04099	1.06155	1.07290
8	1.09000	1.04244	1.07030	1.09009	1.07310
9	1.05593	0.99548	1.02508	1.05594	1.06250
10	1.05098	0.99302	1.02325	1.05096	1.05392
11	1.05691	1.00564	1.03635	1.05335	1.05157
12	1.05519	1.00915	1.04072	1.05955	1.05139
13	1.05038	1.00312	1.03457	1.04940	1.04836
14	1.03553	0.97970	1.01116	1.03543	1.04097

Table 6

Comparison of true and estimated voltage angles for IEEE 14-bus system

Bus no	True voltage angle (degree)	Estimated voltage angle			
		WLS	GM	FFNN	LSTM
1	0.00000	0.00000	0.00002	0.00000	-0.06214
2	-4.98259	-5.12756	-4.84083	-4.98258	-5.02083
3	-12.72510	-13.16706	-12.50966	-12.72502	-12.87459
4	-10.31290	-10.51851	-10.01290	-10.31285	-9.99912
5	-8.77385	-8.96567	-8.53965	-8.77382	-8.69088
6	-14.22095	-14.93620	-14.17390	-14.22541	-14.12077
7	-13.35963	-13.73609	-13.04024	-13.35019	-13.07401
8	-13.35963	-13.73477	-13.03726	-13.35929	-13.13985
9	-14.93852	-15.45928	-14.65665	-14.93769	-14.71624
10	-15.09729	-15.67465	-14.86375	-15.09144	-14.73552
11	-14.79062	-15.43579	-14.64498	-14.82734	-14.53140
12	-15.07558	-15.85483	-15.04993	-15.07530	-14.93506
13	-15.15628	-15.91313	-15.09867	-15.17020	-14.98227
14	-16.03364	-16.76849	-15.89944	-16.03184	-15.53161

The true and estimated values of voltage magnitudes and angles in this case study are shown in Figures 6 and 7, respectively. In these figures, the true values, estimated values by using WLS, estimated values by using GM, estimated values by using FFNN, and estimated values by using LSTM are demonstrated by the solid line (black color), dash line with * marker (red color), dash-dot line with square marker (yellow color), dash-dot line with + marker (blue color), and dash-dot line with pentagram marker (green color), respectively.

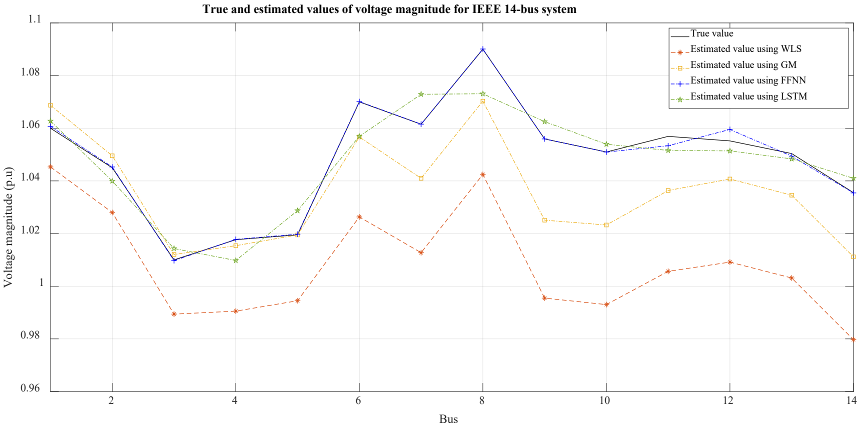


Figure 6

True and estimated values of voltage magnitudes for IEEE 14-bus system

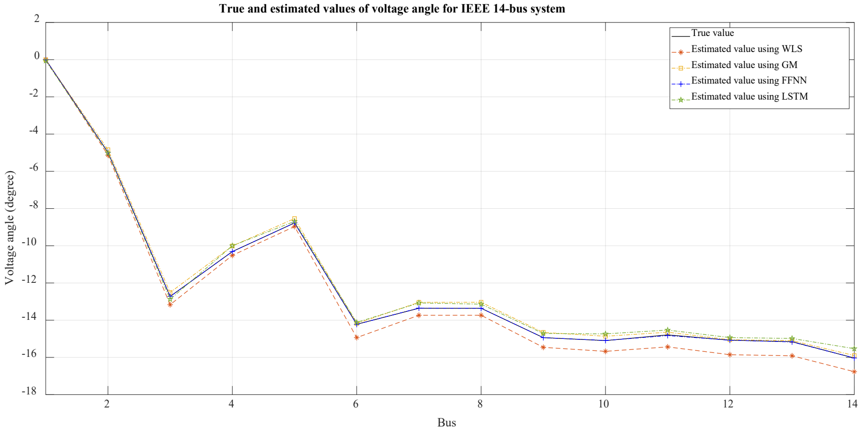


Figure 7

True and estimated values of voltage angles for IEEE 14-bus system

The MAPE and WAPE criteria of the WLS, GM, FFNN, and LSTM approaches for estimating voltage magnitudes and angles in this case study are represented in Tables 7 and 8, respectively.

Table 7

Evaluation criteria of the proposed approaches in voltage magnitude estimation for IEEE 14-bus system

Evaluation criteria	WLS	GM	FFNN	LSTM
MAPE (%)	3.82581	1.38623	0.07419	0.65567
WAPE (%)	3.83760	1.39623	0.07452	0.65874

Table 8

Evaluation criteria of the proposed approaches in voltage angle estimation for IEEE 14-bus system

Evaluation criteria	WLS	GM	FFNN	LSTM
MAPE (%)	3.40259	1.50406	0.03599	1.51901
WAPE (%)	3.82964	1.45638	0.04373	1.72487

As shown in Table 7, the GM-estimator (MAPE = 1.38623%, WAPE = 1.39623%) can provide better-estimated results compared to the conventional WLS method (MAPE = 3.82581%, WAPE = 3.83760%). Moreover, the proposed FFNN (MAPE = 0.07419%, WAPE = 0.07452%) and LSTM (MAPE = 0.65567%, WAPE = 0.65874%) approaches can provide more accurate estimated voltage magnitudes compared to the two above methods. The proposed FFNN approach is the best estimation method in this case study.

As shown in Table 8, the GM-estimator (MAPE = 1.50406%, WAPE = 1.45638%) can provide better-estimated results compared to the conventional WLS method (MAPE = 3.40259%, WAPE = 3.82964%). Moreover, the proposed FFNN (MAPE = 0.03599%, WAPE = 0.04373%) and LSTM (MAPE = 1.51901%, WAPE = 1.72487%) approaches can provide more accurate estimated voltage angles compared to the two above methods. The proposed FFNN approach is the best estimation method in this case study.

Conclusions

The two combined approaches, based on FFNN and LSTM models, for PSSE are proposed. The two proposed approaches are trained to identify the optimal structures. The IEEE 9-bus system and IEEE 14-bus system are used to approve the effectiveness of the proposed approaches in finding the optimal estimate of the states. Two evaluation criteria consisting of MAPE and WAPE are used to specify the better methods. The simulation results indicate that the two proposed combined approaches can provide better solutions compared to the conventional WLS method and the GM-estimator using Projection statistics. For further studies, the weighted Tchebycheff optimization technique and Genetic Algorithm can be applied to solve the PSSE.

References

[1] M. B. D. C. Filho, A. M. L. da Silva, J. M. C. C. Cantera, and R. A. da Silva, “Information debugging for real-time power systems monitoring,” *IEE Proceedings - Generation, Transmission and Distribution*, 136 (3), 1989, pp. 145-152, doi: 10.1049/ip-c.1989.002

[2] A. Saikia and R. K. Mehta, "Power system static state estimation using Kalman filter algorithm", *International Journal for Simulation and Multidisciplinary Design Optimization*, *7*, A7, 2016

[3] C. Pozna and R.-E. Precup, “Plausible Reasoning in Modular Robotics and Human Reasoning,” *Acta Polytechnica Hungarica*, 4 (4), 2007

[4] C. Pozna and R.-E. Precup, “Aspects concerning the observation process modelling in the framework of cognition processes,” *Acta Polytechnica Hungarica*, 9 (1), 2012

[5] N. Ngoc Son and L. The Vinh, “Parameter estimation of photovoltaic model, using balancing composite motion optimization,” *Acta Polytechnica Hungarica*, 19 (11), 2022, pp. 27-46, doi: 10.12700/APH.19.11.2022.11.2

[6] R.-E. Precup, G. Duca, S. Travin, and I. Zinicovscaia, “Processing, neural network-based modeling of biomonitoring studies data and validation on republic of moldova data”, *Proceedings of the Romanian Academy Series A-Mathematics Physics Technical Sciences Information Science*, 2022, 403-410

[7] E.-L. Hedrea, R.-E. Precup, R.-C. Roman, and E. M. Petriu, “Tensor product-based model transformation approach to tower crane systems modeling,” *Asian Journal of Control*, 23 (3), 2021, pp. 1313-1323, doi: 10.1002/asjc.2494

[8] S. Travin and G. Duca, "New opportunities model for monitoring, analyzing and forecasting the official statistics on Coronavirus disease pandemic", *Romanian Journal of Information Science and Technology*, 1, 2023, pp. 49-64, doi: 10.59277/ROMJIST.2023.1.04

[9] F. C. Schweppe and J. Wildes, "Power system static-state estimation, Part I: Exact model", *IEEE Transactions on Power Apparatus and systems*, (1), 1970, pp. 120-125

[10] F. C. Schweppe and D. B. Rom, “Power System Static-State Estimation, Part II: Approximate Model,” *IEEE Transactions on Power Apparatus and systems*, 89 (1), 1970, pp. 125-130, doi: 10.1109/TPAS.1970.292679

[11] F. C. Schweppe, “Power System Static-State Estimation, Part III: Implementation,” *IEEE Transactions on Power Apparatus and systems*, 89 (1), 1970, pp. 130-135, doi: 10.1109/TPAS.1970.292680

[12] E. Caro and A. J. Conejo, “State estimation via mathematical programming: a comparison of different estimation algorithms,” *IET Generation,*

Transmission & Distribution, 6 (6), 2012, pp. 545-553, doi: 10.1049/iet-gtd.2011.0663

[13] E. J. Contreras-Hernandez and J. R. Cedeno-Maldonado, “A self-adaptive evolutionary programming approach for power system state estimation,” in *2006 49th IEEE International Midwest Symposium on Circuits and Systems*, Aug. 2006, pp. 571-575. doi: 10.1109/MWSCAS.2006.382127

[14] R. C. Pires, A. Simoes Costa, and L. Mili, “Iteratively reweighted least-squares state estimation through givens rotations,” *IEEE Transactions on Power Systems*, 14 (4), 1999, pp. 1499-1507, doi: 10.1109/59.801941

[15] E. Caro, R. Mínguez, and A. J. Conejo, “Robust WLS estimator using reweighting techniques for electric energy systems,” *Electric power systems research*, 104, 2013, pp. 9-17, doi: 10.1016/j.epsr.2013.05.021

[16] S. K. Kotha, B. Rajpathak, M. Mallareddy, and R. Bhuvanagiri, “Wide area measurement systems based power system state estimation using a robust linear-weighted least square method,” *Energy Reports*, 9, 2023, pp. 23-32, doi: 10.1016/j.egyr.2023.05.046

[17] G. Bei, “Observability analysis for state estimation using Hachtel’s augmented matrix method,” *Electric power systems research*, 77 (7), 2007, pp. 865-875, doi: 10.1016/j.epsr.2006.07.010

[18] A. Abur and M. K. Celik, “Least absolute value state estimation with equality and inequality constraints,” *IEEE Transactions on Power Systems*, 8 (2), 1993, pp. 680-686, doi: 10.1109/59.260812

[19] C. Rakpenthai, S. Uatrongjit, I. Ngamroo, and N. R. Watson, “Weighted least absolute value power system state estimation using rectangular coordinates and equivalent measurement functions,” *IEEJ transactions on electrical and electronic engineering*, 6 (6), 2011, pp. 534-539, doi: 10.1002/tee.20692

[20] H. Singh and F. L. Alvarado, “Weighted least absolute value state estimation using interior point methods,” *IEEE Transactions on Power Systems*, 9 (3), 1994, pp. 1478-1484, doi: 10.1109/59.336114

[21] S. Sarri, L. Zanni, M. Popovic, J.-Y. Le Boudec, and M. Paolone, “Performance assessment of linear state estimators using synchrophasor measurements,” *IEEE Transactions on Instrumentation and Measurement*, 65 (3), 2016, pp. 535-548, doi: 10.1109/TIM.2015.2510598

[22] L. Fan and Y. Wehbe, “Extended Kalman filtering based real-time dynamic state and parameter estimation using PMU data,” *Electric Power Systems Research*, 103, 2013, pp. 168-177, doi: 10.1016/j.epsr.2013.05.016

[23] F. Shabani, M. Seyedyazdi, M. Vaziri, M. Zarghami, and S. Vadhva, “State estimation of a distribution system using WLS and EKF Techniques,” in

2015 IEEE International Conference on Information Reuse and Integration, Aug. 2015, pp. 609-613, doi: 10.1109/IRI.2015.101

[24] E. Ghahremani and I. Kamwa, “Dynamic state estimation in power system by applying the extended Kalman filter with unknown inputs to phasor measurements,” *IEEE Transactions on Power Systems*, 26 (4), 2011, pp. 2556-2566, doi: 10.1109/TPWRS.2011.2145396

[25] G. Valverde and V. Terzija, “Unscented Kalman filter for power system dynamic state estimation,” *IET generation, transmission & distribution*, 5 (1), 2011, pp. 29-37, doi: 10.1049/iet-gtd.2010.0210

[26] X. Qing, H. R. Karimi, Y. Niu, and X. Wang, “Decentralized unscented Kalman filter based on a consensus algorithm for multi-area dynamic state estimation in power systems,” *International Journal of Electrical Power & Energy Systems*, 65, 2015, pp. 26-33, doi: 10.1016/j.ijepes.2014.09.024

[27] A. Sharma, S. C. Srivastava, and S. Chakrabarti, “A cubature Kalman filter based power system dynamic state estimator,” *IEEE Transactions on Instrumentation and Measurement*, 66 (8), 2017, pp. 2036-2045, doi: 10.1109/TIM.2017.2677698

[28] J. A. D. Massignan, J. B. A. London, and V. Miranda, “Tracking power system state evolution with maximum-correntropy-based extended Kalman filter,” *Journal of Modern Power Systems and Clean Energy*, 8 (4), 2020, pp. 616-626, doi: 10.35833/MPCE.2020.000122

[29] N. Zhou, D. Meng, Z. Huang, and G. Welch, "Dynamic state estimation of a synchronous machine using PMU data: A comparative study", *IEEE Transactions on Smart grid*, 6 (1), 2014, pp. 450-460

[30] G. Durgaprasad and S. S. Thakur, “Robust dynamic state estimation of power systems based on M-estimation and realistic modeling of system dynamics,” *IEEE Transactions on Power Systems*, 13 (4), 1998, pp. 1331-1336, doi: 10.1109/59.736273

[31] J. Zhao and L. Mili, "A framework for robust hybrid state estimation with unknown measurement noise statistics", *IEEE Transactions on Industrial Informatics*, 14 (5), 2017, pp. 1866-1875

[32] J. Zhao, “Dynamic state estimation with model uncertainties using H_ınfinity extended Kalman filter,” *IEEE Transactions on power systems*, 33 (1), 2018, pp. 1099-1100, doi: 10.1109/TPWRS.2017.2688131

[33] Y. Wang, Y. Sun, V. Dinavahi, S. Cao, and D. Hou, "Adaptive robust cubature Kalman filter for power system dynamic state estimation against outliers", *IEEE Access*, 7, 2019, pp. 105872-105881

[34] L. Mili, M. G. Cheniae, N. S. Vichare, and P. J. Rousseeuw, “Robust state estimation based on projection statistics [of power systems],” *IEEE*

Transactions on Power Systems, 11 (2), 1996, pp. 1118-1127, doi: 10.1109/59.496203

[35] J. Zhao, G. Zhang, and M. La Scala, “A two-stage robust power system state estimation method with unknown measurement noise,” in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, Jul. 2016, pp. 1-5, doi: 10.1109/PESGM.2016.7741350

[36] Y. Shi, Y. Hou, Y. Yu, Z. Jin, and M. A. Mohamed, "Robust power system state estimation method based on generalized M-estimation of optimized parameters based on Sampling", *Sustainability*, 15 (3), 2023, pp. 2550

[37] D. M. V. Kumar, S. C. Srivastava, S. Shah, and S. Mathur, “Topology processing and static state estimation using artificial neural networks,” *IEE Proceedings-Generation, Transmission and Distribution*, 143 (1), 1996, pp. 99-105

[38] E. Manitsas, R. Singh, B. C. Pal, and G. Strbac, “Distribution system state estimation using an artificial neural network approach for pseudo measurement modeling,” *IEEE Transactions on power systems*, 27 (4), 2012, pp. 1888-1896, doi: 10.1109/TPWRS.2012.2187804

[39] H. Salehfar and R. Zhao, “A neural network preestimation filter for bad-data detection and identification in power system state estimation,” *Electric power systems research*, 34 (2), 1995, pp. 127-134, doi: 10.1016/0378-7796(95)00966-7

[40] L. Wang, Q. Zhou, and S. Jin, “Physics-guided deep learning for power system state estimation,” *Journal of Modern Power Systems and Clean Energy*, 8 (4), 2020, pp. 607-615, doi: 10.35833/MPCE.2019.000565

[41] K. R. Mestav, J. Luengo-Rozas, and L. Tong, “Bayesian state estimation for unobservable distribution systems via deep learning,” *IEEE Transactions on Power Systems*, 34 (6), 2019, pp. 4910-4920, doi: 10.1109/TPWRS.2019.2919157

[42] F. Shabani, N. R. Prasad, and H. A. Smolleck, “A fuzzy-logic-supported weighted least squares state estimation,” *Electric power systems research*, 39 (1), 1996, pp. 55-60, doi: 10.1016/S0378-7796(96)01107-8

[43] K. E. Holbert and K. Lin, “Reducing state estimation uncertainty through fuzzy logic evaluation of power system measurements,” in *2004 International Conference on Probabilistic Methods Applied to Power Systems*, Sep. 2004, pp. 205-211

[44] V. Kirinčić, E. Čeperić, S. Vlahinić, and J. Lerga, "Support vector machine state estimation", *Applied Artificial Intelligence*, 33 (6), 2019, pp. 517-530

[45] A. A. Abod, A. H. Abdullah, and M. K. Abd, “Support vector machine based approach for state estimation of Iraqi super grid network,” in *2008*

Workshop on Power Electronics and Intelligent Transportation System, Aug. 2008, pp. 252-256, doi: 10.1109/PEITS.2008.102

[46] A. Abur and A. G. Exposito, "*Power system state estimation: theory and implementation*", 2004, CRC press

[47] W. W. Kotiuga and M. Vidyasagar, “Bad data rejection properties of weighted least absolute value techniques applied to static state estimation,” *IEEE Transactions on Power apparatus and systems*, 101 (4), 1982, pp. 844-853, doi: 10.1109/TPAS.1982.317150

[48] T. H. Le, L. Dai, H. Jang, and S. Shin, "Robust process parameter design methodology: A new estimation approach by using feed-forward neural network structures and machine learning algorithms", *Applied Sciences*, 12 (6), (2022), pp. 2904

[49] Z. Zainuddin and O. Pauline, "Function approximation using artificial neural networks", *WSEAS Transactions on Mathematics*, 7 (6), 2008, pp. 333-338, doi:10.5555/1466915.1466916

[50] K. G. Sheela and S. N. Deepa, "Review on methods to fix number of hidden neurons in neural networks", *Mathematical problems in engineering*, 2013, *2013*

[51] A. Zilouchian and M. Jamshidi (Eds.). "*Intelligent control systems using soft computing methodologies*, CRC press, 2001

[52] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, 9 (8), 1997, pp. 1735-1780, doi: 10.1162/neco.1997.9.8.1735

[53] Q. Xiaoyun, K. Xiaoning, Z. Chao, J. Shuai, and M. Xiuda, “Short-term prediction of wind power based on deep long short-term memory,” in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)* Oct. 2016, pp. 1148-1152, doi: 10.1109/APPEEC.2016.7779672

Compliance Risk Assessment – Results of a Comprehensive Literature Review

Petra Benedek, Ferenc Bognár

Department of Management and Business Economics, Faculty of Economic and Social Sciences, Budapest University of Technology and Economics
Műegyetem rkp 3, H-1111 Budapest, Hungary
benedek.petra@gtk.bme.hu, bognar.ferenc@gtk.bme.hu

Abstract: Today's terminology and definitions of compliance risk are various, and the description of compliance risk assessment is heterogeneous in the literature. These differences result in different expectations, processes, and methodologies in practice, which do not support the widespread adoption of standardized compliance management systems. This study is based on a comprehensive literature review. It aims to redefine compliance risk and propose a structured model for the compliance risk assessment process. The study provides a new framework for compliance risk assessment based on findings and gaps in scientific papers, business reports, and relevant standards. It also introduces the Digital Operational Resilience Act and its compliance aspects.

Keywords: compliance risk; risk assessment; risk identification; DORA; PRISM

1 Introduction

Organizations increasingly realize that they must address the issue of compliance in their operations. New rules and regulations go beyond national borders while increasing in quantity and extent. Regardless of size, sector, and other parameters, organizations are affected by a complex, ever-changing regulatory environment and are subject to enforcement actions, sanctions, fines, and reputational risk.

This paper focuses on the various interpretations of compliance risk and the compliance risk assessment process. In this study, the definitions of compliance risk are collected from the literature to answer the following research questions:

RQ1. What meanings does the term compliance risk contain?

RQ2. What does the compliance risk assessment process look like according to the literature?

RQ3. What are the gaps in the current literature that future research might explore?

In this section, a brief introduction to compliance risks is presented. A frequently referred definition of compliance risk states that it is the organization's exposure to potential legal or regulatory sanctions, financial loss, or a loss of reputation due to the organization's failure to comply with laws and regulations [1]. Compliance risk also includes failure to comply with internal policies or best practices on various topics, like data protection, which could lead to the inability to operate the business.

The term “compliance function” refers to the workgroup which carries out the compliance activities [1]. While the independence of the compliance function is necessary to avoid conflict of interest between compliance and other units, close cooperation with other internal control functions and the business units is indispensable [1].

The ISO 19600:2014 “Compliance management systems guidelines” recognize a risk-based approach to compliance [2]. The ISO 19600:2014 guidelines for compliance management are aligned with the ISO 31000:2018 risk management guidelines [3] as described in previous works [4,5]. The ISO 37301:2021 Compliance management systems standard [6] supersedes the ISO 19600:2014. The standard follows the PDCA logic, where risk identification is part of the Plan phase, compliance risk mitigation by controls and procedures is part of the Do phase, and measurement and monitoring activities are included in the Check phase.

Organizations may follow frameworks and mechanisms to control compliance risk. One critical activity of compliance management is monitoring changes in the regulatory environment to ensure that the organization is well informed and up-to-date on the requirements it is facing and in understanding its level of compliance. Business continuity is closely related to compliance management. Organizations that are prepared and able to remain operational even during disruptive events (e.g., cyberattacks) instill confidence in their partners and can expect better cooperation.

Ultimately, the board (the governing body) is responsible for reviewing all aspects of an organization's compliance risk, and senior management is responsible for effectively communicating and managing the risks [1]. Compliance risk consists primarily of penalties and other consequences for regulatory noncompliance and reputational risk. The first includes illegal practices, like fraud, theft, bribery, money laundering, and embezzlement. Violation of data protection laws, pollution, environmental damage, and occupational health and safety violations are also common compliance risks. Cloud computing delivers new compliance risks since cloud services might store sensitive or protected data.

It is necessary, to clarify a few other risks that are close or even partially overlap with compliance risks.

1) Reputational risk is a loss in an organization's perceived trustworthiness or integrity. It has a negative impact, resulting in direct losses in revenue, indirect losses of customers, orders, employees, foregone business opportunities, or perception of the brands. Reputational loss is usually a

consequence of another business risk; negative news spreads quickly and beyond the company's control.

2) Integrity risks are current or future threats to an organization's reputation, capital, or results due to inadequate compliance with applicable laws. Integrity risks are partly the risk of insufficient compliance with the law and, on the other hand, the risk of employees engaging in actions that could seriously damage trust in the organization. Examples of integrity risks are money laundering, corruption, and conflicts of interest between staff and clients.

3) Conduct risk refers to the potential inappropriate, unethical, or harmful behavior (such as misleading advertising, insider trading, market manipulation) that could negatively affect customers, investors, and the market. Conduct risk can have serious consequences, damage the institution's reputation, lead to legal and regulatory sanctions, and cause financial losses to customers or investors. Nicolas and May defined conduct risk as any activity or inaction by an organization's personnel that could lead to unfair outcomes for its clients, affect the integrity of the markets, or otherwise jeopardize the organization's reputation or financial situation [7].

The Digital Operational Resilience Act (DORA) is a legislative proposal of the European Commission which aims to increase the operational resilience of the EU financial sector by creating a harmonized framework for digital operational resilience. The proposal aims to ensure financial institutions can withstand and respond to various operational risks, including cyberthreats, IT disruptions, and other technology-related risks. According to present plans, it shall apply from January 2025 [8]. DORA compliance is a current challenge for thousands of financial entities and ICT service providers operating within the EU and the ICT infrastructure supporting them from outside the EU.

The importance of DORA lies in the financial sector increasingly relying on digital technology, which presents new risks and challenges regarding operational flexibility [9]. Cyberattacks, IT failures, and other technology-related incidents can cause significant disruption to financial institutions and have far-reaching consequences for the financial system and the economy. DORA is expected to significantly impact financial institutions operating in the EU, as they must meet new requirements and standards for digital operational flexibility. This includes establishing and maintaining effective governance and risk management arrangements, conducting regular testing and exercises to assess operational resilience, and reporting significant events to the relevant authorities.

The financial sector witnesses a change in the regulatory perspective from defense and protection to building resistance, resilience, and flexibility [10]. Therefore, an Information and Communication Technology (ICT) risk management framework to manage ICT risks is strongly connected to compliance risk management since some risks may have regulatory, reputational, or both effects.

This paper is organized as follows. Section 2 introduces the methodology, while Section 3 presents the results. In Section 4, the results are discussed, highlighting managerial implications.

2 Methodology

This research is based on a comprehensive literature review. The data extraction process was designed based on the research questions to highlight the similarities and differences among the results of the studies.

For this study, the authors used the Scopus digital database, which many research studies have used, to select and identify the most relevant studies. The selection process was guided by specific keywords included in the following search: compliance risk OR compliance assessment OR compliance risk evaluation. The search was conducted in July 2023 following the logic of the PRISMA 2020 statement.

The search was extended to one regulatory documents outside the Scopus [1], that was used as references in the first set of research studies. Additionally, reports and white papers published by consultancy firms are reviewed in Section 3.3.

The following inclusion criteria have been defined for examining the research questions: (i) journal papers and regulatory reports that dealt with the intersection of compliance management and risk management and included the terms in the title, abstract, or keywords; (ii) documents in English; (iii) documents published since 2005. In addition, papers using the term outside of an organizational perspective (e.g., medical use) were excluded from the research. The selected documents are presented in Table 1.

Table 1

Documents of the literature review

Bibliographic information of the publication	**Country of research**	**Approach/methodology**
Basel Committee on Banking Supervision, 2005 [1]	Switzerland	high-level paper on compliance risk and the compliance function in banks
Birindelli, Ferretti, 2008 [12]	Italy	questionnaire
Sathye, Islam, 2011 [13]	Australia	method of analogy, scorecard of risk assessment based on the literature on credit-scoring models
Birindelli, Ferretti, 2013 [14]	Italy	literature review, theoretical model of an efficient internal control system
Esayas, Mahler, 2015 [15]	Norway	modeling of compliance risk identification and assessment

Losiewicz-Dniestrzanska, 2015 [11]	Poland	literature review and proposal of quantitative indicators in compliance risk monitoring
Nicolas, May, 2017 [7]	USA	practical guidance for developing a compliance risk assessment
Naheem, 2019 [16]	Germany	literature review and surveys
Achkasova et al. 2021 [17]	Ukraine	cognitive modeling method based on the construction of a fuzzy cognitive map

3 Results

The results are presented along with the research questions. Section 3.1 reflects on the definitions of compliance risks and the boundaries of compliance risk management. Section 3.2 provides a detailed insight into the risk assessment process. Finally, Section 3.3 delivers additional information from the business reports and surveys.

3.1 Definitions and Insights on Compliance Risk

The following explicit definition of compliance risk has been collected:

1) The definition of compliance risk is given by the [1] as follows: “The expression "compliance risk" is defined in this paper as the risk of legal or regulatory sanctions, material financial loss, or loss to reputation a bank may suffer as a result of its failure to comply with laws, regulations, rules, related self-regulatory organisation standards, and codes of conduct applicable to its banking activities (together, "compliance laws, rules, and standards").” This definition is widely accepted and used [11, 14].
2) The Bank of Italy provides another definition. "The risk of non-compliance with rules is the risk to incur in judicial or administrative sanctions, material financial losses or loss of reputation as a result of infringement of mandatory rules (laws and regulations) or of self-regulation (that is statutes, codes of conduct, codes of self-discipline)" [12].
3) The Polish Financial Supervisory Commission defined noncompliance risk "as a result of a bank's failure to comply with legal requirements and recommendations set out by the Polish Bank Association" [11].
4) Nicolas and May [7] define compliance risk as the risk of legal or regulatory sanctions or financial loss resulting from failure to comply with applicable laws, regulations, rules, and related market standards.
5) The ISO 37301:2021 standards define compliance risk as likelihood of occurrence and the consequences of not fulfilling the organization’s (mandatory or voluntarily chosen) compliance obligations [6].

While legal risks have an external focus, compliance risks focus on the internal and external environment and include failures to comply with self-regulatory standards [15]. Furthermore, reputational risks that are excluded from legal and operational risks are also included in compliance risks. There is a partial overlapping of operational, legal, and compliance risks [14].

Requirements for the efficient and effective management of compliance risks include (1) establishing an independent function and (2) the definition of the person responsible for compliance risk management [12]. The independence of the compliance function, its formalization of responsibilities, and relationship with other control functions are general requirements [14].

In principle, the compliance function and the internal audit function should be separated to ensure that the activities of the compliance function are subject to an independent review [1]. While compliance risk assessment is primarily the responsibility of the compliance functions, a review (control) responsibility lies within the internal audit, and supervision is the governing body's responsibility [1]. In contrast, [7] emphasized that the business should own the compliance risk assessment process, and the compliance function should only assist in its planning and execution.

With internal audit, some synergies are related to risk and control assessment methods, risk mapping, and promoting a strong "control culture" [12]. Unlike the top-down approach of internal audit, compliance management is a bottom-up activity with an analytical vision of compliance risks and the processes involved [14].

Operational risks, legal risks, and compliance risks often overlap. Cross-cases emerge from a "grey zone" that includes breach of contract (classified as an operational risk event) and the bank's liability for improper conduct leading to legal risk lawsuits [14]. Although the European Network and Information Security Agency [18] have published recommendations on cloud computing risk assessment, there are no specific guidelines for identifying legal risks.

Both compliance and operational risk management are second-level control structures whose task is to identify the risks inherent in the processes implemented by the various functions [14]. A cooperative or integrated approach to risks can create effective synergies facilitated by shared risk identification, risk indicators, business environment analysis, and information exchange and validation [14]. Consultations with business units (e.g., internal audit, operational risk unit, legal or security department) and using the results of their audits and information from their reports can contribute to better compliance monitoring [11].

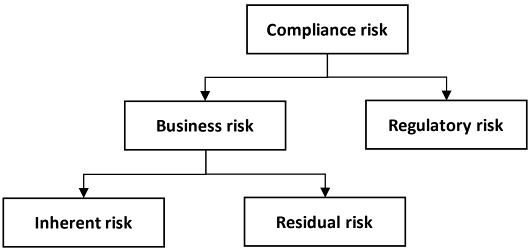


Figure 1
The grouping of compliance risks, own editing based on [13 p. 176.]

Compliance risks can be grouped under two categories: regulatory risk and business risk (Figure 1). Regulatory risks come into being because of the inability to comply with legislation requirements. Australia's financial intelligence unit divides business risk into inherent and residual risk. Inherent risks are identified and managed and come from various sources like customers (i.e., politically exposed people, customer complaints), products or services, and previous compliance reports [7, 13]. Inherent risks are identified, controls to mitigate the risks are listed, and the resulting residual risk calculations are classified in terms of potential financial, regulatory, and public reputational damage to the entity [7]. Methods for identifying inherent risk should include quantifying control effectiveness [7]. The basic idea behind quantifying inherent business risks is to pre-identify key factors and combine or weight them into a quantitative score, which can be directly interpreted as the probability or used as a classification system. Residual risk is the risk left, despite a robust risk management system [13].

3.2 The Compliance Risk Assessment Process

Compliance risk management is a systematic process for identifying, analyzing, and prioritizing an organization's compliance risks. According to Nicolas and May [7], the compliance risk assessment process starts with identifying the main inherent risks within a business or legal entity. In this section, the next steps are risk analysis and evaluation, followed by risk treatment.

3.2.1 Risk Identification

Risk identification examines how a compliance requirement—obligation or prohibition—may lead to risk. Risk identification can be requirements-centered or facts-centered [15], and both approaches are equally relevant. In the requirements-centered approach, experts aim to identify what might trigger the legal norm through guiding questions like what actions could lead to violations.

In contrast, business processes are evaluated in the facts-centered approach to identify potential noncompliance areas. The benefit of a facts-centered approach is that it is possible to reuse previously identified risks from other areas and assess their compliance implications [15].

Nicolas and May [7] recommend combining the above approaches as follows:

1) The regulatory requirements are the starting point.
2) The risk inventory is prepared based on them.
3) The next step is the detailed examination of the risks through the relevant business processes and identifying (yet not assessing) relevant actual controls.

Risk identification builds on the collection of timely and accurate data (even independent third-party data) and on identifying the relevant legal obligations by establishing an applicable legal framework and evaluating the relevance and importance of specific regulations in the organization's business activities. Risk identification is a critical step for the effectiveness of the subsequent stages of the risk management process [11].

According to Łosiewicz-Dniestrzańska [11], risks can be described by four factors: nature (event or incident), source (people or units, like internal audit or operational risk reports, whistleblowing), cause and effect (impact). Measuring risk compliance in banks usually means creating overly simplistic risk matrices determining the risk degree [11, 16].

Esayas and Mahler [15] found that compliance risk identification is usually made in unstructured or semi-structured brainstorming sessions, relying on lawyers' expertise. Instead, they propose a requirements-centered five-step process for the structured identification and assessment of legal and compliance risks:

1) step: identify the source of the requirements
2) step: list of obligations and prohibitions
3) step: structuring a requirements template
4) step: template-based modeling
5) step: instantiation

"We don't talk the same language when we discuss risks" [15]. Esayas and Mahler highlight the importance of language and possible difficulties in communication as experts from different fields use their vocabularies (e.g., IT, legal). The proposed graphical modeling can break down complex regulations into easily understandable elements. Using templates in the risk identification and assessment steps facilitates modeling and monitoring, while it has the risk of missing information while transforming the regulations. One participant in their case study indicated that risk identification is less challenging than risk assessment.

The benefits of a structured approach to risk identification reduce the subjectivity of compliance decisions. In addition, visualization provides focus and facilitates communication between experts from different backgrounds. Furthermore, the structured approach produces reusable results, so the costs of using the approach can be lower in the long-term [15].

3.2.2 Risk Analysis and Evaluation

Compliance risk assessment is a systematic process for identifying, analyzing, and prioritizing an organization's compliance risks. Compliance risk assessment aims to identify areas of significant risk and where controls are required to reduce risks [7]. The Basel Committee [1] proposes using performance indicators to measure compliance risks. According to the document, compliance risk should be incorporated into the internal audit function's risk assessment methodology, and an audit program should be established that covers the testing of controls proportional to the level of perceived risk. Birindelli [12] suggests defining risk models and Key Performance Indicators as part of the boundary setting of the different management areas. Meanwhile, some risks are simultaneously part of operational and compliance risks (i.e. contract breeches).

By the Second Pillar of Basel 2, it is necessary to quantitatively measure compliance risk in banks [12]. In the early stages of risk assessment, there was no general, predefined methodology for assessing compliance risk, and banks used non-statistical methods to calculate risk exposure, such as [12]:

1) Qualitative assessments based on indicators,

2) Self-assessment of the frequency and severity of the risk and the controls. The aim was to calculate the residual risk present after the controls.

Sathye and Islam [13] distinguish the rule-based and risk-based approaches to compliance. The former means establishing the compliance function based on a catalog of regulations. After collecting the legal requirements, they must be evaluated to implement appropriate measures to ensure compliance. The latter approach means that organizations (reporting entities) can develop compliance procedures and processes and allocate resources appropriately to address the specific risks they face [13, 15]. The benefits of the risk-based approach are the efficient allocation of resources, prioritization of risks, and lesser burdens on customers (and eventually lesser costs). The main steps of the compliance risk management process are risk identification, risk assessment, and developing strategies to manage and mitigate the identified risks [13].

Sathye and Islam [13] propose an inherent business risk assessment scorecard based on credit scoring models. Two risk assessment factors are the risk's likelihood (probability) and impact (severity). For example, they propose a 400-point model that consists of 4 main types of risk for money laundering and terrorism financing, where customers over 300 points would be considered high risk. In general, the outcome of the assessment is, on the one hand, the level of risk identified (high, medium, and low) and, on the other hand, mitigation and control procedures relevant to the risk.

For regulatory risk assessment, Sathye and Islam [13] propose a qualitative self-assessment technique, a questionnaire as a checklist to assess compliance with relevant regulations.

Łosiewicz-Dniestrzańska [11] proposes independently determining risk likelihood and impact on 1-to-5 scales and computing the overall risk as a product of impact x likelihood. Next, we can transform numerical values (1-25) to a 5-scale risk rating (minor, moderate, significant, major, catastrophic). In practice, the accepted scale is often narrower and consists of only three categories (green, amber, red), where, like on a heat map, the amber is a warning and requires corrective measures [11].

Risk assessment is generally carried out in teams, which can be facilitated with software inputs [16]. Teams might include members out of the organization, like customs experts or other third parties. Esayas and Mahler [15] highlight that the risk appetite of the individuals performing the risk assessment might differ significantly. Hence, the evaluations are subjective in case of no formalized approach to compliance risk assessment. Historical data can help simplify the estimation of the probability and impact of compliance risks. In their study, violations have a low, medium, or high-level impact on compliance, depending on the level of remediation (individual, business unit or board, respectively) [15].

According to the 2008 Federal Reserve Supervisory Letter [19], the risk assessment should be based on company-wide standards that define the method and criteria for risk assessment throughout the organization. Also, it should consider the risk inherent in the activity and the strength and effectiveness of the controls designed to mitigate the risk [19]. For assessing risk controls, some questions focus on control design, others on implementation (How reliable is the control? Is it easily bypassed? With control operation: how well does control work in practice?) [7].

Naheem [16] distinguished reactionary versus forward-thinking strategies for anti-money laundering (AML) risk assessment. Reactionary focus means following the the state's agenda and managing development according to regulatory requests. It has the disadvantages of not recognizing risks and other legal challenges, like too fast changes in regulation.

Naheem [16] highlights that improved technology facilitates detecting wrongdoing. Also, this study identified three areas for improvement in detecting and calculating risks: training and experience of the team members and communication with management.

3.2.3 Risk Treatment

The compliance risk assessment forms the basis for implementing compliance management systems and allocating appropriate resources and processes to manage the identified compliance risks. Improvements based on compliance risk assessment lead to better compliance with health and safety and other specific regulations. A compliance risk assessment is a real opportunity to initiate new and update old or unused controls to mitigate risk [7]. Establishing and implementing controls aims to reduce the probability of the causes and their negative consequences. The following control mechanisms can be helpful to internal procedures: training, segregation of duties, application of the "four eyes" principle, legal opinions, physical security, and system mechanisms (access rights, exclusions), surveillance and monitoring, and testing [7, 11].

Quantitative tools for compliance risk monitoring are mainly based on simple, readily available indicators, often overlapping with those used by operational risk management. They are based on historical data (e.g., the number of overdue corrective action, the number of customer complaints to regulators, ratio of completion of training, and the number and frequency of detected violations) [11, 20, 21]. Please note that indicators do not measure the risk but are valuable in showing the trends and can signal early warnings.

Given the importance of issues related to compliance risk assessment, it is necessary to develop a theoretical basis and tools to assess the potential growth and realization of compliance risks [17].

3.3 The Business Perspective

Traditionally, compliance has been seen as the responsibility of specific business units or functions (i.e., financial regulation, safety and environmental laws, employment standards). Many businesses used a silo approach to compliance and isolated efforts without aligned intent [22].

According to a KPMG survey in 2006, compliance verifies the consistency of internal and bank regulations and advises on legal risk issues, while the risk management function monitors all risks [14]. KPMG emphasizes the importance of compliance risk assessment in developing effective compliance programs. The report highlights key steps in a compliance risk assessment and provides practical advice for organizations to conduct compliance risk assessments [23]. Advancements in technology and automation present tremendous opportunities to innovate and increase efficiency, as data analytics solutions help to identify alerting data, prevent, detect, and respond to potential violations and make evidence-based decisions. Key risk and performance indicators often predict events that can increase an organization's risk exposure and work as alerting signs of potential problems so they can be monitored and mitigated. KPIs and KRIs enable compliance managers to make better decisions and manage compliance risks more effectively. KPMG also presents a maturity model for the integration of data analytics into compliance management [24].

PwC provides practical guidance on compliance risk assessments, including using risk matrices [25]. Compliance testing should be designed around and focused on the organization's most serious threats and aligned with risk appetite and business risk assessment. Mitigation should respond to test results; the most significant identified risks or weaknesses are subject to increased testing. The compliance function typically performs this type of assessment with data from business areas [26].

Boards must provide tangible evidence that they are effectively managing their compliance risks [27]. According to a recent Ernst & Young report [28], emerging technology could improve the early detection of risks (e.g., using AI in fraud detection, continuous monitoring instead of sampling), contribute to less reliance on manual processes, and enhance risk assessment processes. To manage identified and assessed risks, EY proposes four strategies: risk avoidance, risk transfer (to a third party), risk mitigation (reducing the probability), and risk acceptance (controlling and monitoring expected risks) [28].

Another 2021 Ernst & Young study covered 21 European banks, most implementing compliance functions using traditional compliance risk monitoring models. However, there is much interest in adopting technologically advanced models [29].

Deloitte has issued a report on compliance risk assessment in 2015 [30]. In the methodology, they distinguish the legal, financial, business, and reputational impact of inherent risks. The main practical recommendations are data collection from cross-functional specialists, building on existing content (like reports) and methodology, clear risk ownership for transparency, and delivering useable and actionable risk evaluations (priorities, action plans, monitoring). Further recommendations are using simple language and regularly repeating the risk assessment [30].

Deloitte has also developed a Systematic Integrity Risk Analysis (SIRA) methodology that covers all relevant integrity risks and meets the risk assessment requirements outlined in the 4th Anti-Money Laundering Directive. The main steps of the risk analysis are to determine inherent risk, identify controls, determine managed risk, and define mitigating measures [31]. The SIRA methodology also outlines preparation and closing steps after a risk analysis. Deloitte and EY recommend using Robotics Process Automation (RPA) to reduce compliance costs and increase process reliability and regulatory compliance [28, 31].

4 Discussion and Managerial Implications

4.1 Discussion

While the scientific, and business literature generally agree on a risk-based approach to compliance, it is vital to highlight one condition. The risk-based approach works if the regulators empower the businesses and believe they know the risks they face best and should, therefore, be empowered to decide how to identify, mitigate, and manage those risks. In a legal environment, where the regulators think they know the best will go to detailed regulations where a rule-based approach might be more suitable.

One problem with the requirements-centered approach is that regulations are created to respond to crimes (i.e., cybercrime). Following a strictly reactionary strategy to compliance management will expose the organizations to new risks, for example, due to changes in the organization's digital, social, and legal context. Naheem [16] argues for a holistic approach to risk, as organizational failures and fraud often transcend business unit levels and add up across processes. Therefore, the authors propose using a process-based approach to compliance risk management, which could be supplemented with a requirements- or rule-based approach.

The “explain or comply” approach, required by regulatory supervision in some cases, means that organizations that do not comply with laws or codes must explain each noncompliance. Explaining is only valid if it is about meaningful reasoning and not rhetorical misleading by lessening the severity of potential damages or losses in other terms [32].

Bello and Harvey [33] highlight the difficulties of the risk-based approach to anti-money laundering compliance (e.g., confusion on whether the organization’s risk perception is in line with the regulator’s) and propose the uncertainty-based approach as an alternative. The latter would provide a better understanding of the risk problem within the AML domain and would be more cost-effective while aligning the interests of banks and regulators. The authors of this paper would like to emphasize that AML is a unique field of compliance where the probability assessment of a potential outcome could be even more difficult than in other areas of regulatory compliance.

The answers to the research questions are presented below.

RQ1. What meanings does the term compliance risk contain? The collected definitions mainly reflect on the causes and impact of compliance risks. The Basel Committee [1] definition is widely accepted as a reference. This definition has a cause and an impact part. Causes of noncompliance may be ”failure to comply with laws, regulations, rules, related self-regulatory organisation standards.” The impact is divided into three areas: legal sanctions, financial loss, and loss of reputation.

On the sources side, market standards [7] and voluntarily chosen requirements [6] could be added to the Basel definition.

However, the impact side is significantly different in the ISO 37301:2021 definition. While the consequences are not divided into three, the likelihood of the occurrence is an essential part of the standard’s definition [6].

In this paper, we propose a new definition of compliance risk as follows:

Compliance risk refers to an event with the likelihood of potential regulatory, financial, or reputational losses for the organization due to noncompliance with regulations or voluntary obligations.

RQ2. What does the compliance risk assessment process look like according to the literature?

1) The literature is not uniform, not even in terms of compliance risk management activities (confusing mitigation, control and monitoring activities and the relation of these). Few specific methods and techniques have been developed to identify and model compliance risks. Scientific and business reports hardly refer to the published ISO guidelines and standards relevant to this topic. Adopting general risk assessment approaches and methodology in the specific compliance risk area would be beneficial.
2) Many publications see value in the close cooperation of operational risk management and compliance management. The information generated in the internal control (internal audit, operational risk management, and compliance management) frameworks can be reused using a structured approach. Cooperation with operational risk management and internal audit can reduce compliance costs.
3) Using mitigation levels as a guideline to estimate the noncompliance impact [15] uses the risk management process backward, creating unreliable risk assessment and inconsistency in the whole process. The severity of risk impact should be assessed independently from the analysis of the controls. Organizations need help to quantify risk impact in practice, and making good use of historical data is necessary but insufficient since it can be incomplete or misleading [12].
4) Nicolas and May [7] highlight that controls are part of the inherent risk. In a sense, actual controls create potential risks. Studies show that partial risks might stay hidden if only the traditional risk matrix (probability vs. impact) is applied [4, 5]. The Partial Risk Map (PRISM) methodology adds the detection factor of failure modes and gives a more efficient and detailed view of the risk assessment results. Root [34] emphasizes that the root causes of compliance violations should be identified.
5) Finally, individual risk assessment, besides group assessment, is highly underrepresented in the literature. Visualization of risk assessment and setting up cross-functional teams in compliance risk identification and assessment is concluded from the study of Esayas [15] to compensate for the difficulties of individual, professional, and verbal interpretations of risks. A visual representation of risks can facilitate a shared understanding and more straightforward communication on compliance issues.

Based on the above inconsistencies of the reviewed literature and the relevant ISO guidelines, the authors propose the following structured compliance risk assessment process.

Among others, compliance obligations provide inputs to the compliance risk assessment process, which has three main steps (Figure 2):

1) compliance risk identification,
2) compliance risk analysis (analysis of probability and impact of noncompliance and assessment of ease of detection by current controls),
3) compliance risk evaluation (ranking of risks).

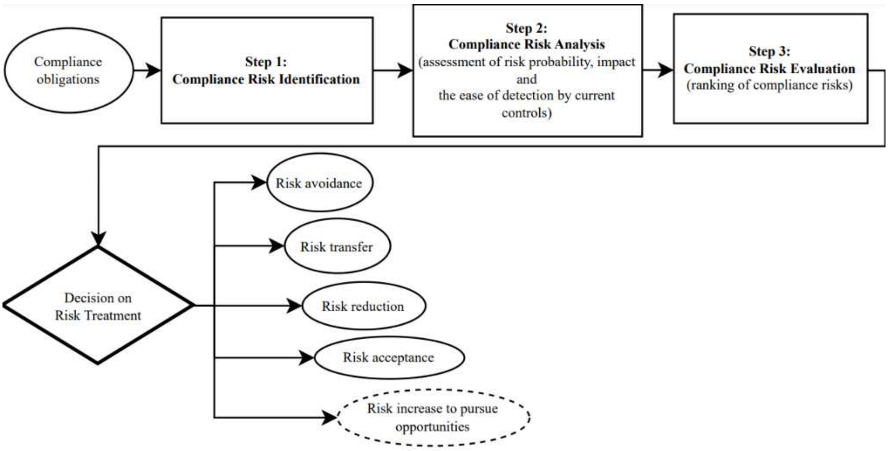

Compliance obligations
Step 1:
Compliance Risk Identification
Step 2:
Compliance Risk Analysis
(assessment of risk probability, impact and the ease of detection by current controls)
Step 3:
Compliance Risk Evaluation
(ranking of compliance risks)
Decision on Risk Treatment
Risk avoidance
Risk transfer
Risk reduction
Risk acceptance
Risk increase to pursue opportunities

Figure 2
A structured compliance risk assessment process

The structured compliance risk assessment process ends with decision-making. A privacy breach example explains the risk treatment strategies. An organization can avoid the risk of a privacy breach by not using certain technologies. Or an organization may transfer the risk to third parties or external service providers with particular expertise in customer data protection. Alternatively, it can reduce the risk of a privacy breach by investing in measures such as encryption or firewalls. If the organization accepts a certain level of risk, it may plan to respond to incidents by, for example, recovery plans and early detection systems. In some cases, organizations may increase the risk to pursue a business opportunity.

RQ3. What are the gaps in the current literature that future research might explore?

1) Robust methodologies are scarce. Developing robust methods for quantifying compliance risk can improve the accuracy of risk assessment. Future research could propose models for quantifying compliance risk.
2) Research that examines the integration of compliance risk assessments with other types of risk assessments (such as operational risks) can provide a more holistic picture of an organization's risk profile.
3) Data analysis for compliance risk assessment is an emerging field. Future research could explore the intersection of technological innovations with compliance risk assessment.

4) Assessing the effectiveness of compliance risk reduction is still in its infancy. Future studies could develop methodologies to measure the impact of different compliance risk management strategies.
5) Finally, DORA compliance lies in the intersection of information security and compliance management. Future research need to explore what are the compliance aspects of ICT and cyber-risks and how the frameworks can be integrated in theory and practice.

4.2 Managerial Implications

First, similarly to quality management, compliance management was mainly seen as a cost rather than a value-creating function [12]. Nowadays, managing integrity and compliance is critical in creating value and improving reputation.

Transforming legal requirements to risks is a challenge in itself. How can regulations be transformed into threat models and later risks? Darimont and Lemoine [35] propose the KAOS methodology, which transforms regulatory requirements into goals and, after modeling the goals, identifies anti-targets as threats. Also, creating and using templates in risk identification and assessment facilitates communication among experts [15].

Objectivity in assessing consequences can be introduced by creating a structure and criteria for assessing compliance risk. A structured approach can reduce subjectivity in making compliance decisions and resource allocations. Better results can be achieved with a structured approach than an unstructured brainstorming session [15].

When developing a remediation and testing plan, being realistic about what can be accomplished in the given time frame is crucial [7].

As for DORA compliance, Chief Information Security Officers working with DORA can use the ISO/IEC 27001: 2022 standard as a starting point [36]. Compliance with ISO/IEC 27001: 2022 means that an organization has implemented a system to manage risks related to the security of data owned or operated by the company. This standard helps organizations recognize risks and proactively identify and address gaps. An information security management system implemented according to the standard is a tool for risk management, cyber resilience, and operational excellence. ISO 27005: 2022 standard [37] provides a framework and approach to information security and cybersecurity risk management. Most risk management methodologies are derived from this international standard.

The cooperation of compliance and risk management in a coordinated manner, based on shared goals, principles, and values, and having the processes and organizational structures in place to monitor the organization's activities continuously can create value [22].

The Ernst & Young report [28] proposes investment in emerging technologies, investment in the right processes and actions, specialized training and reskill of people, and last but not least, "set the right tone at the top".

Internal audit (responsible for reviewing the effectiveness of the compliance controls) needs to have an in-depth understanding of the various compliance risks to judge the appropriateness of the risk assessment strategy and methodology. Likewise, the board might need training on compliance risks to be able to and be motivated to carry out their responsibility of supervising the compliance risk management of the organization. Root [34] points to the multiplicity of reasons for compliance violations, such as difficulties overseeing compliance programs and the lack of an integrated compliance culture in the corporate structure. Companies starting the "compliance journey" may face resistance from first-line business units [7].

Conclusions

Public or private organizations, regardless of size, sector, and geographic location, are subject to certain regulatory compliance risks. This article has collected various definitions of compliance risk that reflect the prevalence, principles, and scope of compliance management based on a review of nine studies on compliance risk assessment from 2005-2021. The main findings of this research are:

1) A new definition of compliance risk was created based on a combination of several previous definitions.
2) Adopting general risk assessment approaches and methodology tailored to the specific compliance risk area facilitates cooperation with other internal control functions, like operational risk management and internal audit.
3) Analysing controls is a significant part of risk assessment since detectability is an essential part of the risk.
4) Compliance risk assessment can be improved by using structured frameworks and methodology. For this, the authors have developed a structured compliance risk assessment process (Figure 2).

The most important limitation of this study is that some studies on compliance risk assessment may have been excluded from this review due to the inclusion and exclusion criteria developed by the researchers. Future research will focus on quantifying compliance risk to improve the accuracy of risk assessment. Further research could study the use of data analysis in compliance risk assessment. Finally, research needs to explore the compliance aspects of ICT and cyber risks and how the frameworks can be integrated in theory and practice, as necessary for DORA compliance in the future.

References

[1] Compliance and the compliance function in banks, Basel Committee on Banking Supervision, Bank for International Settlements, 2005, https://www.bis.org/publ/bcbs113.pdf

[2] ISO: Compliance management systems - Guidelines, 2014, ISO 19600:2014

[3] ISO: Risk management - Guidelines, 2018, ISO 31000:2018

[4] Bognár F., Benedek P.: A novel risk assessment methodology: a case study of the PRISM methodology in a compliance management sensitive sector. Acta Polytechnica Hungarica, 2021, 18, 7, pp. 89-108, https://doi.org/10.12700/APH.18.7.2021.7.5

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

[6] ISO: Compliance management systems – Requirements with guideance for use, 2021, ISO 37301:2021

[7] Nicolas, S., May, P. V.: Building an effective compliance risk assessment programme for a financial institution. In the Journal of Securities Operations & Custody, 2017, 9, 3, https://www.henrystewartpublications.com/sites/default/files/Nicolas%2C%20Stephanie%20%26%20May%2C%20Paul%20JSOC%209-3.pdf

[8] Regulation (EU) 2022/2554

[9] Grima, S.; Marano, P.: Designing a Model for Testing the Effectiveness of a Regulation: The Case of DORA for Insurance Undertakings. Risks 2021, 9, 206, https://doi.org/10.3390/risks9110206

[10] Pavlidis, G.: Europe in the digital age: regulating digital finance without suffocating innovation, Law, Innovation and Technology, 2021, 13:2, 464-477, https://doi.org/10.1080/17579961.2021.1977222

[11] Losiewicz-Dniestrzanska, E.: Monitoring of Compliance Risk in the Bank, Procedia Economics and Finance, 2015, 26, pp. 800-805, https://doi.org/10.1016/S2212-5671(15)00846-1

[12] Birindelli, G.; Ferretti, P.: Compliance risk in Italian banks: the results of a survey, Journal of Financial Regulation and Compliance, 2008, 16, 4, pp. 335-351, https://doi.org/10.1108/13581980810918404

[13] Sathye, M., Islam, J.: Adopting a risk-based approach to AMLCTF compliance: the Australian case, Journal of Financial Crime, 2011, 18, 2, pp. 169-182, https://doi.org/10.1108/13590791111127741

[14] Birindelli, G., Ferretti, P.: Compliance function in Italian banks: organizational issues, Journal of Financial Regulation and Compliance, 2013, 21, 3, pp. 217-240, https://doi.org/10.1108/JFRC-07-2012-0027

[15] Esayas, S., Mahler, T.: Modelling compliance risk: a structured approach. Artif. Intell. Law, 2015, 23, 3 (September 2015), 271-300, https://doi.org/10.1007/s10506-015-9174-x

[16] Naheem, M. A.: Anti-money laundering/trade-based money laundering risk assessment strategies – action or re-action focused?, Journal of Money Laundering Control, 2019, 22, 4, pp. 721-733, https://doi.org/10.1108/JMLC-01-2016-0006

[17] Achkasova, S.; Bezrodna, O.; Ohorodnia, Y.: Identifying the volatility of compliance risks for the pension custodian banks. Banks and Bank Systems, 2021, 16(3), 113-129, https://doi.org/10.21511/bbs.16(3).2021.11

[18] ENISA: Cloud computing: benefits, risks and recommendations for information security. In: Catteddu D, Hogben G (eds) European Network and Information Security Agency, 2009

[19] Federal Reserve Supervisory Letter SR 08-08, 16th October, 2008, https://www.federalreserve.gov/boarddocs/srletters/2008/sr0808.htm

[20] Asenov, E.: Characteristics of Compliance Risk in Banking. Economic Alternatives, 2015, 4, 20-28, https://www.unwe.bg/uploads/Alternatives/2-Asenov.pdf

[21] Compliance in the spotlight, Deloitte LLP, 2013. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/about-deloitte/deloitte-uk-audit-compliance-in-the-spotlight.pdf

[22] Lord, T.; Smith, M.: Compliance + risk management = value, PwC, 2011, https://www.pwc.com/gx/en/oil-gas-energy/publications/pdfs/compliance-plus-risk-management-equals-value.pdf

[23] Matsuo, A.; Staines, K.: Effective compliance programs – Updated DOJ guidance, KPMG Regulatory Alert, 2020, https://advisory.kpmg.us/content/dam/advisory/en/pdfs/2020/effective-compliance-programs.pdf

[24] Gerlach, J., Stryker, N., Matsuo, A., Dookhie, R.: Harnessing data and analytics to transform compliance, KPMG, 2017, https://advisory.kpmg.us/articles/2018/harnessing-data-analytics-to-transform-compliance.html

[25] A practical guide to risk assessment, PricewaterhouseCoopers, 2008, https://web.actuaries.ie/sites/default/files/erm-resources/a_practical_guide_to_risk_assessment.pdf

[26] Franco A., Woolgar, O.: Maximising the benefits from your compliance monitoring programme, PricewaterhouseCoopers, 2022, https://www.pwc.com/jg/en/services/advisory/blogs/maximising-benefits-from-compliance-monitoring-programme.html

[27] Integrity, Compliance & Ethics, Ernst and Young, 2018, https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/assurance/assurance-pdfs/ey-integrity-compliance-ethics.pdf

[28] Reshaping the future of compliance with emerging technologies, Ernst and Young, 2021, https://assets.ey.com/content/dam/ey-sites/ey-com/en_in/news/2021/07/ey-forensics-survey-reshaping-the-future-of-compliance-with-emerging-technologies.pdf

[29] Crotaz, S., Lown, J., Niedbala, C.: Compliance transformation: how banks can leverage opportunities now. Ernst & Young, 2021, https://www.ey.com/en_gl/banking-capital-markets-risk-regulatory-transformation/compliance-transformation-how-banks-can-leverage-opportunities-now

[30] Compliance risk assessments, The third ingredient in a world-class ethics and compliance program, Deloitte, 2015, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-aers-compliance%20riskassessments-02192015.pdf

[31] Compliance Risk Management Powers Performance, Deloitte, 2018, https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-compliance-risk-management-powers-performance.pdf

[32] Shrives, P.; Brennan, N. M.: Explanations for corporate governance non-compliance: A rhetorical analysis, Critical Perspectives on Accounting, 2017, 49, pp. 31-56, https://doi.org/10.1016/j.cpa.2017.08.003

[33] Bello, A. U., Harvey, J.: From a risk-based to an uncertainty-based approach to anti-money laundering compliance. Security Journal, 2017, 30, 24-38, https://doi.org/10.1057/s41284-016-0002-0

[34] Root, V: The Compliance Process, Indiana Law Journal, 2019, 94, 1, Art. 5, https://www.repository.law.indiana.edu/ilj/vol94/iss1/5

[35] Darimont, R.; Lemoine, M.: Goal-oriented analysis of regulations, REMO 2V06: international workshop on regulations modelling and their verification and validation, Luxemburg. 2006, http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-241/paper9.pdf

[36] ISO/IEC: Information security, cybersecurity and privacy protection – Information security management systems, Requirements, 2022, ISO/IEC 27001:2022

[37] ISO/IEC: Information security, cybersecurity and privacy protection – Guidance on managing information security risks, 2022, ISO/IEC 27005:2022

Augmented Reality System for Instructed and Visualized Pallet Loading

Nikolina Dakić, Vladimir Jurošević, Vule Reljić,
Ivana Milenković, Slobodan Dudić, Jovan Šulc

University of Novi Sad, Faculty of Technical Sciences,
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia,
nikolinadakic@uns.ac.rs; vladimirjurosevicvl@uns.ac.rs; vuketa90@uns.ac.rs;
ivanai@uns.ac.rs; dudic@uns.ac.rs; sulc@uns.ac.rs

Abstract: Optimally loading pallets is a challenge and can be a time-consuming task for many industrial and logistics applications. To help solve this problem, we proposed an augmented reality (AR) application that assists warehouse workers during the pallet loading process. For this reason, an algorithm for solving the Pallet Loading Problem (PLP) was defined first. All input data (name, dimensions and quantity of boxes that need to be loaded on the pallet), which actually form a loading order, are then transferred to the warehouse with the generated QR code. A warehouse worker scans the QR code and gets a visual representation of how to load the boxes onto the pallet on the lens of their smart glasses. The experimental results show that the proposed system successfully facilitates and speeds up the process of shipping products from the warehouse to customers.

Keywords: pallet loading; augmented reality; packing problem; optimization

1 Introduction

When delivering a certain amount of products to customers, PLP often occurs. It is often the case that requests are received from several customers at the same time or that one customer requests the delivery of several different types of products. In order to optimize costs and meet multiple requirements with one delivery, suppliers try to make maximum use of the available cargo space of a pallet, truck or other means of transport [1]. Therefore, it is necessary to carefully plan the loading of transport boxes on pallets. However, usually the time for loading pallets is very limited because the customer expects delivery within the appropriate time. In the background, many optimization methods need to be applied to define the best way to load pallets, which can be done by workers in production facilities, who are usually counted as “unskilled labour”. The goal is for the worker to receive simple, unambiguous instructions on how to load the transport boxes onto the pallet, and all this in a very short time.

By carefully searching the literature and following modern technological achievements, it can be concluded that AR is emerging as a promising solution [2, 3]. Through the integration of virtual objects into the real space, workers in warehouses can be provided with clear guidance on the placement of shipping boxes in a simple way.

In order to successfully develop a system that would meet all the mentioned requirements, it is, first of all, necessary to have a good and efficient algorithm that solves the PLP. By searching the literature, it is possible to find a few of them, which will be discussed in Section 1.1. The algorithm used in this paper, with all constraints and defined optimization criterion is presented in detail in Section 2.1.

For the successful functioning of any algorithm, it is necessary to provide appropriate input data. In this paper, the input data is contained in a QR code generated by a desktop application as a part of a loading order, which is then forwarded to the worker. The developed desktop application that generates loading orders is presented in Section 2.2.

Finally, it is crucial to visually show the worker how to load the boxes on the pallet. In this paper, this is made possible by the use of AR, through the developed AR application. Upon receiving the loading order, the worker runs the AR application and scans the QR code, starting the algorithm, which is then executed in the background of the application.

A detailed description of the developed AR application is given in Section 2.3. Testing of the developed applications and the most important results of the research are shown in Section 3, along with the discussion and directions for further research. The most important conclusions are given in the last Section.

1.1 Literature Review

AR has emerged as a transformative technology in the logistics sector, offering innovative solutions to enhance process optimization, efficiency, accuracy, and overall operational flexibility and effectiveness [4-6]. Significant contributions can be observed in various activities, such as:

1) *Warehouse management*: AR facilitates real-time inventory tracking and management. By providing instant information about stock levels and locations, AR helps in reducing stockouts, overstock situations, and streamlining the replenishment process. In addition, AR is employed for hands-free picking and packing in warehouses. Workers equipped with AR devices, such as smart glasses, receive real-time information on item locations, reducing errors and increasing productivity [7].

2) *Navigation and route optimization*: AR assists in route planning and navigation for logistics and transportation. By overlaying digital information on the real environment, AR aids drivers in identifying optimal routes, reducing delivery times, and minimizing fuel consumption [8].
3) *Maintenance and repairs*: In field service operations, technicians use AR to access relevant information and step-by-step instructions while on-site. This ensures accurate and timely execution of repairs and maintenance tasks, minimizing downtime [9].
4) *Training, onboarding and collaborative operations*: AR is employed to train logistics personnel, offering immersive experiences for simulating diverse scenarios and thereby enhancing skills and preparedness for real-world situations. Additionally, AR fosters collaboration within logistics teams by enabling remote experts to deliver real-time guidance to on-site workers through AR devices. This not only improves communication but also enhances problem-solving capabilities [10].
5) *Quality control*: AR empowers inspectors to overlay digital information onto physical products, facilitating the identification of defects and ensuring strict adherence to quality standards [6].
6) *Augmented packaging*: AR is used to enhance packaging processes. By overlaying digital instructions on packaging materials, workers can ensure accurate and efficient packing, reducing errors and optimizing the use of packaging materials [3].

Overall, AR is based on the insertion of virtual objects into real 3D human space. The user perceives the inserted virtual objects as if they were real. A variety of practical applications within warehouse settings are explored in existing literature, showcasing the diverse potential uses of AR in this context.

Porter et al. used spatial AR for prototyping new human-machine interfaces, such as control panels or car dashboards [11]. Within the prototyping system projectors are used to present the visual appearance of controls on a product prototype. In order to provide real-time interactions with the controls, finger tracking is applied. This kind of technology can be used for the quick and inexpensive creation and evaluation of device interface prototypes.

One of the possible applications of AR for material and product handling purposes is shown in [12], where AR is proposed as a helpful tool for picking the items listed on the order. A simulation was given, using Head Mounted Displays (HMDs) and virtual storage for this purpose. Schwerdtfeger and Klinker performed a similar experiment where a real storage environment and smart glasses were used by many participants, in order to obtain useful data regarding the adaptation of people to AR and new technologies [13]. It is proven that users were quick to adapt to AR technology.

Study [14] shows an application of AR in a warehouse, designed to facilitate the work of the driver that needs to be focused both on their surroundings while driving, and the process of loading pallets. The driver gets information about their position in space via a 3D model, the path to follow, and the contents of the pallet that needs to be loaded.

In the realm of intralogistics, the application of AR holds potential benefits, particularly in the context of palletization. The literature encompasses diverse approaches and suitable algorithms for solving PLP, suggesting opportunities for guidance and optimization within this field.

Martins and Dell analysed Multidimensional Packing Problems, such as Bin Packing problem, Knapsack Problem, Strip Packing Problem, and PLP in order to understand the differences among the execution times of the algorithms created for solving the mentioned loading problems [15]. In the case of PLP, the items of the same height are loaded in vertical layers. Therefore, the problem is reduced to finding the two-dimensional (2D) arrangement which maximizes the number of packed items in a layer. It is also considered that PLP is an example of an unconstrained 2D Knapsack Problem. Considering dimensionality, the type of assignment, the assortment of boxes, and the assortment of small items, the PLP is identified as 2D. When solving the problem, the aim is to minimize unused space and find the optimal arrangement of the boxes. Heuristics and approximation algorithms are discussed, which also include a tree search with Graph-theory. In this case, the PLP is classified as a non-deterministic polynomial time (NP) problem.

A real-world application intended for packing boxes of heterogeneous size, shape, and weight is discussed in [16] to optimize their arrangement. Constraints such as geometry (border of pallet, overlap, height threshold), stability, and fragility are taken into account. To solve the PLP, a metaheuristic framework is used to try to escape from local minimums, and achieve reasonable response time. Also, it is said that by using only heuristic techniques, the algorithm will get trapped in a local optimum. To find an optimal solution, a hybrid genetic algorithm is used.

The PLP is defined as the problem to find the optimal layout for packing a set of identical boxes on a rectangular pallet in [17]. Furthermore, it is said that it is justified to consider the height of the boxes as fixed, which reduces a three-dimensional (3D) problem to a 2D PLP. As a result, the PLP algorithm is applicable in logistics where distributed goods need to be packed in layers on uniform pallets. Also, pallet area utilization and stability were marked as important, with an emphasis on stability. An Integer Programming (IP) formulation of PLP is used.

Ha and Nananukul analysed an air cargo 3D packing model with constraints: overlapping, the practical position of cartons, cargo priority, and weight limitation applied on a forwarder [18]. The idea was to improve the model to be able to operate with a large number of input data in a short time. The improvement of the original model is achieved by relaxing the original constraints, in order to address unsolved cases from the original model. As small deviations are now allowed, minor movements of the box are made in order to meet the additional criteria.

A carton-to-pallet loading problem is discussed in [19]. Products are grouped in boxes. Also, boxes containing similar products (soft drinks and beer) are grouped together in one class. These boxes have similar dimensions as well. In this way, different groups of products are loaded on the same pallets. When an order from a customer is ready to be packed, boxes of similar size form a group which is then used as input to the algorithm, reducing the number of possible combinations. The algorithm sequentially allocates product boxes to pallets using ten sub-steps. Each box, when allocated to a pallet, will take up a certain percentage of the pallet volume. A product group/class minimal percent rule is given.

Constraints of dynamical stability are taken into account in [20], so that there is no box movement caused by forces such as gravity and horizontal forces that occur during the transportation of the container. To solve this problem, an approach based on the metaheuristic BRKGA (Biased Random-Key Genetic Algorithm) is used in order to analyse the geometrical structure of the problem. Static and dynamic stability is discussed. During loading/unloading and transportation, it is important to pack the cargo in a physically stable manner, forbidding the movement of the cargo, especially rotating or falling.

To define the positioning of the boxes in [21] a coordinate origin in the bottom-left-behind of the bin (Bin packing problem) is used. Solving the loading problem begins with tackling it in sub-problems. It is necessary to define the box location and the optimal way to place it in the defined location. Constraints that are taken into account are stability, rotation, weight, different products, and product agglomeration. An algorithm based on Extreme Points (EP) is used. The height of a layer depends on the first box loaded in the layer. The weight constraint reflects in placing boxes with heavier products at the bottom, to prevent them from crushing the lighter ones. There are, however, some deviations where heavier boxes do not need to necessarily be at the bottom. In loading with extreme points, instead of packing single boxes, the boxes are grouped and packed in rows whenever possible. The floor boxes are the ones whose top is at the same level as the EP where the present box is to be stored.

Different ways of loading are given in [22]. The loading problem is divided into manufacturer's and distributor's problems. The task is to find a loading pattern using identical boxes for each layer of the pallet. A more complex formulation brings the Distributor's PLP (DPLP), which implies loading non-homogeneous items onto one or more pallets. The paper is based on solving the distributor's problem. In order to minimize delivery costs, each pallet must be loaded efficiently with respect to maximal volume utilization and inner-layer support, making it stable for safe transport. An algorithm for loading different types of rectangle-shaped boxes onto a pallet is defined as NP-hard. The main objective was to maintain stability in addition to volume utilization with a nested beam search.

The PLP characterized as a 3D packing problem with additional requirements to ensure that pallets are stable and that cargo may be handled and transported safely is discussed in [23]. The DPLP and Manufacturer's PLP (MPLP) are both considered. The DPLP considers shipment from a distribution centre to retailers where pallets are composed of mixed cases. On the other hand, MPLP considers shipment from a manufacturer to a distribution centre and deals mostly with similar cases. It is similarly concluded as in [22], that the DPLP deals with heterogeneous items, while the MPLP deals with homogeneous items. It is indicated, to build stable pallets that it is essential to guarantee adequate bottom support for all items. The recommendation is to use 70% of bottom support.

What is particularly interesting for analysis is how to apply the PLP solution in real working conditions, i.e. how to help and show the worker to properly load the boxes on the pallet or how to visualize the packaging solution [24]. AR can be used for this purpose [25].

2 Method

This section presents the entire method for creating AR system for instructed and visualized pallet loading is presented. This method starts with the PLP algorithm in order to solve DPLP where different products are loaded into boxes of different sizes according to customer requirements [26]. Based on this algorithm, two applications were developed: Desktop application for creating the LOADING ORDER and AR application for pallet loading, which then together enable guided and visualized loading of pallets.

2.1 PLP Algorithm

The state of the art indicates the complexity of the PLP issue. Various researchers take into account a number of different factors, such as the dimensions of the boxes, their masses, the position of the centre of mass, the orientation, the order in which the boxes are removed from the pallet, etc.

For the efficiency of the real-time execution of the algorithm, only the planar problem is considered in this study. Precisely, all rectangular boxes are assumed to have at least one edge of the same length, be it length, width, or height. This reduces the 3D problem to 2D one [17]. Given that the PLP resolution process consists of consecutive steps, the flow diagram of instructions is presented in Figure 1.

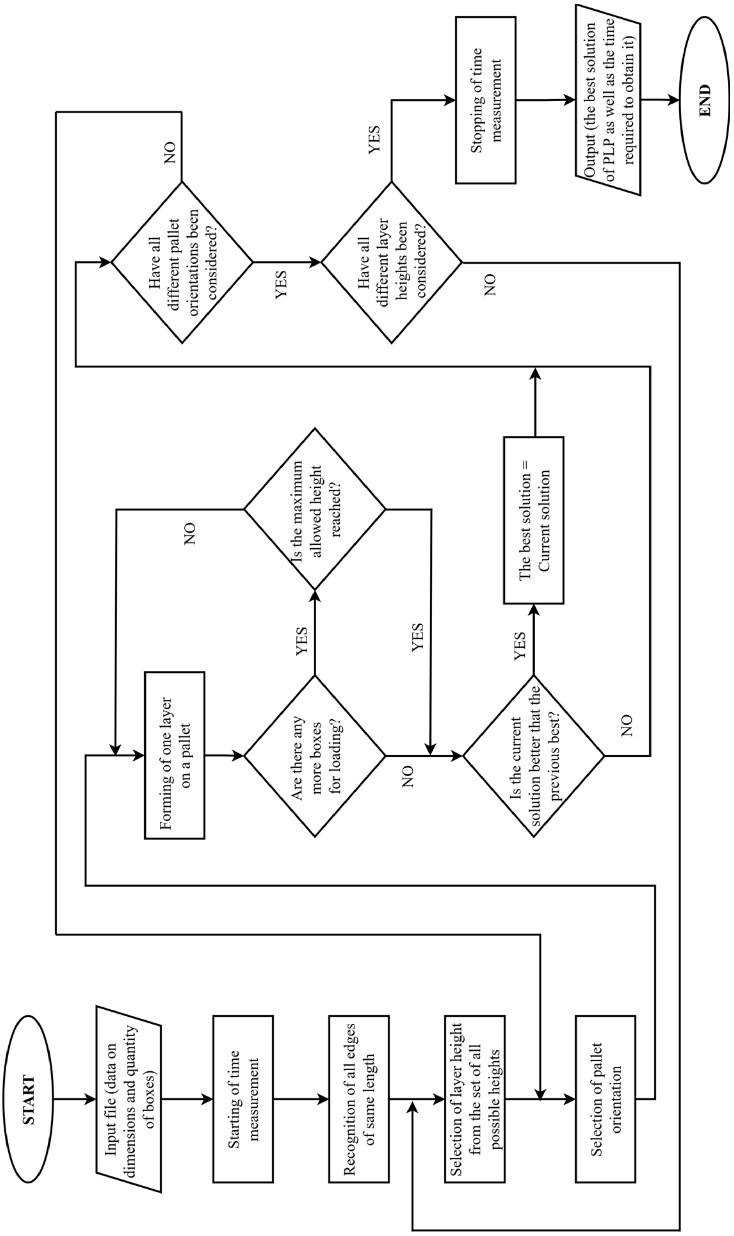

START
Input file (data on dimensions and quantity of boxes)
Starting of time measurement
Recognition of all edges of same length
Selection of layer height from the set of all possible heights
Selection of pallet orientation
Forming of one layer on a pallet
Are there any more boxes for loading?
YES
NO
Is the maximum allowed height reached?
NO
YES
Is the current solution better that the previous best?
YES
NO
The best solution = Current solution
Have all different pallet orientations been considered?
NO
YES
Have all different layer heights been considered?
NO
YES
Stopping of time measurement
Output (the best solution of PLP as well as the time required to obtain it)
END

Figure 1
Instructions flow diagram

As can be seen from Figure 1, a series of steps is undertaken:

1) In the first step, the algorithm recognizes all edges that have the same length. These edges represent the potential height of some layers of boxes on the pallet.
2) In case all the boxes have more than one edge of the same length, the algorithm shows all those edges as a result in the background.
3) For each of the recognized edges (height of the layers), a pallet loading simulation (layer by layer) is performed for two different pallet orientations, until the maximum allowed height is reached or until all the boxes are loaded. The maximum height is predefined in the desktop application.
4) After the first performed simulation, the current solution is declared as the best possible one.
5) In the case that the next solution is better than the current one in terms of the best utilization of the pallet volume, i.e. the existence of the minimum possible "empty space", that solution is declared the best possible.
6) The simulation is repeated until all combinations have been tested.
7) At the very end, the algorithm returns the final optimal solution as a result, as well as the time required to obtain it.

As noted earlier, when the boxes are properly oriented and loaded on the pallet, their common dimension actually becomes the height of a single layer of loaded boxes. In this way, the problem of loading boxes on a pallet is reduced to finding the optimal layout for placing different-sized rectangles on the surface of the pallet or a previous layer of loaded boxes. The idea is to minimize the free space, or "slots" that could occur after placing these rectangles on a given surface. To achieve this, the box with the largest base area is placed first in the new layer of boxes. The formation of one layer ends when no new boxes can be added to it. As discussed earlier, the pallet is then loaded layer by layer until all the required boxes are loaded on the pallet or until the maximum height of the loaded boxes is reached. An example illustrating these algorithmic steps is presented, specifically showcasing the sequence for loading four boxes, as presented in Figure 2. Two different dimensions of the given boxes (a, b) are shown in Table 1. The third dimension of these boxes is the same and as such represents the height layer.

Table 1

Contains the result of comparing in pairs with the final result

Dimensions (mm)		Colour
a	b	
430	280	Yellow
420	250	Green
320	300	Red
300	180	Blue

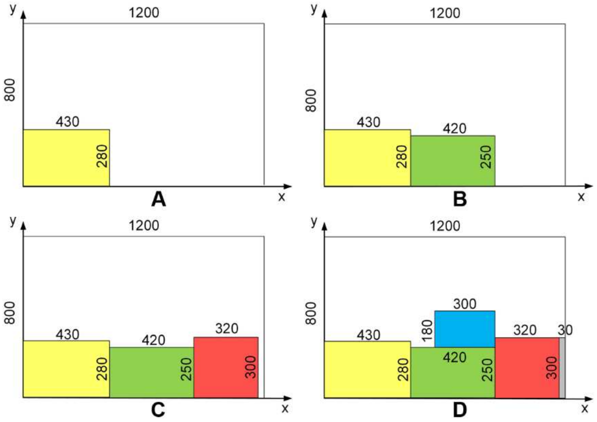


Figure 2
Algorithm functioning - forming one layer on the pallet

The algorithm first selects one of the pallet edges. Since EUR-pallet is used, the dimensions of the pallet are 1200 mm x 800 mm. The box with the longest edge is selected first when forming the layer of boxes, and its long edge is aligned with the long edge of the pallet, i.e. the x axis (Figure 2A). As it can be seen from the Table 1, the box that meets this condition is the yellow box with a 430 mm long edge. This step is repeated, placing new boxes with their long edge parallel to the x axis next to each other, until no more boxes can fit in the current row of the layer. Accordingly, Figure 2A and 2B show the addition of the boxes that have the second and third longest edge – the green box with the edge of 420 mm and the red box with the edge of 320 mm, respectively. After adding these three boxes, a gap, i.e. “slit” appeared between the third box and the edge of the pallet, visible in Figure 2C. As the length of this gap is 30 mm and there are no boxes with appropriate dimensions that could “fill” the gap, it is obvious that no other boxes can fit in the first row and new boxes are to be added to the second one. Further, the first box of the following row needs to be placed above the box from the previous row whose shorter edge, i.e. the dimension parallel to the y axis is minimal, in this case 250 mm (Figure 2D). If this isn’t feasible because of the difference in the longer edge values of the two boxes, the algorithm seeks the box with the next higher value of the shorter edge in the previous row.

If it is detected that placing the next box in the row would result in exceeding the pallet dimensions, the box cannot be added to the row as described. In these cases, the algorithm searches for the next shorter edge that would not exceed the pallet dimension when added to the row, which can either be the short edge of the given box, or any edge of some of the other boxes. In case the short edge of the current box is chosen as optimal, the box is rotated for 90 degrees and placed onto the pallet, making its short edge parallel to the x axis. Otherwise, the observed box is no longer considered for placement in the current row and the next box to be added to the row is searched for among remaining boxes in the same way. Also, if the dimensions of a single box exceed the pallet dimensions, the algorithm does not allow the addition of that box to the pallet.

2.2 Desktop Application for Creating the LOADING ORDER

To seamlessly incorporate the PLP algorithm into an AR application, it is crucial to establish the methodology for generating input data. These inputs include the name, dimensions and quantity of boxes that need to be loaded onto the pallet. If there is a need, it is possible to add different information to each article depending on the needs of the warehouse itself. The colour of the virtual boxes has also been added for easier identification. Based on the input data, a loading order is created, which is then sent to the worker in the warehouse, either electronically or in paper form. To achieve this, a desktop application was developed to generate loading orders whose main menu is shown in Figure 3.

In the first step, the user needs to select the item name from the drop-down list (Figure 3, position 1) and enter the number for each of the selected items (Figure 3, position 2). The general term “item” is used, while the actual product names are entered for each warehouse/production facility individually. One item refers to one shipping box. After entering the item name, the corresponding data related to the dimensions and colour of the boxes are generated in the background. In the same way, after entering the item quantity, the required amount of boxes is generated. In cases of incorrect data entry, e.g. by entering decimal numbers or characters in the “NUMBER OF ITEMS” field, a warning appears and the input field is cleared for re-entry. Confirming the entry of one item is done by clicking the “ENTER” button (Figure 3, position 3), after which a message about successful entry appears on the display. The user needs to confirm that they have read the message by clicking on “OK” or on “X” to close the window. Following this step, the input fields are cleared and the application is ready to enter a new item. By clicking on the “SHOW THE LIST OF ITEMS” button (Figure 3, position 4), the user can view the entered items (Figure 3, position 5) at any time. After completing the list of items, the user generates a QR code by clicking the “GENERATE QR CODE” button (Figure 3, position 6) and the QR code appears on the main form (Figure 3, position 10). The key thing in this step is to generate the QR code that contains the input data for the algorithm related to the products placed inside the boxes. In this way, the QR code represents a means of communication between the desktop application in the back office of the warehouse and the AR application in the warehouse itself. Before generating a QR code, the list of items needs to be displayed at least once. If the user tries to generate a QR code before all items are displayed the user will be warned that execution is not possible and that it is necessary to display the content of the loading order first. Also, if the user tries to enter an already entered product, the application will notify them that the item has already been entered, and that the number of these items can be changed in the entry table named “CURRENT

INPUT”. Some items that have been entered by mistake can easily be deleted from this table by simply deleting the row in which the item is located.

The *LOADING ORDER* is generated by clicking the “GENERATE A LOADING ORDER” button (Figure 3, position 7). Clicking on the “NEW ORDER” button (Figure 3, position 8), all data is deleted and a new loading order is created. In case the user wants to close the program, they need to click on the “EXIT” button (Figure 3, position 9).

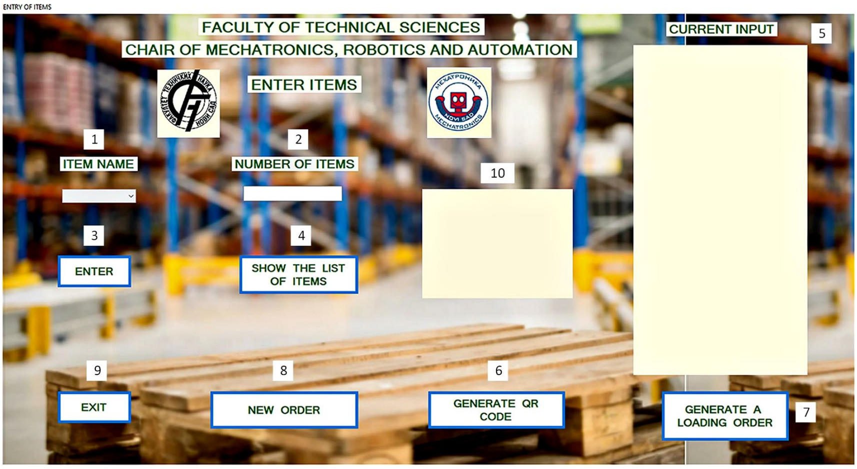
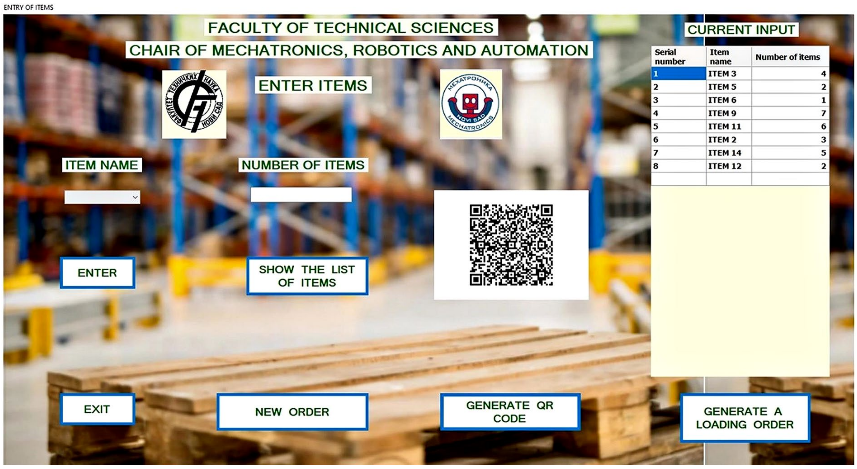


Figure 3
Main menu of the desktop application

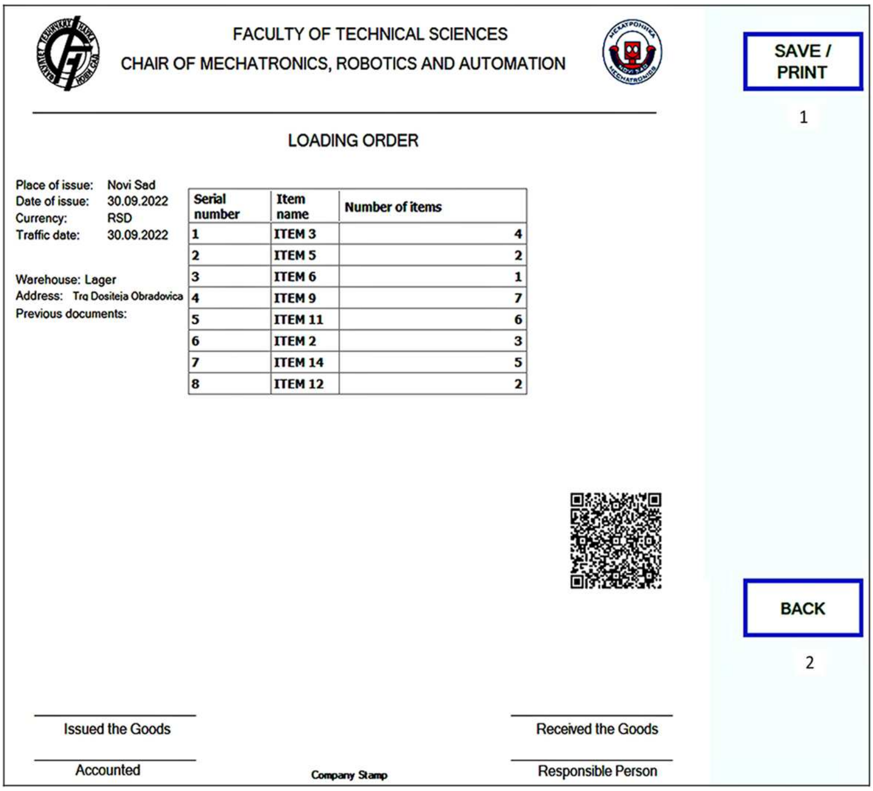
The desktop application with entered items (products) and the generated QR code is shown in Figure 4.



Serial number	Item name	Number of items
1	ITEM 3	4
2	ITEM 5	2
3	ITEM 6	1
4	ITEM 9	7
5	ITEM 11	6
6	ITEM 2	3
7	ITEM 14	5
8	ITEM 12	2

Figure 4
Desktop application with entered items and generated QR code

As a result of, executing the desktop application, a “LOADING ORDER” document is generated (Figure 5) containing all the entered data. This document appears in a new window and allows the user to view the order before clicking the “SAVE/PRINT” button (Figure 5, position 1). By selecting the printer icon, it is possible to save the order in *.pdf format or print it on paper. If the user wants to create the next order, they must first return to the main form by clicking the “BACK” button (Figure 5, position 2), and then select the “NEW ORDER” option.



FACULTY OF TECHNICAL SCIENCES
CHAIR OF MECHATRONICS, ROBOTICS AND AUTOMATION

LOADING ORDER

Place of issue: Novi Sad
Date of issue: 30.09.2022
Currency: RSD
Traffic date: 30.09.2022

Warehouse: Lager
Address: Trg Dositeja Obradovica
Previous documents:

Serial number	Item name	Number of items
1	ITEM 3	4
2	ITEM 5	2
3	ITEM 6	1
4	ITEM 9	7
5	ITEM 11	6
6	ITEM 2	3
7	ITEM 14	5
8	ITEM 12	2

Issued the Goods

Accounted

Company Stamp

Received the Goods

Responsible Person

Figure 5
Output document “LOADING ORDER” with generated QR code

2.3 AR Application for Pallet Loading

The AR application for Android platform was created using the *Unity* programming environment and the *Vuforia* plugin which enabled creating recognition markers. The marker is a CAD model of a standard EUR-pallet. Depending on the used smart device, it is possible to use different markers that imply different pallet views. For the initial testing of the application, in order to check only the functionality, the pallet marker 1 was created (Figure 6A) [27].

On the other hand, under the real working conditions, the real boxes hid the pallet marker which resulted in impossibility of tracking the marker. For this reason, the pallet marker 2 with a changed view was created (Figure 6B).

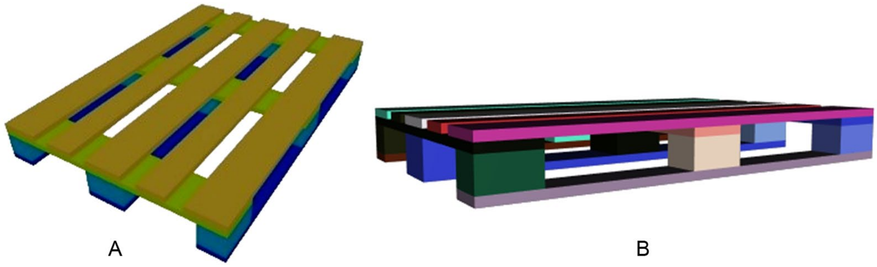


Figure 6
A) Pallet marker 1 and B) Pallet marker 2

The camera of smart device should be oriented in a way that the spatial orientation of the pallet matches the spatial orientation of the markers created based on the CAD model. In other words, it is necessary that the view of the pallet obtained from the camera coincides with the view of the marker created within the application. A certain deviation from the complete coincidence of the two views defined within the scope of recognition is allowed. Once the marker is spotted using the correct camera orientation, there is an area within which the marker will still remain recognized, although there is a slight change in the camera angle relative to the pallet [27]. For both pallet markers, the recognition angle in relation to the vertical axis is from -180° to + 180°, and the recognition angle in relation to the horizontal axis from 0° to 60°.

During testing, the importance of the angle at which the camera is directed towards the pallet was confirmed. If the orientation of the pallet in relation to the camera position is not appropriate, the pallet markers will not be recognized and the display of AR elements will not be generated. If the user moves the camera so that it cannot capture the pallet markers, the display of the AR elements on the screen is lost, but the captured state is preserved. When the user points the camera back at the pallet, the previously recorded state of the application is displayed again on the screen.

As a result of, the application of the PLP algorithm, after recognizing the appropriate marker, the developed AR application provides a display of the positions for placing the boxes on the pallet. These positions are represented as virtual square boxes that are sharp and coloured and as such are easily distinguishable from real boxes. Each virtual box corresponds to exactly one real box.

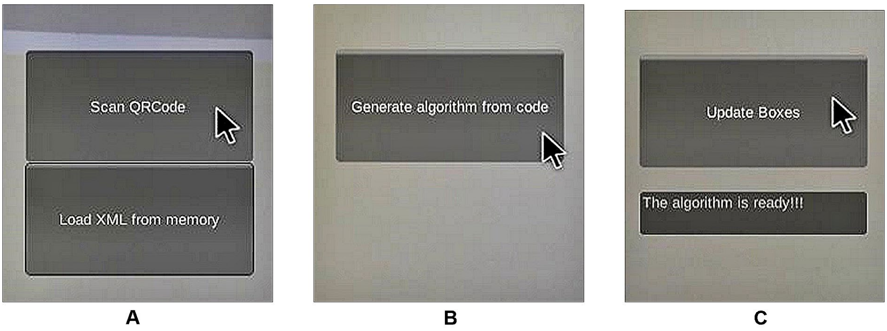
When the generated “Load Order” document arrives at the warehouse, the worker turns on the glasses and puts them on and then scans the QR code on the document by clicking on the button “Scan QRCode” (Figure 7A).

Confirmation of successful code scanning appears in the form of a new button on the screen labelled “Generate algorithm from code” (Figure 7B). At that point, the AR application starts executing. The content of the application is displayed on the screen located on the right lens. The final version of the AR application titled “Slaganje na paletu” is implemented on Vuzix Blade 494 smart glasses [28]. The advantage of using smart glasses, compared to other smart devices such as smartphones or tablets, is that the user’s hands remain free to manoeuvre [25].

Based on the scanned QR code, the application reads the data necessary to generate the algorithm (dimensions, quantity, colour of the boxes). After the successful generation of the algorithm, the process of loading the transport boxes on the pallet begins. To enable the display of virtual boxes on the screen, the pallet marker needs to be detected. A new virtual box is displayed with each new click on the “Update Boxes” button (Figure 7C). It is important to note that boxes that have identical dimensions also have the same colour. The number of identical packing boxes corresponds to the quantity of the item in the loading order.

When the “Update Boxes” button disappears from the application, it means that all boxes have been successfully arranged or the layer height limit has been reached. In this way, it becomes visually clear to the worker that there is no possibility of adding new boxes, i.e. that all boxes are arranged according to the application's algorithm.

All screenshots of the smart device are taken indirectly, from the computer using the “Vuzix View” software [29].



The algorithm is ready!!!

Figure 7

Sequence of steps in the AR application for pallet loading

3 Results and Discussion

The research related to the application's functionality is divided into two phases. The first phase involved the development and testing of the first version of and algorithm that provides the display of virtual boxes. Pallet marker 1 was developed for this purpose [27]. Algorithm execution ran flawlessly while executing the guidance and visualizing loading of the virtual boxes. However, a problem arose with placing the first real box on the pallet. The placed box partially covered the pallet marker 1 and thus prevented the detection of the marker and the execution of the algorithm. For this reason, the second version involved first the development of an additional marker (pallet marker 2), and then testing the developed applications in reality.

3.1 The First Version of the Developed AR Application

After following all the necessary steps to successfully launch the AR application and recognize the pallet marker 1, the application was tested using different cardboard boxes (Figure 8). The first set of test data is randomly defined and is given in Table 2. The common dimension of all boxes is equal to *200 mm*. Figure 8A shows the first two boxes (since boxes that have identical dimensions are represented by the same colour and there is no obvious boundary between the boxes, it may not be apparent) and the cursor that allows the user to add the next box. An illustration of the third box added to the pallet, identical to the previous two, is given in Figure 8B. It should be noted that the combination of the first three identical boxes was not chosen intentionally, but such a combination was determined by an algorithm according to the set of input data. Figure 8C shows the first two layers of boxes (four cyan and four yellow boxes) loaded on the pallet, and Figure 8D shows the pallet after loading four layers. The final result, after loading all the boxes given in Table 2, is given in Figure 8E.

Table 2

First set of test data

Item	Dimensions [mm]			Number of items	Color
	Width	Length	Height		
Item 1	900	200	360	4	Cyan
Item 2	440	460	200	2	Navy
Item 3	200	330	600	5	Yellow
Item 4	490	420	200	4	Magenta
Item 5	580	200	500	3	Lime
Item 6	200	380	460	6	White

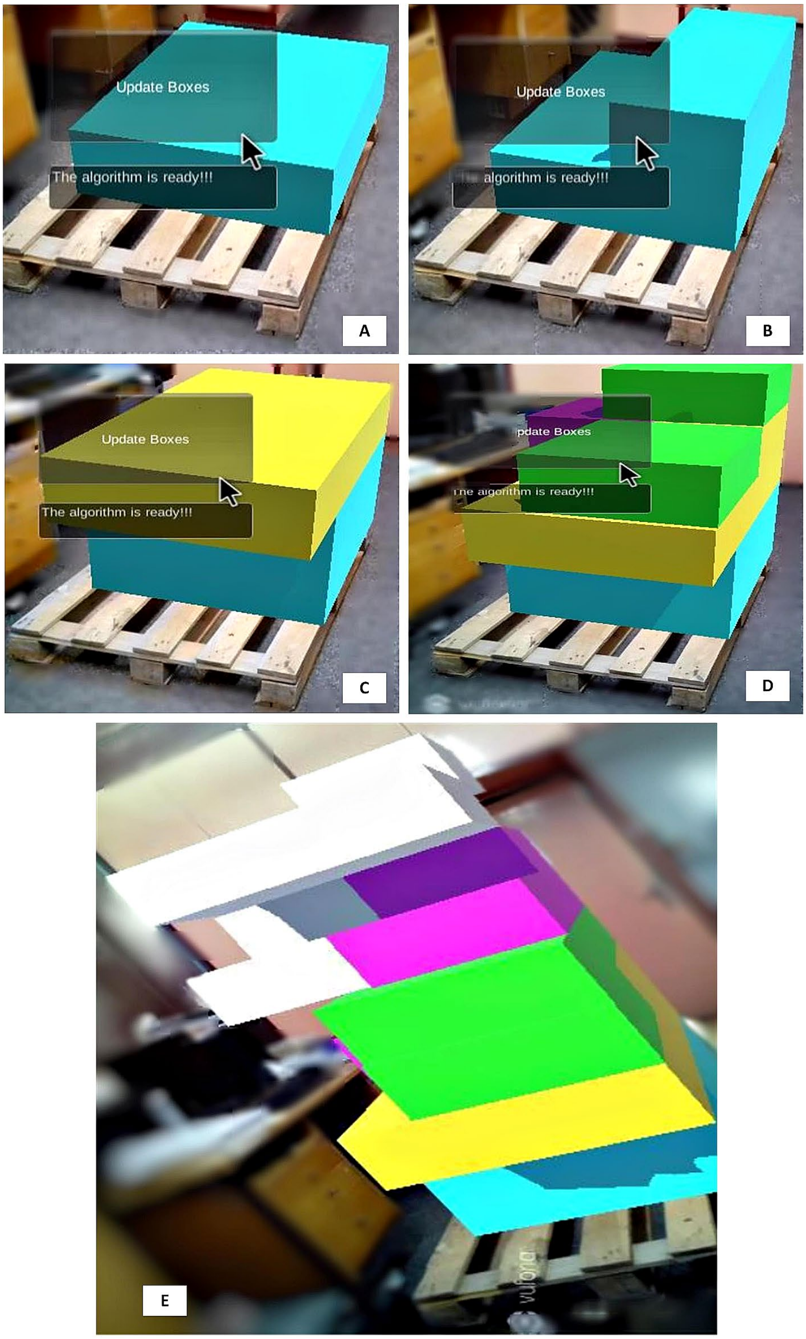


Figure 8
One example of the first test of an AR application

Relying on pallet marker 1 was found to be unreliable. There was an interruption in the operation of the AR application, due to the marker being covered by real boxes in real working conditions. To overcome these issues, within the second version of the AR application, all tests were conducted using pallet marker 2.

3.2 The Second Version of the Developed AR Application

The goal of the second version of the application was to enable the simultaneous display of real boxes loaded on a pallet and virtual boxes on the screen of smart glasses.

The second set of test data was obtained by collecting real boxes. Deviations ranged from several millimetres to one centimetre are neglected, as they do not impair the stability of a formed layer. The second set of test data is given in the Table 3. The QR code containing the data related to these boxes was generated. By scanning this QR code, the algorithm in the background application receives the data and is ready for operation.

In this case, the virtual boxes are set to be transparent so that the positions of real and virtual boxes can be compared, otherwise the virtual boxes would overlap or cover the real boxes loaded on the pallet. Additionally, unlike the first case, the transparency of the boxes allows the user to see the boundaries between boxes of the same colour (identical boxes). The result of applying the second set of input data is given in Figure 9. The successfully solved PLP in this case can be seen in Figure 9E (virtual boxes are shown in different colours). Using this approach, the actual loaded box corresponds to the virtual box. A white line is also added as the contour line of the virtual boxes. A view of the physically loaded boxes is given in Figure 9F. The arrow points to the user's point of view when displaying the loading sequence given in Figure 9E as well.

Table 3

Second set of test data

Item	Dimensions [mm]			Number of items	Color
	Width	Length	Height		
Item 1	400	580	140	1	Red
Item 2	280	400	280	2	Orange
Item 3	400	530	240	3	Yellow
Item 4	130	630	400	2	Green
Item 5	400	250	250	1	Blue
Item 6	130	400	160	2	Cyan

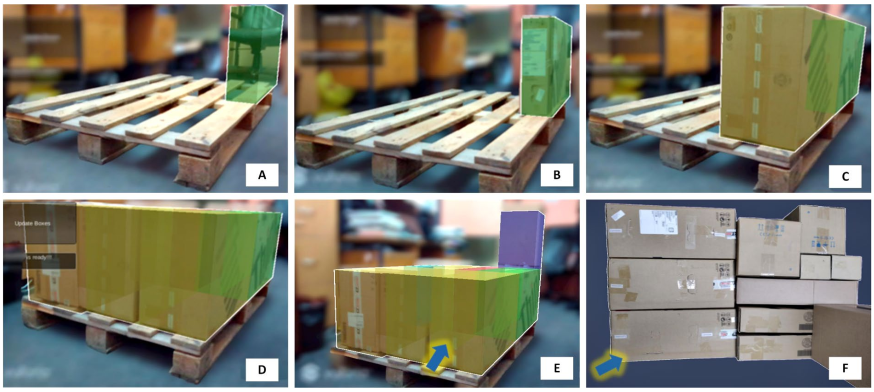


Figure 9
One example of the second test of an AR application

3.3 Discussion

By testing the developed applications in real working conditions, it can be said that AR can be successfully implemented in material handling processes in production plants, distribution centres, warehouses, etc. Further analysis can confirm that AR can be used as support to solve the PLP. In real working conditions, the execution time of the proposed algorithm for solving PLP is relatively short (several tens of milliseconds) [27] which provides the fast implementation of the algorithm into the AR application, where "step-by-step" movement occurs. This movement allows the warehouse worker to display one box at a time, then to pick it up and place it on the pallet (before displaying another box). For this purpose, the best solution is using smart glasses because in that case the worker has free hands to manipulate the real boxes. It should also be noted that the virtual boxes in the current version of the AR application are displayed in the corresponding colours, which implies the fact that the worker knows which virtual box (which colour) corresponds to which real box. To address the issue where the worker is unaware of the correspondence between the colors of virtual boxes and real boxes, it is possible to improve the application and automatically providing the worker with information about the real box after displaying the virtual box.

Live testing pointed out that the recognition angle of the desired marker is crucial to the app's functionality. One pallet marker was not enough to run the application smoothly, so two were created. Comparing these two markers, it can be seen that the pallet marker 1 is more often observed in real situations, so the upper surface is the dominant part of the marker. However, by placing the boxes on the pallet, they obscure the marker and the application is very prone to interruptions. For this reason, it was necessary to develop the pallet marker 2. The dominant part of the pallet marker 2 is the lateral side, so that the load boxes do not affect the loss of display. It was also observed that it is more difficult to recognize the marker the first time due to the viewing angle and to make an initial connection in this case. The positive thing is the preservation of the captured state of loaded boxes, so the application continues to work from the step where it stopped as soon as the marker is detected again.

One limitation in the current version of the AR application is the absence of feedback regarding whether the worker correctly positioned and oriented the actual box on the pallet. In the event of an error, which could occur randomly, subsequent steps may result in overlapping boxes, making it impossible to place the next box accurately. When such an error is detected, the worker would need to reset the system. Following this, they would have to revisit all previous steps, rectifying the position and/or orientation of the inaccurately placed box in the ongoing work.

Future research will be related to the improvement of the implemented algorithm for PLP solving. In this observed case, a planar problem is practically considered since all the boxes have one common dimension, that is, the 3D problem is reduced to 2D. In reality, many boxes may have different dimensions. In that case, it would be necessary to create a new optimality criterion, i.e. take into account other parameters, such as the mass of the boxes, the position of the centre of mass, orientation, etc. Furthermore, the desktop application for generating the loading order would have to be changed so that the newly formed QR code contains all the necessary information. In addition, the AR application will be enhanced with integrated feedback on whether the worker placed the actual box in the correct position and orientation.

It is extremely important to follow the described methodology when developing an AR environment regardless of specific conditions in a plant/warehouse. This development path consists of three steps: 1) user algorithm; 2) a desktop application for creating loading order and 3) an AR application. Depending on the case, only partial adjustments are allowed. These adjustments imply adaptation to the requirements of specific users in some of the mentioned steps.

Conclusions

The main goal of this paper is to provide insight into how AR can be used to support material handling processes. A realistic setup was prepared and experimental proof was confirmed by implementing the AR application with the integrated algorithm for the pallet loading process. This novel application includes two user applications: 1) a desktop application, and 2) an AR application for pallet loading.

Desktop application generates a loading order with a QR code, which contains all the necessary information for the successful execution of the algorithm. The proposed algorithm provides one way of solving PLP. The AR app scans this QR code, executes the algorithm in the background and visually show the worker how to load the boxes onto the pallet.

Experimental testing has shown that AR is competitive both in terms of time efficiency (several tens of ms) and error rates (thereby increasing production). Using the developed system facilitates and speeds up the process of shipping products from the warehouse to customers. In real working conditions, during loading boxes, there is a possible situation in which a worker places one box and then removes it from the pallet, realizing that a box of other dimensions fits better. Applying the proposed solution avoids this situation.

Of particular interest is the adaptability of the PLP algorithm for application in various scenarios, particularly in “pick and place” operations executed by industrial robots. In these situations, implementation of the algorithm would solely require integration during the programming phase of the robot's operation.

Augmented reality can be a viable option for pallet loading, especially where more information needs to be displayed. Future work will be related to the development of a new algorithm that allows overcoming existing limitations, in a way that provides the possibility of loading boxes of completely different dimensions, possibly also shapes. In addition, future research will be related to improvement of AR application in order to allow a feedback information.

References

[1] Balogh, A., Gyenge, B., Szeghegyi, Á., & Kozma, T. (2020) Advantages of simulating logistics processes. *Acta Polytechnica Hungarica*, *17*(1), 215-229, https://doi.org/10.12700/APH.17.1.2020.1.12

[2] Masood, T., & Egger, J. (2020) Adopting augmented reality in the age of industrial digitalisation. *Computers in Industry*, 115, 103112, https://doi.org/10.1016/j.compind.2019.07.002

[3] Plewan, T., Mättig, B., Kretschmer, V., & Rinkenauer, G. (2021) Exploring the benefits and limitations of augmented reality for palletization. *Applied Ergonomics*, 90, 103250, https://doi.org/10.1016/j.apergo.2020.103250

[4] Rejeb, A. (2019) The challenges of augmented reality in logistics: a systematic literature review. *WSN*, *134*(2), 281-311

[5] DHL, Ricoh, & UBiMAX. (2015, April) In *DHL Global Technology Conference 2015 - Augmented Reality in Logistics*

[6] Wang, W., Wang, F., Song, W., & Su, S. (2020) Application of augmented reality (AR) technologies in inhouse logistics. In *E3S Web of Conferences* (Vol. 145, p. 02018) EDP Sciences, https://doi.org/10.1051/e3sconf/202014502018

[7] Akbari, M., Ha, N., & Kok, S. (2022) A systematic review of AR/VR in operations and supply chain management: maturity, current trends and future directions. *Journal of Global Operations and Strategic Sourcing*, *15*(4), 534-565, https://doi.org/10.1108/JGOSS-09-2021-0078

[8] Gehring, M., & Mosler, P. (2022) Indoor Navigation with Augmented Reality and BIM: A Marker-Based Approach for Locating Logistics Areas on Construction Sites. In *Proceedings of 33. Forum Bauinformatik*

[9] Konstantinidis, F. K., Kansizoglou, I., Santavas, N., Mouroutsos, S. G., & Gasteratos, A. (2020) Marma: A mobile augmented reality maintenance assistant for fast-track repair procedures in the context of industry 4.0. *Machines*, *8*(4), 88, https://doi.org/10.3390/machines8040088

[10] del Amo, I. F., Erkoyuncu, J. A., Roy, R., Palmarini, R., & Onoufriou, D. (2018) A systematic review of Augmented Reality content-related techniques for knowledge transfer in maintenance applications. *Computers in Industry*, *103*, 47-71, https://doi.org/10.1016/j.compind.2018.08.007

[11] Porter, S. R., Marner, M. R., Smith, R. T., Zucco, J. E., & Thomas, B. H. (2010, October) Validating spatial augmented reality for interactive rapid prototyping. In *2010 IEEE International Symposium on Mixed and Augmented Reality* (pp. 265-266) IEEE, https://doi.org/10.1109/ISMAR.2010.5643599

[12] Reif, R., & Walch, D. (2008) Augmented & Virtual Reality applications in the field of logistics. *The Visual Computer*, *24*, 987-994, https://doi.org/10.1007/s00371-008-0271-7

[13] Schwerdtfeger, B., & Klinker, G. (2008, September) Supporting order picking with augmented reality. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality* (pp. 91-94) IEEE, https://doi.org/10.1109/ISMAR.2008.4637331

[14] Pettersson, M., & Stengård, M. (2015) The Impact of Augmented Reality Support in Warehouse Trucks. Linköping University, The Institute of Technology

[15] Martins, G. H., & Dell, R. F. (2008) Solving the pallet loading problem. *European Journal of Operational Research*, *184*(2), 429-440, https://doi.org/10.1016/j.ejor.2006.11.012

[16] Ancora, G., Palli, G., & Melchiorri, C. (2022, April) Combining Hybrid Genetic Algorithms and Feedforward Neural Networks for Pallet Loading in Real-World Applications. In *Human-Friendly Robotics 2021: HFR: 14th International Workshop on Human-Friendly Robotics* (pp. 1-14) Cham: Springer International Publishing

[17] Kocjan, W., & Holmström, K. (2008) Generating stable loading patterns for pallet loading problems. In *The Fifth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems CPAIOR08*

[18] Ha, H. T. H., & Nananukul, N. (2017) Air cargo optimization models for logistics forwarders. *Advanced Science Letters*, *23*(5), 4162-4167, https://doi.org/10.1166/asl.2017.8327

[19] Lel, V. T., Creighton, D., & Nahavandi, S. (2005, August) A heuristic algorithm for carton to pallet loading problem. In *INDIN'05. 2005 3rd IEEE International Conference on Industrial Informatics, 2005* (pp. 593-598) IEEE, https://doi.org/10.1109/INDIN.2005.1560443

[20] Bracht, E. C., de Queiroz, T. A., Schouery, R. C., & Miyazawa, F. K. (2016, August) Dynamic cargo stability in loading and transportation of containers. In *2016 IEEE International Conference on Automation Science and Engineering (CASE)* (pp. 227-232) IEEE, https://doi.org/10.1109/COASE.2016.7743385

[21] Beirão, J. F. (2009) Packing problems in industrial environments: Application to the expedition problem at INDASA. *Private Thesis*

[22] Schuster, M., Bormann, R., Steidl, D., Reynolds-Haertle, S., & Stilman, M. (2010) Stable stacking for the distributor's pallet packing problem. In *2010 IEEE International Conference on Intelligent Robots and Systems (IEEE/RSJ)* (pp. 3646-3651) IEEE, https://doi.org/10.1109/IROS.2010.5650217

[23] Gzara, F., Elhedhli, S., & Yildiz, B. C. (2020) The pallet loading problem: Three-dimensional bin packing with practical constraints. *European Journal of Operational Research*, *287*(3), 1062-1074, https://doi.org/10.1016/j.ejor.2020.04.053

[24] Techasarntikul, N., Ratsamee, P., Orlosky, J., Mashita, T., Uranishi, Y., Kiyokawa, K., & Takemura, H. (2020) Guidance and visualization of optimized packing solutions. *Journal of Information Processing*, *28*, 193-202, https://doi.org/10.2197/ipsjjip.28.193

[25] Reljić, V., Milenković, I., Dudić, S., Šulc, J., & Bajči, B. (2021) Augmented reality applications in industry 4.0 environment. *Applied Sciences*, *11*(12) 5592, https://doi.org/10.3390/app11125592

[26] Singh, M., Almasarwah, N., & Süer, G. (2019) A two-phase algorithm to solve a 3-dimensional pallet loading problem. *Procedia Manufacturing*, *39*, 1474-1481, https://doi.org/10.1016/j.promfg.2020.01.301

[27] Reljić, V., Dudić, S., Trišić, Ž., Jekić, B., Jurošević, V., & Milenković, I. (2022, February) Augmented Reality as a Support to Solve Pallet Loading Problem. In *International Conference on Remote Engineering and Virtual Instrumentation* (pp. 261-270) Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-17091-1

[28] VUZIX. (2022) Vuzix Blade Upgraded Smart Glasses. https://www.vuzix.com/products/vuzix-blade-smart-glasses-upgraded#/blade-technical-specs (accessed Jun. 12, 2022)

[29] VUZIX. (2022) VUZIX View Software Downloads. https://www.vuzix.com/pages/vuzix-view-software (accessed Jun. 12, 2022)