ACTA TECHNICA JAURINENSIS
Vol. 15, No. 1, pp. 1-6, 2022
10.14513/actatechjaur.00581

MLOps approach in the cloud-native data pipeline design

István Pölöskei[1,*]

[1]Adesso Hungary Kft.
Infopark sétány 1, 1117 Budapest, Hungary
*e-mail: istvan.poeloeskei@adesso.eu

Submitted: 25/12/2020 Accepted: 21/03/2021 Published online: 09/04/2021

Abstract: The data modeling process is challenging and involves hypotheses and trials. In the industry, a workflow has been constructed around data modeling. The offered modernized workflow expects to use of the cloud's full abilities as cloud-native services. For a flourishing big data project, the organization should have analytics and information-technological know-how. MLOps approach concentrates on the modeling, eliminating the personnel and technology gap in the deployment. In this article, the paradigm will be verified with a case-study in the context of composing a data pipeline in the cloud-native ecosystem. Based on the analysis, the considered strategy is the recommended way for data pipeline design.

Keywords: MLOps; cloud-native; data pipeline; machine learning

I. INTRODUCTION

Classical algorithms are apparent; they obey a clear logic in a human-readable manner. A machine learning model results from the training process by using unusual methods to reach the best result. The model evaluation might answer what the best result indicates. The way to the model is challenging, comprises assumptions and experiments.

The data and its structure has been evolving. In an enterprise, a workflow has been built around data modeling. This base helps mine the information from the data swiftly and efficiently, providing the organization's agility. The proposed modern workflow requires to use of the cloud's full capabilities. Planning a workflow according to the infrastructure, makes the operation more affordable and the implementation more powerful. Since the public cloud providers serve on-demand invoicing, the reserved resources should be connected to the running tasks [1].

II. CLOUD-NATIVE

1. Cloud benefits

The brand-new cloud technology innovations make computing more affordable and manageable than in the on-premised environment [2]. The training process of a deep learning model takes some time. It could occur that the local settings are not able to manipulate massive datasets. The training quality can often be efficiently increased by committing more resources, like attaching computation-intensive hyperparameter optimization measures [3].

2. Containers

Container technology is a state-of-the-art virtualization platform [4]. This approach presents an isolated, transportable bundle of the application with its dependencies. The currently applied (Docker) containers are originated from Linux primitives, implementing a process level separation [4].

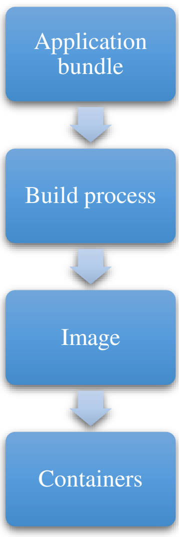


***Figure 1**. Container building process in Docker*

Docker containers are formed (**Fig. 1**) from layered immutable images based on their descriptor (Dockerfile) [5]. In the enterprise where containers are started and stopped dynamically by the runtime environment, it is advised to reduce the startup time; the container should be light-weight for better usability in the cloud-native circumstances [5]. In application development, the building process can be supported by assigned containers; the production container contains only the binaries with the runtime environment.

According to the 12 factors application principle [6], the conventional container-level application design must utilize the cloud-native's potential opportunities. The terms of the pattern, like stateless processes, are the primary entry points for cloud-native use-cases.

3. Cloud-native benefits

The Cloud Native Computing Foundation was established in 2015 to promote innovative container technologies [7]. According to their interpretation: the cloud-native is the set of technologies for providing scalable applications in the cloud, including ideas like micro services, DevOps, and agile [8]. Furthermore, cloud-native is more than architecture; it is a radical change in providing and developing a service. The cloud-native applications use the cloud's traits with its full potential.

Containers and micro services are the fundamental characteristics of cloud-native applications [8]. The functionalities are broken down into tiny autonomous bits, including an API-based interface. The pieces can be operating individually; each part has its deployment and version. The containers' concept proves each service process-wise sovereignty and lessens the entire system's complexity [8]. By wholly practicing cloud principles, it is feasible to serve compliant and scalable software.

Implementation's rate is vital. Fleet delivering features for the business is a strategic advantage. Introduction with a DevOps mindset and Agile, state-of-the-art deployment system can accomplish the business's demands more efficiently than before [8].

By the usage of the cloud-native best practices, the suggested architecture makes the application reliable and fault-tolerant. The framework replaces the broken element with a fresh instance, operating like a self-healing system. It preserves the functionalities by allocating more instances process-wise; the resources' dynamic usage is possible [9].

The state-of-the-art cloud-native solution is the Kubernetes, presented by Google [10]. In the Kubernetes, the containers are merged into POD, the atomic primitive in the Kubernetes ecosystem. Through the POD approach, the lifecycle of the processes can be executed through the orchestrator.

III. DATA PIPELINE

1. Pipeline

The data pipeline is an adhesive code between the data sources and machine learning algorithms [11]. The code segment, as mentioned earlier, is continuously growing because the data composition strongly influences it. The data needs to be reconstructed and conveyed before the training commences. A workflow is a feasible solution, gathering and standardizing the required data, for providing high-level input for the machine learning algorithms. Since the entire process is divided into numerous levels, parallelized computing procedures are used [12].

The pipeline workflow design is not an uncomplicated task. Realizing the whole data processing requires some boilerplate codes. Workflow builders standardize the orchestration of data pipelines by using workflow engines and frameworks.

2. MLOps

Some prototypes are not deployed in the machine learning use-cases in the production situation because they may have concerns with the more prominent data-load or scaling, scheduling, or the integration to the data sources. For a successful big data project, the organization should have analytics and infrastructure expertise; they need to have a comprehensive plan about deploying the model and the production dataset's training cycles.

As a bridge linking the data scientist and software engineer, a unique responsible role is a viable solution to tackle machine learning models to production as soon as possible. This specialist needs to have a global picture of the model lifecycle, like organizing each stage's main interactions [13]. Infrastructural coordination is necessary for having efficient architectural usage and automatized, high-quality input data from diversified sources. It makes the deployment cycles of the models regular.

The role of MLOps as a Machine Learning branch of the DevOps is accepted in the machine learning domain [13]. It is a compound of operation, machine learning, and business understanding. All of them are distinct competencies. The DevOps methodology is for the speed and effectiveness of development and deployment, providing the project's adaptability in line with the agile [13]. MLOps is the same approach but focuses on the model development for eliminating the personnel and technology gap in the deployment. It concentrates on model formulation, evaluation, deployment, and monitoring. The new cloud-based solutions promote this approach vigorously.

3. Data platform

The data solution performs some steps in the data use-case, like the data-preparation, training, model verification, feature selection [12]. Achieving a data solution is challenging because it involves some perspectives like mathematics, informatics, and business [14]. It was stimulating to present a solution for data scientists in the last years, granting the possibility to deploy their model without code modifications in a productive environment. The complexity should be diminished through the modern data frameworks, giving a portable, scalable, and efficient infrastructure for the data experts. Managing the existing, same locally tested code, in the cloud environment, without an extensive perception of the cluster is an immense business value. The team can focus on the business, not on the cloud-native infrastructure or deployment. The effort to the productive data solution has been simplified, but the actual performance is to build an integrated fail-tolerant data infrastructure.

If the model has been constructed, the business should apply it. Conventional components also support the deployment section, like the TensorFlow Serving library [15]. The implementation unit or product support should permanently observe the usage of the model. Sometimes, the prototype should be aligned with the current data structure or obscure data patterns (**Fig. 2**). The model should be retrained for being up-to-date.

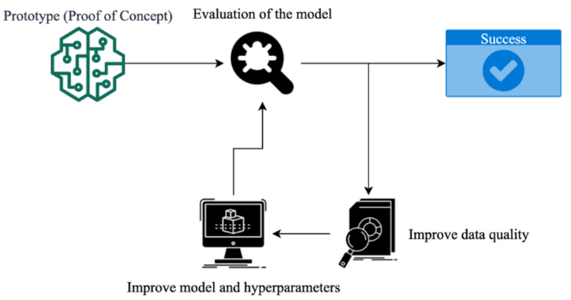


***Figure 2.** Prototyping process*

A CI/CD pipeline like a data platform is an added advantage because it accommodates the rapid prototyping process. Data science can explore their new idea directly in production. The new model can be deployed smoothly through some automatisms, producing business value as soon as possible in the continuously varying business environment. According to the MLOps principle, the Machine Learning tasks and artifacts can be combined with mainstream DevOps instruments [13].

4. Cloud-native pipelines

Kubernetes and TensorFlow are the prime open-source brands that were started by Google [16]. By them, some modern infrastructure has been established for giving improved solutions for novel problems. As their union, the KubeFlow is a state-of-the-art toolkit in a self-managed cloud-native ecosystem [17].

Kubeflow is an open-source cloud-native platform for developing data pipelines and workflows. It is a potent mixture of pipeline approach with cloud-native foundation [17]. It clarifies and normalizes the whole machine learning sequence in the cloud.

Machine learning and deep learning should be used with active data-processing in the cloud to determine contemporary data problems. According to the use-case discussed earlier, the KubeFlow may work as a fundamental part of its cloud-native strategy.

5. Workflow as graph

A pipeline can operate on the cluster. This procedure is cost-effective because the scaling-out is more affordable than the scaling-up [18]. Since the workflow is composed as a DAG (directed acyclic graph), each job can be executed as a DAG node [19].

Each task can be grouped by components (like python functions). The components, like regular atomic elements, can interact through their inputs and outputs. The low-level segments are not responsible for the cloud-naiveness. In the framework, the orchestrator is responsible for the supervision of the elements. The data specialists must not implement the infrastructure; they can only focus on their business processes [19].

Other pipelines can re-use each operator or component. By them, the entire pipeline can be controlled on the tasks level. The flow is visualized graphically, and each part has its log-stream. That facilitates the bug finding and fixing in the production.

For the coding, the KubeFlow accommodates jupyter notebooks as a standard coding context [17]. It is a popular conventional solution in the machine-learning society, strongly supports feature engineering and model fabricating. Through the built-in visualization, the data scientist can adequately decipher the information of the data. By a jupyter service provision, the most popular dependencies and libraries can be included by default; the resource quotas can also be configured. The configured runtimes can use GPU assistance as well (**Fig. 3**).

6. Infrastructure based on the pipeline

Establishing the infrastructure on the POD level of the nodes is complex. New frameworks, like Kubeflow, allow constructing pipelines at the code level without explaining them in any descriptor files. The script can control the whole infrastructure; the designed program can be scaled based on the load in a portable style [16]. This is propitious for machine learning use-cases when the training cycles are more resource-demanding than the prediction use-cases. The billing is based on consumption in the cloud; the corporation should not keep any additional resources when it is not necessary.

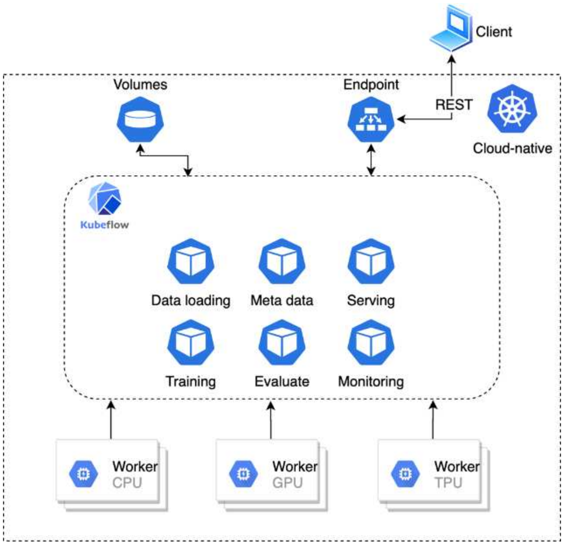


Figure 3. *Kubeflow landscape*

The critical advantage of cloud-native infrastructure, the built-in features for the provision of the services. ReplicaSets, monitoring, heartbeat. All these factors and patterns are primary for consistent and fail-tolerant infrastructure [10]. The abstraction and standardization of the frameworks enhance the security and the overall maintainability.

7. GPU based pipelines

The data pipelines in the industry expect regular training cycles. This is resource consuming, but this resource should be granted only on demand. Cloud computing is a valid option for this use-case.

The workflows are compute-intensive but present a choice for distributed computing. Deep learning is a potent use-case of GPU programming [20]. By the native integration of GPU, the training can be accomplished more adequately without complicated GPU scripts.

8. TFX

Tensorflow is the state-of-the-art deep learning and computational framework [21]. Tensorflow Extended (TFX) library and specification help the Tensorflow model implementations in the production. A proper pipeline can be built around the deep learning model through its elements, like an end-to-end solution (**Fig. 4**). It is providing a high-performance application in the cloud-native environment.

Through Tensorflow Extended, a standardized TensorFlow pipeline can be created for particular orchestrators like Airflow and Kubeflow [21]. Having a multi-architectural approach has an advantage because it cannot be guaranteed that the project can obey the cloud-native principles. Integration and workflow management is a necessity, but the target system does not always use Kubernetes. In that case, traditional alternatives (like Airflow) must be used.

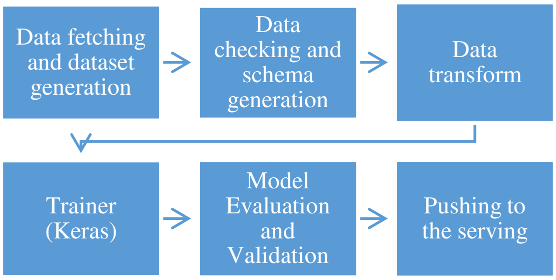


Figure 4. *Tensorflow Extended pipeline*

IV. Case study: data platform in manufacturing

1. Situation faced

Data science's fundamental duty is the observation and optimization of business processes. This journey is like a discovery of the production. The base-line situation is the progressed scrapping rate in manufacturing; the selected data scientist should investigate and analyze the root causes.

For the production-ready data-analytics result, the data team wants to constitute a data platform. The platform should utilize the public cloud approach; natively, it needs GPU support besides high availability. The cloud infrastructure should be sized according to the actual load; more resources should be involved in the exacting training steps. This architecture should support the data science department's standard data tools like Jupyter notebooks and python with deep learning support. The integration layer should be capable of loading the required data sources in the organization without any difficulty.

2. Actions taken

The MLOps approach adds the DevOps toolstack and cloud-native philosophy to the data workloads. An MLOps engineer represents the way how the model should be applied and deployed. Some orchestration difficulties have already been resolved through the introduction of KubeFlow in the cloud-native environment (**Fig. 5**).

The chosen strategy was shifting from the simple to more complex features. The base-line model can be improved for better accuracy through hyperparameter optimization and feature engineering. With a simplistic proof of concept, the data platform architecture can be validated for usability. Based on the evaluation outcome, the model's overall performance has been improved and verified by mainstream measurements (like accuracy, F1-score, etc.) [22]. If the model's correctness meets the requirements, the deployment begins.

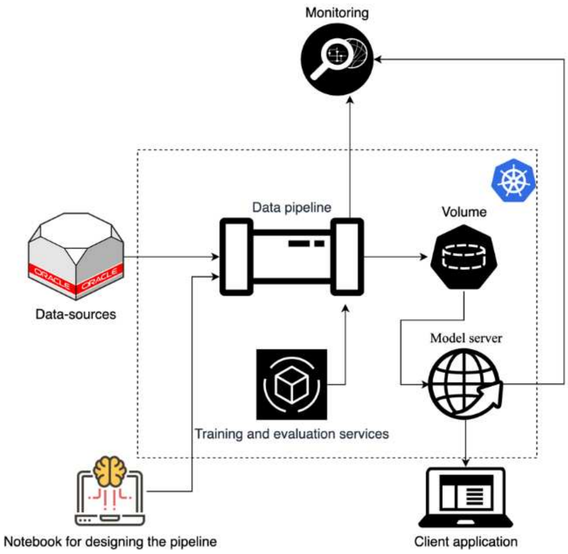


Figure 1. *Realized pipeline. Cloud services have processed the input data. The created model has been published*

Through the prototype has been adapted several times, a version controlling is required. The model selection is also essential because the existing models could have more excellent performance than the new release. A basis artifact library can serve the previously built best model.

The workflow can be altered; it triggers container level cloud-native fundamental changes. The produced components can manage the complete training and evaluation workflow. Monitoring and logging operations are assigned to primary activities. The pipeline steps in details:

- Establishing the pipeline cloud-native architecture based on the repository
- Data pre-processing and data digestion
- Machine learning (classical or deep learning approach) in dedicated PODs
- Evaluation and model selection
- Deployment to the REST endpoint

The data scientists have built a promising offline model. The deployment and extension of the pipeline remain challenging. It takes time to reach the production level.

3. Results achieved

It is justified to set up a pilot project for a reduced scope for experiment purposes. The immediate feedback can solidify the concept shortly (**Fig. 6**). A production-level prototype verifies the theoretical approach and achieves the management's acceptance. The testing charge is lower in that case, and the design can be examined without higher production risks.

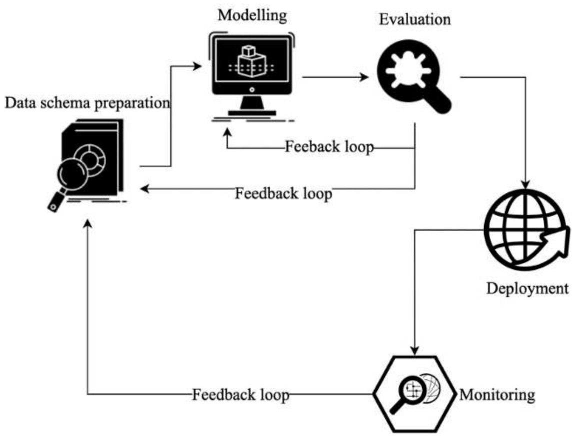


Figure 2. *The quick feedback loop in production*

By acceptance of the MLOps mindset, the research loops become more expeditious. Through continuous integration, the pipeline and model can be arranged rapidly in production. Model optimization, like hyperparameters harmonizing, claims some trials and failures. The prompt reactions yield real business content in production.

The MLOps gives the dynamic construction of event-based workflows, with the native support of distributed data processing frameworks, implementing scale-out opportunities on the top of the cloud-native design.

4. Lessons learned

At the first iteration, the pipeline was composed of Jenkins jobs. It was complicated and not resembling the data science mindset. Between the prototyping and deployment, there was a considerable gap. KubeFlow with Tensorflow Extended resolved this concern. The framework accommodates a firm context for modeling without narrowing the data experts' opportunities. The utilized technology grants high-level automatization possibilities by the usage of elements that implement specific subsequent responsibilities. The Tensorflow Extended represents a current, comprehensive ML pipeline, from implementing the new business scenarios to managing the traditional machine learning activities.

V. Conclusion

Some designs never reach production in the machine learning area. A pipeline, like a structure, supports loading and digesting the data for machine learning procedures. The scalability is crucial because the training expenses are soaring in the latest neural network-based architectures; the complete workflow should be performed in an on-demand manner in the cloud. Cloud-native seems to be a convenient and future-proof solution. Based on the study, the examined strategy is the advised way for data pipeline configuration.

Acknowledgement

The publishing of this paper was supported by XY.

Author Contributions

I. Pölöskei: Conceptualization, Experiments, Writing

Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Gao, L. Wang, R. Teti, D. Dornfeld, S. Kumara, M. Mori, M. Helu, Cloud-enabled prognosis for manufacturing. CIRP Annals - Manufacturing Technology 64 (2) (2015) pp. 749-772. https://doi.org/10.1016/j.cirp.2015.05.011

[2] D. A. Tamburri, M. Miglierina, E. Di Nitto, Cloud applications monitoring: An industrial study. Information and Software Technology 127 (2020) 106376. https://doi.org/10.1016/j.infsof.2020.106376

[3] L. Franceschi, M. Donini, P. Frasconi, M. Pontil, On hyperparameter optimization in learning systems. 5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings.

[4] M. De Benedictis, A. Lioy, Integrity verification of Docker containers for a light-weight cloud environment. Future Generation Computer Systems 97 (2019) pp. 236-246. https://doi.org/10.1016/j.future.2019.02.026

[5] A. Martin, S. Raponi, T. Combe, R. Di Pietro, Docker ecosystem – Vulnerability Analysis. Computer Communications 122 (2018) pp. 30-43. https://doi.org/10.1016/j.comcom.2018.03.011

[6] M. Mohamed, R. Engel, A. Warke, S. Berman, H. Ludwig, Extensible persistence as a service for containers. Future Generation Computer Systems 97 (2019) pp. 10-20. https://doi.org/10.1016/j.future.2018.12.015

[7] D. Gannon, R. Barga, N. Sundaresan, Cloud-Native Applications. IEEE Cloud Computing, 4 (2017) pp. 16-21. https://doi.ieeecomputersociety.org/10.1109/MCC.2017.4250939

[8] S. Peltonen, L. Mezzalira, D. Taibi, Motivations, Benefits, and Issues for Adopting Micro-Frontends: A Multivocal Literature Review. https://arxiv.org/abs/2007.00293

[9] M. Malawski, A. Gajek, A. Zima, B. Balis, K. Figiela, Serverless execution of scientific workflows: Experiments with HyperFlow, AWS Lambda and Google Cloud Functions. Future Generation Computer Systems 110 (2020) pp. 502-514. https://doi.org/10.1016/j.future.2017.10.029

[10] S. Kho Lin, U. Altaf, G. Jayaputera, J. Li, D. Marques, D. Meggyesy, S. Sarwar, S. Sharma, W. Voorsluys, R. Sinnott, A. Novak, V. Nguyen, K. Pash, Auto-Scaling a Defence Application across the Cloud Using Docker and Kubernetes. Proceedings - 11th IEEE/ACM International Conference on Utility and Cloud Computing Companion, UCC Companion 2018. https://doi.org/10.1109/UCC-Companion.2018.00076

[11] D. Wu, L. Zhu, X. Xu, S. Sakr, D. Sun, Q. Lu, Building pipelines for heterogeneous execution environments for big data processing. IEEE Software, 33 (2) (2016) pp. 60-67. https://doi.org/10.1109/MS.2016.35

[12] Z. Peng, Stocks Analysis and Prediction Using Big Data Analytics. Proceedings - 2019 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2019. https://doi.org/10.1109/ICITBS.2019.00081

[13] I. Karamitsos, S. Albarhami, C. Apostolopoulos, Applying DevOps Practices of Continuous Automation for Machine Learning. Information 11 (7) (2020) 363. https://doi.org/10.3390/info11070363

[14] J. S. Saltz, S. Yilmazel, O. Yilmazel, Not all software engineers can become good data engineers. Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016. https://doi.org/10.1109/BigData.2016.7840939

[15] D. Baylor, K. Haas, K. Katsiapis, S. Leong, R. Liu, C. Menwald, M. Trott, H. Miao, M. Zinkevich, N. Polyzotis, Continuous training for production ML in the tensorflow extended (TFX) platform. Proceedings of the 2019 USENIX Conference on Operational Machine Learning, OpML 2019. https://www.usenix.org/conference/opml19/presentation/baylor

[16] Google, Kubeflow [cited 2020-12-25]. https://www.kubeflow.org/

[17] Z. Li, R. Chard, L. Ward, K. Chard, T. J. Skluzacek, Y. Babuji, A. Woodard, S. Tuecke, B. Blaiszik, M. J. Franklin, I. Foster, DLHub: Simplifying publication, discovery, and use of machine learning models in science. Journal of Parallel and Distributed Computing 147 (2021) pp. 64-76. https://doi.org/10.1016/j.jpdc.2020.08.006

[18] C. Avci Salma, B. Tekinerdogan, I. N. Athanasiadis, Domain-Driven Design of Big

Data Systems Based on a Reference Architecture. Software Architecture for Big Data and the Cloud (2017) pp. 49-68. https://doi.org/10.1016/b978-0-12-805467-3.00004-1

[19] R. Mitchell, L. Pottier, S. Jacobs, R. F. Da Silva, M. Rynge, K. Vahi, & E. Deelman, Exploration of Workflow Management Systems Emerging Features from Users Perspectives. Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019. https://doi.org/10.1109/BigData47090.2019.9005494

[20] F. Rouzbeh, P. Griffin, A. Grama, M. Adibuzzaman, Collaborative Cloud Computing Framework for Health Data with Open Source Technologies. https://doi.org/10.1145/3388440.3412460

[21] Google, Tensorflow [cited 2020-12-25]. https://www.tensorflow.org/

[22] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, P. Pławiak, A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine 179 (2019) 104992. https://doi.org/10.1016/j.cmpb.2019.104992

ACTA TECHNICA JAURINENSIS
Vol. 15, No. 1, pp. 7–14, 2022
10.14513/actatechjaur.00639

Investigating the throughput performance of the MPT-GRE network layer multipath library in emulated WAN environment

Szabolcs Szilágyi[1,*], Imre Bordán[1]

[1]University of Debrecen, Faculty of Informatics
26 Kassai Way, 4028, Debrecen, Hungary
*e-mail: szilagyi.szabolcs@inf.unideb.hu

Submitted: 10/07/2021; Accepted: 27/08/2021; Published online: 18/10/2021

Abstract: Nowadays there is a growing demand for a much faster and more secure communication without borders through the internet, which is provoking more and more both network designers and manufacturers of communication devices. Thanks to the BYOD trend, our communication devices can be used at work. They generally have several built-in network interfaces (e.g. Ethernet, Wi-Fi, 4G). Theoretically, using interface/connection in parallel, we could speed up data transmission, and thus communication, by aggregating the channel capabilities of the interfaces. On the other hand, we could make data transmission more reliable by applying redundancy to the system. Unfortunately, traditional IP-based communications do not allow the use of parallel interfaces in a given communication session, leaving the hardware capabilities of our communications devices virtually untapped. To address this issue, we have developed a multipath communication solution called MPT-GRE, which we have already tested in several laboratory environments. The measurement results were published in our previous articles. In this paper we are going to test it in a much more realistic environment, using the Dummynet WAN emulation software. The measurement results confirmed that the MPT-GRE multipath solution is able to aggregate the performance of physical connections efficiently in the emulated Fast Ethernet IPv4 WAN environment as well.

Keywords: MPT-GRE; multipath communication; Dummynet; throughput; WAN Emulator

I. INTRODUCTION

This paper is an extended version of our former conference paper [1].

Multipath communication technologies are one of the hot research topics nowadays. What better proof of this than Apple and Cisco integrating MPTCP[1], considered as the flagship of multipath technologies, into their operating systems. With the help of multipath communication, we can increase throughput, while also employing redundant data paths.

In our earlier publications (see e.g. [2], [3], [4], [5], [6], [7], [8], [9]) we have presented a multipath communication technology (MPT-GRE[2] [10]) developed by our research group, which we have built on the standardized GRE-in-UDP tunneling technology[3]. The layered architecture of the MPT-GRE can be seen in **Fig. 1**.

[1]The MPTCP Project official website: https://www.multipath-tcp.org/
[2]MPT-GRE is a MultipPath Technology developed by our research group at the University of Debrecen, Hungary. Its operating principle is based on the Generic Routing Encapsulation in UDP standard. Practically, MPT-GRE is a multipath extension of the GRE-in-UDP standard described in RFC8086.
[3]GRE-in-UDP Encapsulation standard: https://tools.ietf.org/html/rfc8086
[4]The MPT-GRE Project official website: https://irh.inf.unideb.hu/˜szilagyi/index.php/en/mpt/

The MPT-GRE[4] software reads the input packet (IPv4 or IPv6 packet) from the tunnel interface at the sender site. This packet is encapsulated into a new UDP segment, and it is sent out to a path chosen from the multiple paths possibilities. At the receiver site the header of the incoming UDP segment is de-encapsulated, and the data (which is the original packet coming from the sender's tunnel interface) is transmitted to the tunnel interface of the receiver host. The (logical) connection between the tunnel interfaces of the peers is a direct, point-to-point connection.

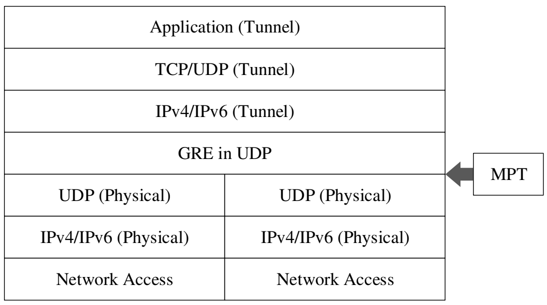


Figure 1. Layered architecture of the MPT-GRE.

We have examined its effectiveness with the help of numerous scenarios in our test environment, comparing results with MPTCP as a reference. All of the scenarios have showed that our MPT-GRE solution is capable of efficient path-aggregation both in Fast Ethernet and Gigabit Ethernet IPv4/IPv6 environments. The testing environments we have used previously (see **Fig. 2**) can be considered ideal in the sense that they did not contain any network environment parameters that could negatively affect network performance (e.g. delay, jitter, packet-loss). For this reason, we find it important to further examine the effectiveness of MPT-GRE in a more realistic environment (see **Fig. 3**).

Using the internet can be a given in case of single-path communication. However, to be able to test multipath systems, the presence of dual-home technology is essential, i.e. we need to have multiple ISP connections available.

Network simulation aims to replicate the key parameters of desired network environments with the help of mathematical models, with greater or lesser success.

The essence of network emulation is to replicate real network behaviors. Two main types exist:

- hardware realization (see **Fig. 4**),
- software implementation.

The first contains advanced technological solutions, but it is in turn quite a costly method. The latter is not always capable of providing a reliable and precise test environment, but it is cost-effective. Some examples of network emulator software: Dummy Cloud[5], Dummynet[6], NETEM[7], NIST Net[8], SoftPerfect Connection Emulator[9], WANem[10].

Given that a hardware implementation of a WAN emulator sufficient for our goals would be around 6000 EUR + VAT[11], after having reviewed the software solutions, our choice was Dummynet.

The Dummynet WAN emulator was developed in 2010 at the University of Pisa, and later got integrated into the FreeBSD operating system[12]. It provides a suitable framework for testing multipath solutions, enabling the setup of packet-delay, jitter and packet-loss network parameters [11]. It also has good documentation, including numerous code examples[13]:

When wanting to test a newly developed networking software in a realistic environment, we practically have three possibilities:

- internet,
- network simulation,
- network emulation.

II. MEASUREMENT ENVIRONMENT

To perform our measurements, we created two types of environments, namely a dual-path Fast Ethernet IPv4 and a dual-path Gigabit Ethernet IPv4 WAN

[5] Dummy Cloud official website: http://www.dummycloud.com/
[6] Dummynet Project official website: http://info.iet.unipi.it/ luigi/dummynet/
[7] NetEm's manual page: https://man7.org/linux/man-pages/man8/tc-netem.8.html
[8] NIST Net home page: https://www-x.antd.nist.gov/nistnet/
[9] SoftPerfect Connection Emulator: https://www.softperfect.com/products/connectionemulator/
[10] WANem official web page: http://wanem.sourceforge.net/
[11] Linktropy Mini-G's price at June 2021: https://www.digital-hands.eu/products/apposite-network-emulation/linktropy-mini-g/
[12] FreeBSD Manual Pages: https://www.freebsd.org/cgi/man.cgi?dummynet
[13] Using Dummynet in FreeBSD: http://noahdavids.org/self_published/using_dummynet.html

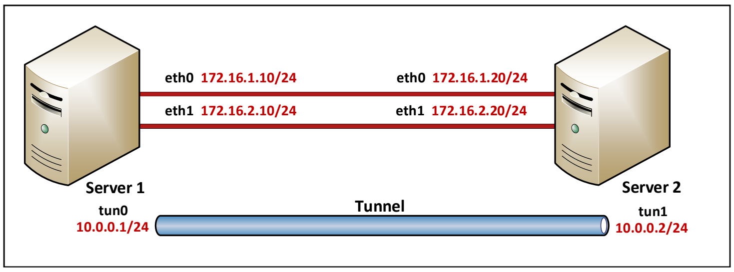


Figure 2. Our previous "ideal" testbed.

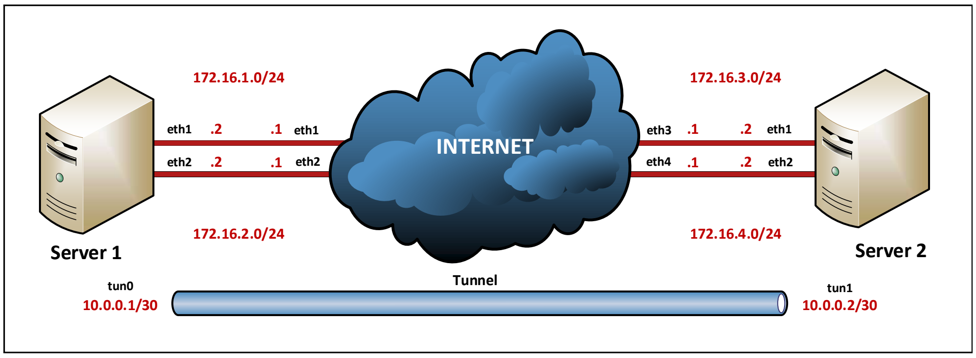


Figure 3. A real network testbed example.

emulated measurement environment (see **Fig. 5**). We downloaded a 1 GB file from the fileserver on the left onto the server on the right. Network parameters were controlled on the intermediate server that had Dummynet installed on the kernel level.

All three machines were running the Linux Ubuntu operating system. We examined the effect of packet-delay, jitter and packet-loss on file download speed, download time, and CPU performance. Bash and Python scripts – available on our website[14] – were used to automate the measurement process. We repeated each series of measurements ten times.

III. MEASUREMENT RESULTS

1. Measurement results in dual-path Fast Ethernet IPv4 WAN emulated environment

First, we checked how packet-delay affected download speed (see **Fig. 6**).

We gradually increased delay values on a scale of 0-190 ms on the first path using Dummynet. Everything proved to be stable until 100 ms. Above 100 ms, we experienced a continuous decrease in file download speeds. Using the 190 ms delay value, the download speed decreased to 107 Mb/s, while download time

[14]Our test scripts can be downloaded from: https://nas01.inf.unideb.hu/share.cgi?ssid=03CsniS

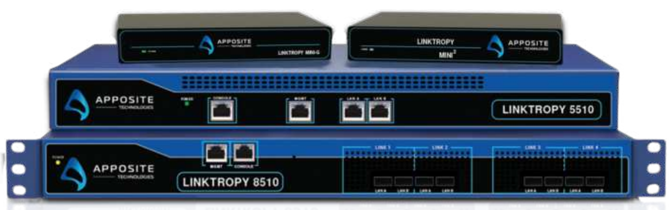


Figure 4. Hardware-based network emulators (Source: https://www.apposite-tech.com).

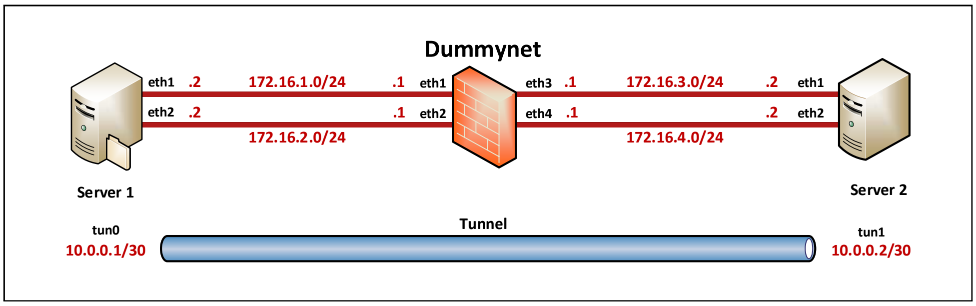


Figure 5. Our new measurement testbed with Dummynet.

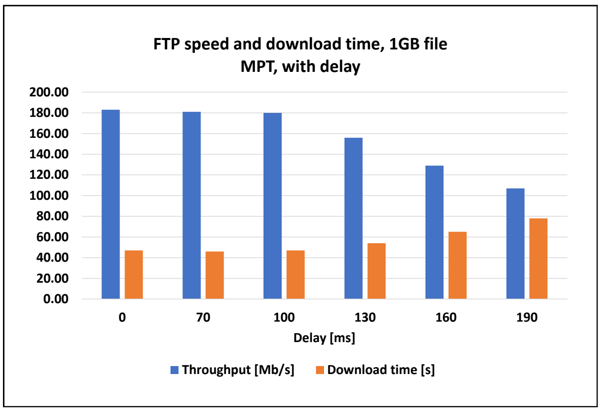


Figure 6
The effect of the delay on the FTP throughput and download time.

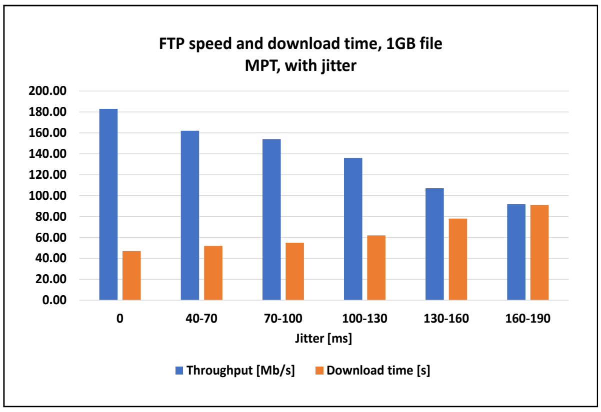


Figure 7
The effect of the jitter on the FTP throughput and download time.

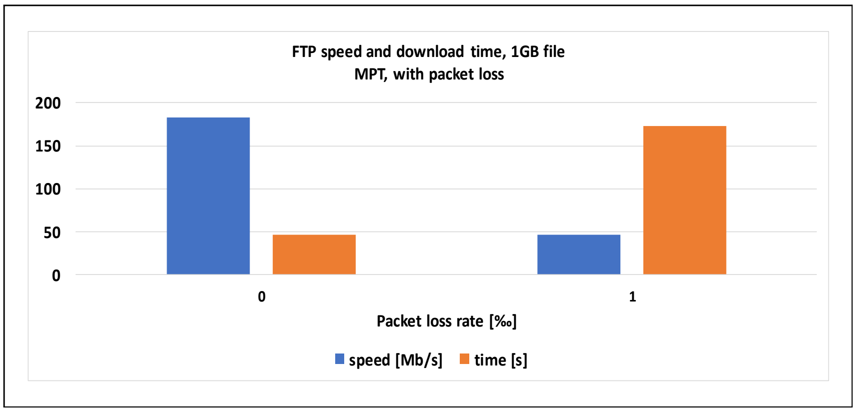


Figure 8. The effect of the packet loss on the FTP throughput and download time.

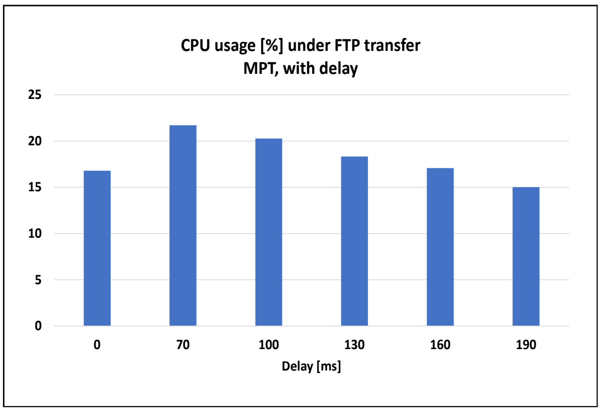


Figure 9
The effect of the delay on the CPU usage on Server 2.

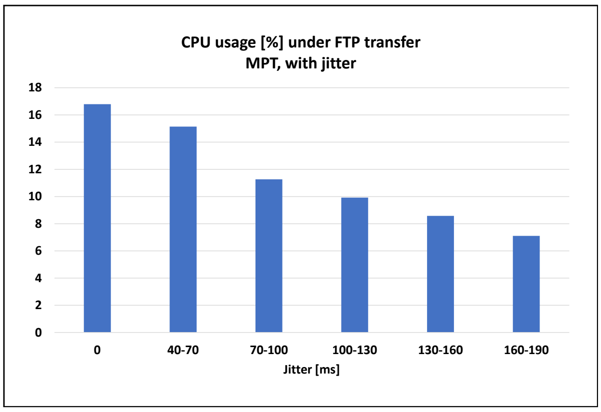


Figure 10
The effect of the jitter on the CPU usage on Server 2.

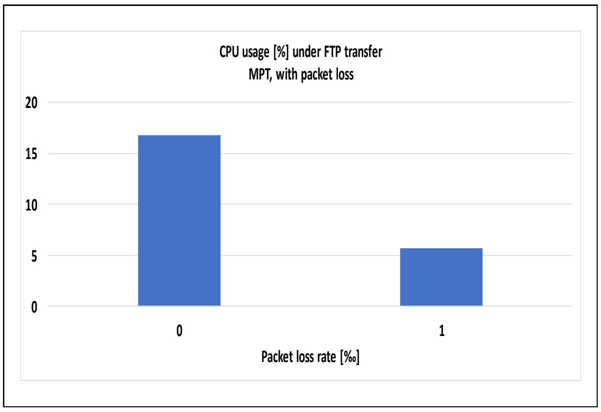


Figure 11
The effect of the packet loss on the CPU usage in case of MPT-GRE.

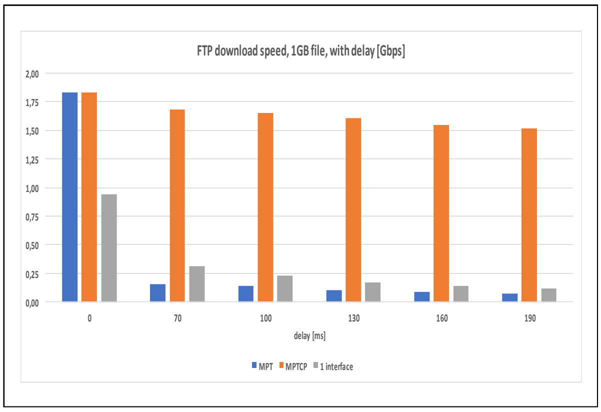


Figure 12
The effect of the delay on the FTP throughput in case of MPT-GRE and MPTCP.

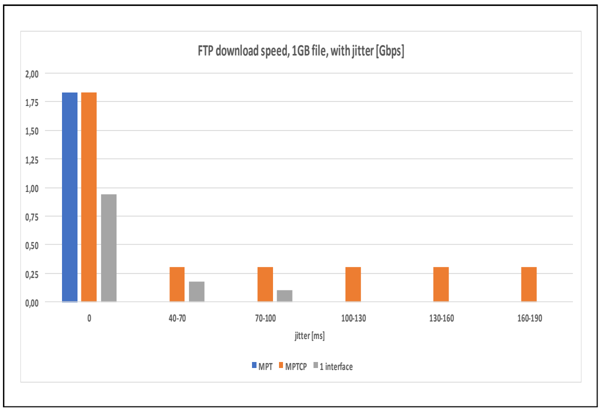


Figure 13
The effect of the jitter on the FTP throughput in case of MPT-GRE and MPTCP.

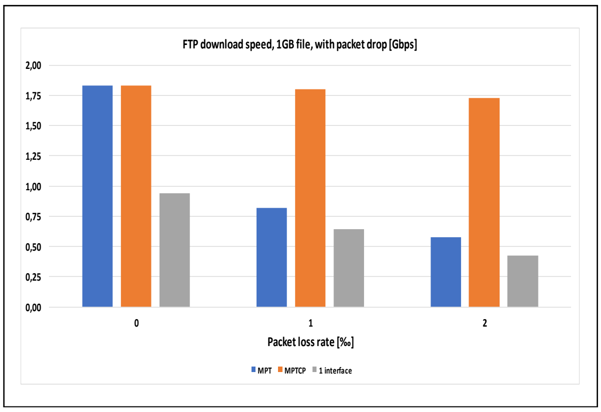


Figure 14
The effect of the packet loss on the FTP throughput in case of MPT-GRE and MPTCP.

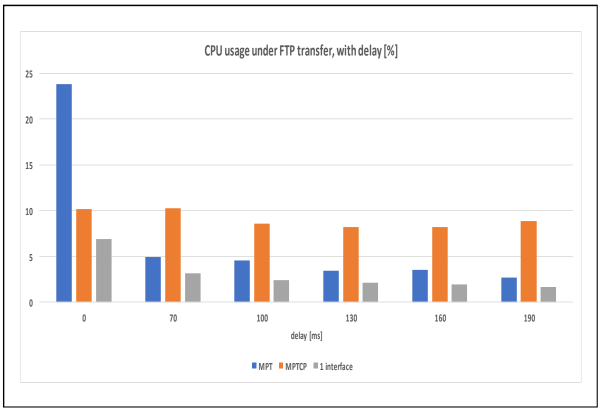


Figure 15
The effect of the delay on the CPU usage in case of MPT-GRE and MPTCP.

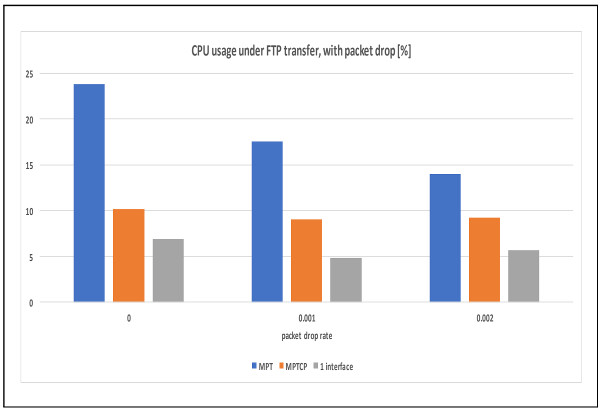


Figure 16
The effect of the packet loss on the CPU usage in case of MPT-GRE and MPTCP.

increased from 47 seconds to 78 seconds

A similar effect could be experienced in the case of increasing jitter values on a single path (see **Fig. 7**). With a 160-190 ms delay fluctuation, the download speed practically decreased by half, while the download time doubled.

Applying even a minimal data-loss rate (1 ‰), we witnessed a drastic performance decline (see **Fig. 8**). The download speed fall to a quarter, while the download time quadrupled. Therefore, we did not experiment with further data-loss rate values.

Regarding the effect of packet-delay on CPU performance, we did not experience significant fluctuation (see **Fig. 9**). CPU utilization hovered between 15-22% in every case.

Introducing jitter however, had noticeable effects on CPU utilization (see **Fig. 10**). With higher jitter values, we experienced a drop in CPU utilization.

While examining CPU loads, the effect of packet-loss also proved to be drastic (see **Fig. 11**). Using a data-loss rate of 1 ‰, utilization dropped from 17 to 5.7 percent.

We also carried out further measurements, mixing the parameters of the different paths. E.g. using only delay on one path, while using only jitter on the other. These scenarios brought similar results as well.

2. Measurement results in dual-path Gigabit Ethernet IPv4 WAN emulated environment

We also extended our measurements to a Gigabit Ethernet emulated WAN environment, examining the same network parameters as for the Fast Ethernet one. Unfortunately, we found from the outset that, for some reason, MPT is unable to function effectively in such a medium. A minimum (1 ms) delay also reduced the file download speed to less than 1Gb/s, and when using jitter, the transfer rate was a few kb/s. For this reason, we installed MPTCP, which was used as a reference for our measurement series. In the following, we make a performance comparison of the two multipath communication solutions.

In the first round, we examined the effect of the delay on the file download speed, indicating the reference value measured in single path communication session in each case. As shown in **Fig. 12**, in a delay-free environment, both environments effectively summed the available route capacity. Gradually increasing the rate of latency, MPT performance dropped below 250 Mbps, while MPTCP performed excellently.

With the use of jitter (see **Fig. 13**), the download speed of the MPT was reduced to 400 kbps already in the first step (applying 40-70 ms jitter), therefore we didn't even try any more measurements. On the other hand, MPTCP was able to produce throughput around 260 Mbps.

As for the effect of the data loss rate (see **Fig. 14**), even a 1‰ rate drastically degrades MPT's performance. In fact, it achieves almost the performance of a single-path environment.

Based on our measurements, it can be stated that the delay itself has a minimal effect on CPU performance (see **Fig. 15**). Using a delay of 70-190 ms, MPT performed slightly better than MPTCP. In both cases, the CPU utilization rate is below 10%.

However, the rate of data loss has a more significant effect on MPT's CPU utilization (see **Fig. 16**). Applying a data loss rate of 1‰, the CPU utilization of the MPT is 18%, while that of the MPTCP is around 9%.

IV. CONCLUSION

In our current paper, we extended the performance-analysis of our own multipath solution, MPT-GRE, using an emulated WAN environment. Firstly, we examined the effect of different network parameters, like e.g. packet-delay, jitter and data-loss rate on file download speed, download time, and CPU utilization in dual-path Fast Ethernet IPv4 WAN emulated environment. The worst performance we experienced was with the application of the 1 ‰ packet-loss rate. We further extended our measurements to a Gigabit Ethernet environment. In this case, the MPT performed very poorly, even with the introduction of a minimum delay, jitter or data loss rate. Because of these, we extended our measurements to the reference MPTCP multipath environment (see e.g. [12], [13], [14], [15], [16]). In his case, too, the results of the measurements were much weaker than expected, but he definitely performed better than the MPT. As a self-check, MPT- and MPTCP-free baseline measurements were performed in a single-path environment. Even so, very poor results were obtained. Finally, we concluded that the Dummynet WAN emulation environment may not be the most suitable solution for testing multipath Gigabit Ethernet environments and that MPT-GRE needs to be further optimized for more efficient performance.

Among our future goals we plan to improve our MPT solution, testing other software-based WAN emulators, and we also would like to get hands-on experience with the capabilities offered by hardware WAN emulators.

ACKNOWLEDGEMENT

This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

AUTHOR CONTRIBUTIONS

S. Szilágyi: Conceptualization, Theoretical analysis, Writing, Review and editing, Supervision.
I. Bordán: Experiments, Scripting, Measurements.

DISCLOSURE STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

S. Szilágyi: https://orcid.org/0000-0003-3562-5062

REFERENCES

[1] S. Szilágyi, I. Bordán, Throughput Performance Measurement of the MPT-GRE Multipath Technology in Emulated WAN Environment, in: Proceedings of the 1st Conference on Information Technology and Data Science: CITDS 2020, Debrecen, Hungary, 2020, pp.187–195.
http://ceur-ws.org/Vol-2874/short17.pdf

[2] S. Szilágyi, I. Bordán *et al.*, Throughput Performance Comparison of MPT-GRE and MPTCP in the Gigabit Ethernet IPv4/IPv6 Environment, *Journal of Electrical and Electronics Engineering* 12 (1) (2019) pp. 57–60.
https://electroinf.uoradea.ro/images/articles/CERCETARE/Reviste/JEEE/JEEE_V12_N1_MAY_2019/10%20paper%200414%20SZILAGYI.pdf

[3] B. Almási, S. Szilágyi, Throughput Performance Analysis of the Multipath Communication Library MPT, in: Proceedings of the 36th International Conference on Telecommunications and Signal Processing: TSP 2013, Rome, Italy, 2013, pp.86–90.
https://doi.org/10.1109/TSP.2013.6613897

[4] B. Almási, S. Szilágyi, Investigating the Throughput Performance of the MPT Multipath Communication Library in IPv4 and IPv6, *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems* 5 (1) (2016) pp. 53–60.
https://doi.org/10.11601/ijates.v5i1.148

[5] B. Almási, G. Lencse, S. Szilágyi, Investigating the Multipath Extension of the GRE in UDP Technology, *Computer Communications* 103 (1) (2017) pp. 29–38.

https://doi.org/10.1016/j.comcom.2017.02.002

[6] S. Szilágyi, I. Bordán *et al.*, MPT-GRE: A Novel Multipath Communication Technology for the Cloud, in: Proceedings of the 9th IEEE International Conference on Cognitive Infocommunications: CogInfoCom 2018, Budapest, Hungary, 2018, pp.81–86. https://doi.org/10.1109/CogInfoCom.2018.8639941

[7] S. Szilágyi, F. Fejes, R. Katona, Throughput Performance Comparison of MPT-GRE and MPTCP in the Fast Ethernet IPv4/IPv6 Environment, *Journal of Telecommunications and Information Technology* 3 (2) (2018) pp. 53–59. https://doi.org/10.26636/jtit.2018.122817

[8] S. Szilágyi, I. Bordán *et al.*, Throughput Performance Analysis of the Multipath Communication Technologies for the Cloud, *Journal of Electrical and Electronics Engineering* 12 (2) (2019) pp. 69–72. https://electroinf.uoradea.ro/images/articles/CERCETARE/Reviste/JEEE/JEEE_V12_N2_OCT_2019/13%20paper%201008%20SZILAGYI.pdf

[9] S. Szilágyi, I. Bordán, The Effects of Different Congestion Control Algorithms over Multipath Fast Ethernet IPv4/IPv6 Environments, in: Proceedings of the 11th International Conference on Applied Informatics: ICAI 2020, Eger, Hungary, 2020, pp.341–349. http://ceur-ws.org/Vol-2650/paper35.pdf

[10] G. Lencse, S. Szilágyi *et al.*, MPT Network Layer Multipath Library, Internet Draft, version 8, 2021. [cited 2021-06-30] https://datatracker.ietf.org/doc/html/draft-lencse-tsvwg-mpt-08

[11] D. T. Tieu *et al.*, Automatic test framework for video streaming quality assessment, in: 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science: NICS, IEEE, 2015, pp. 214–218. https://doi.org/10.1109/NICS.2015.7302193

[12] T. Zhang, S. Zhao, B. Cheng, Multipath Routing and MPTCP-Based Data Delivery Over Manets, *IEEE Access* 8, (2020) pp. 32652–32673. https://doi.org/10.1109/ACCESS.2020.2974191

[13] J. Huang *et al.*, Tuning high flow concurrency for MPTCP in data center networks, *Journal of Cloud Computing* 9 (13), (2020) pp. 32652–32673. https://doi.org/10.1186/s13677-020-00160-3

[14] P. Ignaciuk, M. Morawski, Discrete-Time Sliding-Mode Controllers for MPTCP Networks, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, (2019) pp. 1–11. https://doi.org/10.1109/TSMC.2019.2959630

[15] H. Tong, T. Wang *et al.*, Mobility-Aware Seamless Handover With MPTCP in Software-Defined HetNets, *IEEE Transactions on Network and Service Management* 18 (1), (2021) pp. 498–510. https://doi.org/10.1109/TNSM.2021.3050627

[16] M. Chen, T. Dreibholz *et al.*, Improvement and Implementation of a Multi-Path Management Algorithm based on MPTCP, in: Proceedings of the 2020 IEEE 45th Conference on Local Computer Networks: LCN 2020, 2020, pp.134–143. https://doi.org/10.1109/LCN48667.2020.9314778

Acta Technica Jaurinensis
Vol. 15, No. 1, pp. 15–21, 2022
10.14513/actatechjaur.00628

Comparison of data augmentation methods for legal document classification

Gergely M. Csányi[1,2,*], **Tamás Orosz**[1,3]

[1]Montana Knowledge Management Ltd.,
Hársalja Str. 32., H-1029, Budapest, Hungary
[2]Budapest University of Technology and Economics, Department of Electric Power Engineering
J. Egry Str. 18., H-1111, Budapest, Hungary
[3]University of West Bohemia,
Univerzitni 26, 306 14 Pilsen, Czech Republic
*e-mail: csanyi.gergely@montana.hu

Submitted: 20/05/2021; Accepted: 09/07/2021; Published online: 16/07/2021

Abstract: Sorting out the legal documents by their subject matter is an essential and time-consuming task due to the large amount of data. Many machine learning-based text categorization methods exist, which can resolve this problem. However, these algorithms can not perform well if they do not have enough training data for every category. Text augmentation can resolve this problem. Data augmentation is a widely used technique in machine learning applications, especially in computer vision. Textual data has different characteristics than images, so different solutions must be applied when the need for data augmentation arises. However, the type and different characteristics of the textual data or the task itself may reduce the number of methods that could be applied in a certain scenario. This paper focuses on text augmentation methods that could be applied to legal documents when classifying them into specific groups of subject matters.

Keywords: *text augmentation; augmenting legal cases; legal document classification; data augmentation*

I. INTRODUCTION

The digitalization of the judicial systems needs to process, categorize and pseudonymize many sensitive legal documents before they are published online [1–3]. This paper focuses only on the automatic categorization of legal documents. Text classification or text categorization is an essential branch in Natural Language Processing (NLP) [4–6], where the role of the different machine learning-based automatic text classification procedures is to automatically assign predefined labels for different documents (**Fig. 1**). For instance, the different legal documents can be sorted into different classes by their subject matter, such as theft, embezzlement, fraud, etc. [7].

Nevertheless, the classification of legal documents belongs to the class of multi-labeled categorization, which means that a legal document can belong to more than one legal category. This is a recent and relevant topic of research [7, 8]. There are different mathematical methods proposed to handle this task. Some of them use strategies – such as label powerset or binary relevance transformations [7, 9] – to convert back this selection into a single label classification task, while others extend the numerical methodologies to handle these kinds of tasks. Some examples of machine learning techniques from this latter group are: multi-label k-nearest neighbours, multi-label Naive Bayes, or multi-label AdaBoost [10].

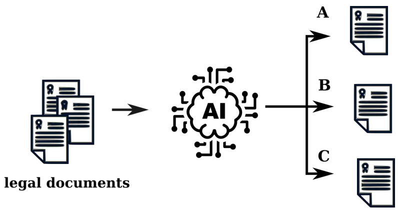


Figure 1. General process of legal document categorization

Fig. 1. shows the general process of legal document categorization, where the A, B, and C categories refer to specific subject matters, e.g., a crime of theft, the invalidity of a contract, causing a traffic accident, etc. Subject matters are generally highly imbalanced, which is illustrated on the left side of **Fig. 2**. Here, the diagram represents the number of documents (training samples) belonging to different categories. Generally, machine learning models tend to perform better when having approximately the same amount of training data for each class. However, as **Fig. 2**. shows, in practice, usually this is not the case. In this case, when the dataset is highly imbalanced, or the minority class has only few members, data augmentation techniques can help balancing the dataset.

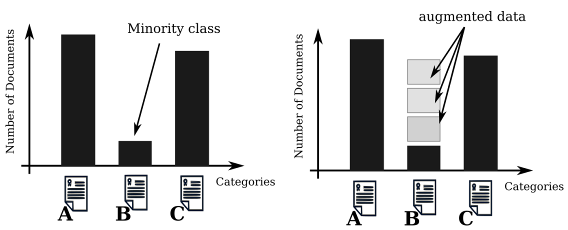


Figure 2. Augmenting imbalanced datasets

Text augmentation is a useful technique, which can automatically generate training samples. Hence, it can help to improve the performance of machine learning applications in this case. Data augmentation [11] is a well-known technique for training more robust machine learning models that have been successfully applied in the field of computer vision [12]. Generally, augmenting images is done by mirroring, rotating, cropping, scaling, flipping, etc., but these methods often cannot be applied on texts [13]. This is because the order of words is important in texts, and changing it could end up in sentences with completely different meanings. However, the need for augmenting data is not present only in the field of computer vision since imbalanced or small datasets occur in all types of machine learning applications. In case of imbalanced or small datasets, it is likely that the machine learning models overfit or do not fit well on training data. Hence augmentation can improve the robustness and performance of the models. Recently, many studies have been published to tackle the problem of data augmentation in the NLP field [14–16]. Some approaches depend more on the language or language models [14, 17], while others are (almost) independent [15, 18]. However, when applying text augmentation, one must pay attention to the characteristics of the text and the problem to be solved, since both of these may affect what type of augmentation techniques can be applied.

The main contribution of this work is providing a comprehensive survey on the possible text augmentation techniques in the case of legal document classification. Legal cases are relatively long (around 1,000-2,000 words), semi-structured texts. Therefore, certain parts of a case can be found in every document, generally well-spelled but not faultless documents. It is important to point out that matter of facts often consist of legal terms. Hence, certain words cannot be replaced or removed by any means. For example, theft, embezzlement, and fraud might be considered relatively close synonyms in general language, but they refer to three completely different types of crime in the legal context. These specialties of legal texts require special attention during the selection of the appropriate text classification methodology.

The paper is organized as follows: Section II. shows similar studies, Section III. presents a brief overview of typical text augmentation techniques and in Section IV a discussion of the useful techniques can be found.

II. SIMILAR STUDIES

Yan et al. presented a solution for augmenting legal documents [13]. Their work aimed to tackle a crime prediction problem, predicting the accusations of a case when the matter of fact is given. They applied three different techniques on sentence level:

- randomly scramble sentences in the sample,
- randomly delete sentences in the sample,
- randomly insert the sentences with the same label in other samples

The neural-based classifiers gained a lot of performance (11 %, F1-score) by the above-mentioned augmentation techniques when the training data was relatively small (10 thousand documents) but significantly less when 10-15 times more training data was available.

Another solution for augmenting legal documents is the TauJud framework, designed for augmenting Chinese legal cases [19]. The solution performs a two-step augmentation process, namely the Universal Augmentation and Judicial Augmentation steps. The former includes stop word deletion, back translation (RTT), and clipping, while the latter includes counterfactual data augmentation [20], and synonym replacement. It is possible to set what kind of augmentation steps have to be done and to choose multiple from these steps simultaneously. However, protecting words from augmentation is missing from the framework's repertoire.

III. TYPICAL AUGMENTATION TECHNIQUES

1. Easy Data Augmentation

Easy Data Augmentation (EDA) has gained interest after defining four simple methods for augmenting textual data and showing the efficiency of this approach on five different classification tasks [15]. The methods were the following:

- Synonym Replacement (SR) – Select n pieces of words from the sentence that are not stopwords. Replace each of these words with one of its randomly chosen synonyms.
- Random Insertion (RI) – Pick a random word in a sentence. Add the randomly selected synonym of this word to a random position of the sentence.
- Random Swap (RS) – Swap position of randomly chosen word pairs n times.
- Random Deletion (RD) – Randomly delete words from the document with probability p.

The latter two methods are completely language-independent, while the first two require a language-dependent WordNet database [21]. Nevertheless, these methods do not require a pretrained language model like GPT-2 [22] or word embeddings [23–25]. The power of these techniques lies in the simplicity of the solution, while the authors reported significant gain (around 3% on average) by using these techniques on text classification tasks.

2. Round-trip translation

Round-trip Translation (RTT) is an augmentation technique that harnesses the fact that translating a text to a random language and translating back to the original one in the majority of the cases (depending on the length of the text) results in a slightly different text, yet preserving the original meaning [26–30]. The technique is also known as recursive, back-and-forth, and bi-directional translation.

3. Semantic similarity augmentation

By means of distributed word representations, semantically similar words can be identified [23]. Hence, textual data can be easily augmented by replacing a fraction of the original text with the nearest neighbours of the chosen words. This approach requires either pre-trained word embedding models for the language in question or enough data from the target application to build the embedding model [16]. Thus, this approach does not require access to a dictionary or thesaurus for a language to find synonyms [16]. This can be advantageous for languages where such resources are more difficult to obtain, but there is enough unsupervised text data to be able to build the embedding models [16]. As word embedding models e.g. Word2Vec [23], GloVe [25], FastText [24] could be used, but these models may not handle words that are homonyms (multiple-meaning words) properly. By transformer-based language models such as BERT [31], the representation for words can be obtained in a context-dependent manner, providing a better solution for homonyms but with significantly more effort.

4. Text generation

Another approach for text augmentation is to use pre-trained language models to generate random texts. While this could be made by an LSTM-based encoder-decoder network, this type of solution would require a significant amount of training data and would generate grammatically incorrect sentences [32]. Another, more sophisticated approaches would be using generative adversarial networks (GANs) [33], variational autoencoders (VAE) [34], or paraphrasing [35].

GPT [36], GPT-2 [22] models are capable of producing grammatically correct, high-quality texts even when fine-tuned on small training data [14]. Nevertheless, the lack of ability to preserve or protect certain words from the original text cannot be assured by this method either.

5. Synthetic Minority Oversampling Technique

This technique is somewhat different from the already mentioned ones. Synthetic Minority Oversampling Technique (SMOTE) cannot be applied of the original text, but on its representations [37].

The basic idea of this method is, that by assuming that in the representation space the points between samples from minority class also belong to the minority class. Hence, creating synthetic samples is done by selecting random points from the lines between original data points as **Fig. 3**. shows. These synthetic data points serve as extra data for training a classifier.

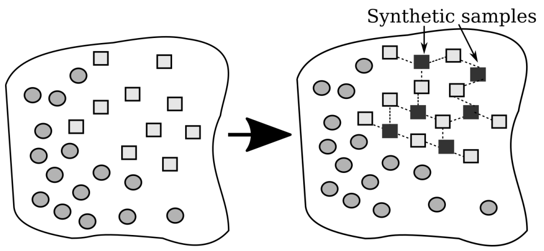


Figure 3. Applying Synthetic Minority Oversampling Technique on an imbalanced dataset to generate new training samples and create a more balanced training dataset.

IV. EVALUATION OF AUGMENTATION TECHNIQUES FOR SUBJECT MATTER CLASSIFICATION

Due to the special characteristics of legal cases, namely the length of the documents and the presence of legal terms, some methods are mentioned in Section III. cannot be applied. In the case of Easy Data Augmentation solutions **Random Deletion** could be used but with a restriction of not deleting specific words, called protected words. These are the words that refer to a certain subject matter, but in other applications, the need for having a list of protected words may also arise. **Random Swap** could be used without restrictions. However, one may have to keep in mind that in certain cases, this would not make much sense, e.g., when using word-order-independent document representation forms like tf-idf vectorization, or calculating document vectors by averaging word embeddings [38, 39]. Both **Synonym Insertion** and **Synonym Replacement** can be applied, but only with caution since adding synonyms of legal terms may end up in changing labels that are not wanted during augmenting data. One solution for this would be the same as mentioned before: by defining a list of protected words, these words cannot be used during random selection, ensuring that the most important words will not change. WordNet [21] provides different domains that can be used to restrict the scope of words for searching synonyms. For instance, synonyms for time-related data, colors, geographical locations, etc., will not change the label of the augmented legal case, yet provide another case that is somewhat different from the original, in other words augmenting the original case. It is important to point out that the quality of methods based on synonyms is highly dependent on whether the words in a text are stemmed or lemmatized or left as they are. This is true especially for highly inflected languages (e.g. German, Spanish, Russian, Hungarian), since WordNet databases contain lemmas only, so the number of words that can be selected for synonym modification (replacement or insertion), is usually significantly higher when the input text has been lemmatized beforehand.

The method of **Round-trip Translation** cannot be used as an augmentation method since there is no control over the words in the document. One cannot define protected words. As mentioned before, this would be important, since this way, after the double translation, the augmented document would not have the same label as the original one. The same problem arises with **text generation** solutions that can be useful in many applications, but there are no guarantees that the legal terms will be generated properly or will be kept intact.

Semantic similarity augmentation method is also an option that could be used during the augmentation of legal cases to classify the subject matters. The principle of using a list of protected words is also important here, since the most similar words to a given word depend on factors, like what kind of text the model was trained on etc., so keeping certain words intact otherwise would be practically impossible. These embedding models capture semantic similarities by assuming that words occurring in the same or similar contexts have similar meanings. An advantage of this method is that nearest neighbours of a certain word are not only synonyms, but words that occur in similar contexts, so a wider, more general augmentation is possible with these methods. However, it is important to emphasize that the nearest neighbours are highly dependent on the corpus on which the data was trained on and the size of the embedding vector. Another drawback of this method is the question of out-of-vocabulary (OOV) words. If the given document contains many OOV words, the quality of the augmented document may be affected. There are solutions to handle this problem. One of them is using FastText embeddings that can map any word into the embedding space, even if they are OOV words, by harnessing the power of subword information. Moreover, this type of embedding proved to perform well on highly inflected languages [24]. However, using FastText embeddings can be a double-edged sword since applying it on OOV words can result in useful embeddings and completely useless ones. Hence, careful analysis has to be made before deciding for or against using FastText to deal with OOV words.

V. CONCLUSIONS

Data augmentation is a very important technique that has proven its effectiveness in a high variety of machine learning tasks. This paper provided a brief overview of textual data augmentation techniques, putting more emphasis on classification of legal cases and comparing the available solutions. Current solutions can be applied on short texts e.g. on sentence level and may be ineffective on longer texts in terms of run time. Legal texts contain a lot of legal terms that have to be handled with caution when augmenting data since synonyms of these terms may refer to a completely different legal term, adding a significant bias to the augmented data. It can be stated that none of the mentioned solutions deal with this issue, however, some of them can be extended to solve it but not all of them. Generally, the more augmentation steps are applied, the better the expected quality of the augmented dataset will be reached, but there is no golden rule, which can be applied for every single problem.

ACKNOWLEDGEMENT

This project (no. 2020-1.1.2-PIACI-KFI-2020-00049) has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2020-1.1.2-PIACI KFI funding scheme.

AUTHOR CONTRIBUTIONS

G. M. Csányi: Conceptualization, Software, Writing, Review, Editing.

T. Orosz: Supervision, Software, Writing, Review and Editing.

DISCLOSURE STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

G. M. Csányi http://orcid.org/0000-0001-8475-5969

T. Orosz http://orcid.org 0000-0002-8743-3989

REFERENCES

[1] E. Comission, “Digitalisation of justice.” "https://ec.europa.eu/info/policies/justice-and-fundamental-rights/digitalisation-justice_en", ”(accessed: 01.02.2021)”.

[2] E. Hyvönen, M. TAMPER, E. IKKALA, S. SARSA, A. OKSANEN, J. TUOMINEN, and A. HIETANEN, “Lawsampo: a semantic portal on a linked open data service for finnish legislation and case law,” in *Proceedings of ESWC*, 2020.

[3] “Pseudonymization according to the gdpr [definitions and examples].” "https://dataprivacymanager.net/pseudonymization-according-to-the-gdpr/", ”(accessed: 02.15.2021)”.

[4] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, “Text categorization: past and present,” *Artificial Intelligence Review*, vol. 54, no. 4, pp. 3007–3054, 2021.

[5] P. Jackson and I. Moulinier, *Natural language processing for online applications: Text retrieval, extraction and categorization*, vol. 5. John Benjamins Publishing, 2007.

[6] P. J. Hayes, L. E. Knecht, and M. J. Cellio, “A news story categorization system,” in *Second Conference on Applied Natural Language Processing*, pp. 9–17, 1988.

[7] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, “Categorizing feature selection

methods for multi-label classification," *Artificial Intelligence Review*, vol. 49, no. 1, pp. 57–78, 2018.

[8] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[9] G. E. Tsekouras, C. Anagnostopoulos, D. Gavalas, and E. Dafhi, "Classification of web documents using fuzzy logic categorical data clustering," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 93–100, Springer, 2007.

[10] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*, pp. 667–685, Springer, 2009.

[11] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," in *2016 international conference on digital image computing: techniques and applications (DICTA)*, pp. 1–6, IEEE, 2016.

[12] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[13] G. Yan, Y. Li, S. Zhang, and Z. Chen, "Data augmentation for deep learning of judgment documents," in *International Conference on Intelligent Science and Big Data Engineering*, pp. 232–242, Springer, 2019.

[14] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? deep learning to the rescue!," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7383–7390, 2020.

[15] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[16] V. Marivate and T. Sefara, "Improving short text classification through global augmentation methods," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 385–399, Springer, 2020.

[17] Y. Li, T. Cohn, and T. Baldwin, "Robust training under linguistic adversity," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 21–27, 2017.

[18] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," *arXiv preprint arXiv:1805.06201*, 2018.

[19] Z. Guo, J. Liu, T. He, Z. Li, and P. Zhangzhu, "Taujud: test augmentation of machine learning in judicial documents," in *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 549–552, 2020.

[20] R. Zmigrod, S. J. Mielke, H. Wallach, and R. Cotterell, "Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 1651–1661, Association for Computational Linguistics, July 2019.

[21] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[25] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[26] A. Gerasimov, D. Nogueira, K. Semolini, S. Firoozkoohi, R. A. Rivera, T. a Patent, K. K. Zerling, L. García-Santiago, M.-D. Olvera-Lobo, M. Aiken, *et al.*, "The efficacy of round-trip translation for mt evaluation,"

[27] S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 90–97, 2018.

[28] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[29] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," *arXiv preprint arXiv:1705.00440*, 2017.

[30] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[32] S. Sharifirad, B. Jafarpour, and S. Matwin, "Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs," in *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pp. 107–114, 2018.

[33] F. H. K. d. S. Tanaka and C. Aranha, "Data augmentation using gans," *arXiv preprint arXiv:1904.09135*, 2019.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[35] A. Kumar, S. Bhattamishra, M. Bhandari, and P. Talukdar, "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3609–3619, 2019.

[36] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[38] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf* idf, lsi and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[39] D. Mekala, V. Gupta, B. Paranjape, and H. Karnick, "Scdv: Sparse composite document vectors using soft clustering over distributional representations," *arXiv preprint arXiv:1612.06778*, 2016.

ACTA TECHNICA JAURINENSIS
Vol. 15, No. 1, pp. 22-35, 2022
10.14513/actatechjaur.00638

The interplay between the physical internet and logistics: A literature review and future research directions

Safwen Ben Neila[1], Abderahman Rejeb[1,*], Péter Németh[1]

[1] Széchenyi István University, Department of Logistics and Forwarding,
Egyetem tér 1. 9026 Győr, Hungary
*e-mail: abderrahmen.rejeb@gmail.com

Submitted: 07/07/2021 Accepted: 09/08/2021 Published online: 16/09/2021

Abstract: In recent years, there is a need for new methods and frameworks for planning transport systems, improving their efficiency, and addressing globalisation and sustainability challenges. In addition, the use of existing capacities and infrastructure has raised significant issues in the transport sector. To achieve an economically, environmentally, and socially sustainable logistics ecosystem, Benoit Montreuil introduces the concept of "Physical Internet" (PI) to increase the efficiency and effectiveness of logistics networks. As a ground-breaking transportation philosophy, the PI aims to revolutionise freight and logistics transport. Moreover, the PI can enhance logistics productivity through the organisation of large-scale pooling. Similar to Digital Internet that conveys data, the PI strives to connect, synchronise, and ship regular modular containers from the point of origin to an exact destination, thereby creating robust and collaborative logistics networks. While the literature on the PI is relatively growing, there is still a lack of reviews that synthesise this knowledge body, identify current trends and gaps, and advance the research more broadly. Therefore, this study aims to investigate the potentials of the PI for the development of sustainable logistics networks. Overall, 59 studies are selected from leading academic databases and further analysed. The review findings reveal that most scholars focus on the optimisation of transport at the tactical and organisational stage while devoting little attention to the contribution of the PI to the social sustainability of logistics compared to the economic and environmental aspects.

Keywords: physical internet (PI); logistics; transportation; sustainability

I. INTRODUCTION

Over the last decade, the fight against climate change has become one of the controversial topics and is currently identified as one of the severe problems societies have ever faced [1]. The main reason for climate change is attributed to Greenhouse Gases (GHG) emission and improper human activities. With the business and logistics community's mounting attention to this matter, there is a need for business talent, investment, problem-solving skills, and global reach to address the intractable problem of climate change and the inequality to equip workers for a new age of work [2]. Similarly, new management approaches are necessary to reduce socioeconomic and ecological imbalances, prevent economic crises, and achieve sustainability [3]. In response to increasing globalisation and customer requirements, the focus of organisations has shifted to developing optimisation initiatives that reduce logistics and transport costs [4]. One way to achieve this objective is to integrate the Physical Internet (PI) in the logistics sector. According to [5], the PI contributes to enabling order-of-magnitude improvements in logistics efficiency and sustainability. Thanks to its interconnections and open characteristics, the PI has the potential to help organisations meet sustainability challenges [6]. The approach of PI is based on establishing global cooperation and interoperability to facilitate both vertical and horizontal integration of transport systems globally and streamline sharing of resources, such as vehicles, data, spaces and warehouses [7]. Several scholars have started to consider the PI and scrutinise its possibilities in multiple stages of transport network planning.

During the realisation of the PI, strategic-level decisions include the design of distribution centres and the determination of the number of dock doors, flow racks, the material handling equipment used, and the travel distances. Whereas tactical-level decisions include the number of inbound trailers received, the number of unloaders and hauliers, the storage location assignment, etc. Operational-level decisions are concerned with inbound intralogistics, including unloaders' assignment and scheduling, loads-to flow rack assignment, and assignment and hauliers' scheduling [8]. Within the PI, the working principles and basics of the Digital Internet metaphorically correspond to handling, transportation, distribution, and storage of physical items. Breaking up the conventional business models, the PI creates a new setting of decentralised, collaborative, and cooperative logistics, spanning organisations, industries, and inter-organisational networks [9].

While the PI is receiving significant attention from researchers, there is still a lack of studies discussing the concept from the decision-making perspective. For example, [10] examine the state-of-the-art of the PI and provide several comprehensive future research directions. [11] discuss the synchromodal and PI state-of-the-art models and their architectures, methodologies, and observations in the academic literature. The authors' main goal is to analyse and investigate the connections between these synchromodal and PI models and explain how they are mutually reinforcing. Furthermore, [12] propose a PI structure recognising seven PI themes that can inform scholars and practitioners on the enablers, barriers, and policies of PI policies. Lastly, [13] study the PI literature using a qualitative content analysis method.

Unlike previous studies, we synthesise extant research on the PI using a coding framework consisting of three axes; the PI themes, the decision stages, and the sustainability aspects. The review contributes significantly to the literature because it increases researchers' understanding of the PI, the themes discussed in this literature, and the impact of this novel paradigm on sustainability. The systematic literature review includes reviews, articles and conference proceedings, business studies, and white papers. In addition, we carry out a descriptive and content analysis to have an overview of the PI literature's history and the themes discussed.

The structure of the paper is as follows. Section 2 presents a brief overview of the PI concept. Section 3 describes the methodology used for the review and content analysis. Section 4 provides an in-depth discussion of the results and findings. In the subsequent section, several recommendations for future research are highlighted, followed by a brief conclusion.

II. Overview of the PI

The PI is a metaphor of the Digital Internet applied to logistics networks and their services. The purpose behind the PI is to link up delivery networks, storage centres, vendors in an open global logistics system founded on physical, digital, and operational interconnectivity through encapsulation interfaces and protocols [5]. It is a perpetually evolving system driven by technological, infrastructural and business innovation [14]. Today, several players in the logistics industry are operating in isolation. Companies are autonomous and have their networks. However, with the support of the PI, the stakeholders of the logistics industry can maintain a high level of interoperability.

The PI provides an optimal solution on a broad scale, streamlining the logistics system, standardisation of resources (hardware and software), and the pooling of means. Through a standardised set of collaboration protocols, modular containers and smart interfaces, it is possible to develop a global logistics system and support the interconnection of logistics networks that can increase efficiency and sustainability [15]. Like the flow of data packages through the Internet, the PI is based on the use and handling of standardised smart boxes. These boxes can be visualised easily when moving throughout the supply chain, making logistics more efficient and sustainable. [5] argues that the way physical objects are handled, moved, stored, realised, supplied and used in the PI. He also affirms that it is a revolutionary concept, which is strikingly different from the currently prevailing theory and practice of transport, logistics and the supply chain. The PI takes advantage of the Digital Internet metaphor to shape the physical world and ensure global logistics efficiency and sustainability.

III. Research Methodology

In general, literature reviews aim to map, consolidate, and analyse a particular research area and find gaps to fill in to expand further the current information base [16]. Given that the PI literature exists in different outlets and spans several fields such as logistics, supply chain, and business, a comprehensive analysis of the current state of the PI is urgently needed. This research will provide better conceptualisation by updating the standard terminology of the PI and suggesting recommendations for future PI studies. As part of this focus, we carried out a systematic literature review to provide a rigorous and detailed analysis of the current research on the PI. Five steps were followed to ensure the extraction of suitable studies and their systematic inspection. Firstly, we expressed the problem of analysis and researchers' choice. Secondly, we identified the inclusion and exclusion criteria. Thirdly, we consulted the academic databases, filtered the studies according to our selection criteria, and chose suitable papers. Fourthly, we discussed and analysed the findings, and finally, we reported the results.

1. Definition of keywords

In the titles, abstracts, and keywords fields, we looked for the term "Physical Internet". Given the subject's specificity, this was the only word appropriate for a complete quest. Next, we set the period of the search from 2011 to the beginning of 2021. We chose 2011 as the starting year because the PI was systematically discussed and introduced in this year.

2. Inclusion and Exclusion Criteria

The first aim was to screen publications based on their titles, abstracts, and keywords. During this process, we only included publications with an exclusive focus on the PI and excluded those concerned with the Digital Internet's physical infrastructure or device hardware. We also reviewed the chosen publications' sources to help to identify further publications. Since several databases were used, we merged all publications and eliminated the duplicates. The publications in question were read very carefully and debated between the reviewers until an agreement was reached.

3. Data search

We identified three sub-steps in the data search process. The first concerns the selection of databases. We chose Google Scholar, Web of Science, Scopus, Springer, Science Direct, and Wiley Online Library as our root databases. We then started searching in the different databases using the keywords identified in this first phase, i.e., "Physical Internet" and combined it with "logistics," "transport," or "freight." In the first stage, we collected the first list of publications based on the title and abstract.

4. Selection of the most relevant papers

In this stage, we implemented a filter based on the inclusion/exclusion criterion to identify all potentially relevant publications. This task attempts to rank the most important publications that dealt with the PI concept and exclude irrelevant studies. We read each paper during the selection phase and checked the list of references for each item to identify potentially relevant publications. After the full-content reading, 59 publications were identified as relevant for the final review and analysis.

5. Coding process of selected papers

We developed a coding framework based on the study objectives raised in the introduction following three main axes:

Axis 1: Themes

After the full-content reading and analysis of all selected publications, thirteen themes emerged:

- - Conceptualisation of the PI
- Detailed literature review of the PI
- Standardised modular containers, space/volume utilisation, handling cost, intelligent boxes.
- Inventory, optimised inventory levels, warehousing services, reduced inventory costs, maximised utilisation, etc.
- Distribution and transport, network optimisation, optimised routing, loading patterns, and truck scheduling
- Production, intelligent and dynamic manufacturing
- Dynamic pricing in the PI
- Auction trading in the PI
- Interconnected city logistics
- Cloud logistics platform
- Synchromodality
- Blockchain technology for the PI
- Horizontal collaboration in the PI

Axis 2: Decision-making level

- Strategic level: L1
- Tactical level: L2
- Operational level: L3

***Table 1**. Decision-making level*

Levels	***L1***	***strategic level***
	L2	***tactical level***
	L3	***operational level***

Axis 3: Sustainability dimensions

- Economic performance (D1)
- Environmental performance (D2)
- Social performance (D3)

***Table 2.** Sustainability dimensions*

Dimensions	***D1***	***Economic***
	D2	***Social***
	D3	***Environmental***

IV. RESULTS AND FINDINGS

1. Publications per year

Fig. 1. depicts the annual distribution of the selected publications. In recent years, 76% of the studies were written since 2016, and the number of publications experienced inconsistent growth. We identified twelve articles (20%) published in 2020 and the beginning of 2021; thus, we can expect further progress in the research areas for this year and the next coming years.

2. Publications by journal

The selected publications were published in 39 international journals. As shown in **Fig. 2.** the International Journal of Production Research published over 38% of the publications, followed by

Table 3. *Classification of the literature based on the PI themes, the decision levels, and the sustainability dimensions*

Authors	Themes	Levels	Dimensions
[9]	Conceptualization of PI	L1	D1+D2+D3
[5]		L1	D1+D2+D3
[17]		L1	D1+D2+D3
[18]		L1	D1+D2+D3
[19]		L1	D1+D2+D3
[20]		L1	D1+D2+D3
[11]	Detailed literature review of PI		
[10]			
[12]			
[13]			
[21]	Standardised modular containers—space/volume utilisation, handling cost, intelligent containers. . .	L1+L2+L3	D1+D2+D3
[22]		L3	D1+D2+D3
[23]		L3	D1+D2+D3
[24]		L1	D1+D2+D3
[8]		L3	D1+D2+D3
[25]		L3	D1+D3
[26]		L2	D1+D3
[27]	Inventory—optimised inventory levels, warehousing services, reduced inventory costs, maximised utilisation, etc.	L2	D1
[28]		L2+L3	D1
[29]		L2	D1
[30]		L2+L3	D1
[31]		L2	D1
[32]		L2	D1
[6]		L3	D1+D2+D3
[43]	Production—intelligent and dynamic manufacturing	L2	D1+D3
[44]		L2	D1
[45]		L1	D1+D2+D3
[46]	Dynamic pricing in PI	L3	D1
[47]		L3	D1
[48]	Auction trading in PI	L2+L3	D1
[49]		L2	D1+D3
[50]		L2	D1
[33]	Distribution and transport—network optimisation, optimised routing, loading patterns, truck scheduling, etc.	L2+L3	D1+D3
[34]		L3	D1
[35]		L2+L3	D1+D2+D3
[36]		L3	D1+D2+D3
[37]		L2+L3	D1
[38]		L2+L3	D1
[39]		L3	D1
[15]		L1	D1+D2+D3
[40]		L1	D1+D2+D3
[41]		L3	D1
[42]		L1+L2+L3	D2
[51]	Interconnected city logistics	L1+L2+L3	D1+D2+D3
[52]		L3	D1+D2+D3
[53]		L3	D1+D2+D3
[54]		L3	D1+D2+D3
[55]		L1	D1+D2+D3
[56]		L1	D1+D2+D3
[57]	Cloud logistics platform	L1	D1+D2+D3
[58]		L3	D1+D3
[59]	Synchromodality	L3	D1+D3
[60]		L3	D1
[61]	Blockchain technology for PI	L2	D1+D2+D3
[62]		L1	D1+D2+D3
[7]		L1	D1+D2+D3
[63]		L1	D1+D2+D3
[64]	Horizontal collaboration in PI	L2	D1+D2+D3
[4]		L1	D1+D2+D3

the Journal of Business and Logistics, and the Journal of Intelligent Manufacturing. The rest of publications were published in 36 separate journals. These journals include, among others, "Transportation Journal," "Journal of Physical Distribution and Logistics Management," "The International Journal of Logistics Management," "International Journal of Advanced Logistics," "International Journal of Signal Processing, Image Processing, and Pattern Recognition," "Journal of Enterprise Information Management," "International Journal of Logistics Systems and Management," "International Journal of Computer Integrated Manufacturing" and "International Journal of Transport Development and Integration. "

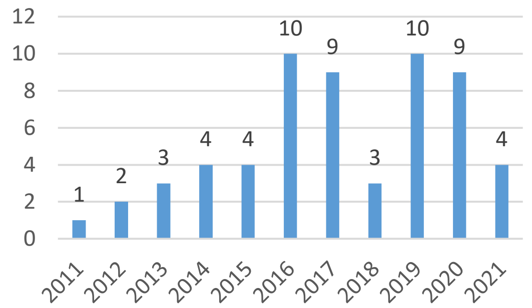


***Figure 1**. The annual distribution of selected publications*

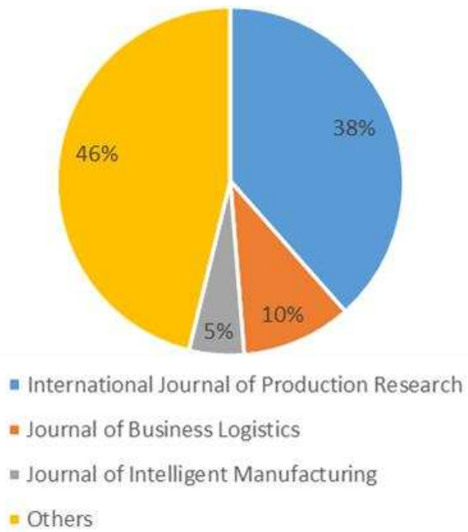


***Figure 2**. The distribution of publications according to journals*

3. Analysis of coding framework

A. Themes of PI

Several previous reviews focused on diverse PI themes, such as [4], [12]. In our study, we considered these themes. Additionally, we considered other categories in our coding framework, namely decision levels and sustainability dimensions, aiming to inform researchers on how the decisions taken within the PI implementation and the impact of this paradigm on logistics sustainability.

B. Decision levels

Freight transport networks are dynamic systems containing various human, material, and capital resources that increased complexities of decisions and management strategies. According to [65], various freight networks need to be structured on three levels:

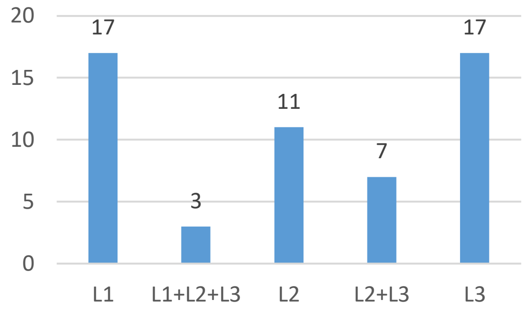


***Figure 3.** Distribution of studies according to the decision-making levels addressed*

Strategic level

The strategic level relates to the business strategy that typically involves top management and needs considerable long-term commitment. At this decision level, the fundamental choices apply to the facilities required to ensure the movement of goods in diverse modes to reduce transportation costs [66]. Several difficulties arise at the strategic level, such as selection of the terminal site (e.g., multimodal cross-docking station), platform facilities, the construction of additional lines (roads, railways), and the demolition of specific infrastructure. On the one hand, we found out that many authors conceptualise the PI by using a business model innovation strategy, which led us to classify their publications into the strategic level of decision-making. For example, [9] introduces an audacious paradigm on how the physical object of transport moves forward and encourages the replication of the Digital Internet ambition to achieve sustainable global logistics. Later, [5] present a systematic description of the PI as an open global logistics infrastructure founded by encapsulation, interfaces, and protocols on physical, digital, and operational interconnection. [17] explore the potential impacts of PI-enabled business models on creativity. [18] discuss analogies and similarities between computer networks and logistic networks to examine the global structure of the interconnection of logistics networks that enables the PI. [20] offer an external and technical viewpoint on PI research to better inform researchers and policymakers. More recently, [19] suggest a model-based conceptual framework to introduce a feasible PI.

On the other hand, various articles classified into the categories production, transportation, distribution, city logistics, cloud logistics, and blockchain technology discuss long-term decision planning during the PI implementation. For example, [15] and [40] cover the cross-docking problem by conceptualising a road-rail Pi-hub and a road-rail Pi-hub, respectively, showing how these models could contribute to economic, environmental, and social sustainability. [56] discuss the idea of Hyperconnected City Logistics, along with nine basic concepts, which provide a rich context for the design of effective urban logistics and transport systems. [57] aim to create a PI-enabled Building Information Modelling System (PI-BIMS) incorporating Auto-ID technology, BIM technology, and cloud storage. [62] introduce an intellectual paradigm using blockchain to provide a workable solution for the fundamental obstacles hindering the exchange of values and physical properties in logistics networks and decentralised management frameworks in the PI. Furthermore, [63] investigates the integration of the PI and Blockchain technologies in a conceptual paper and proposes several research directions for investigating the supply chain's institutional issues during the implementation of these two paradigms. [7] also examine blockchain as an enabling technology for the PI. The authors summarise the core requirements of the PI and explain how blockchain can be used to facilitate the implementation of the PI. [55] and [56] focus on conceptualising Interconnected city logistics, while [24] offers strategic analysis of freight and PI-containers' total volumes and balances in a terminal-to-terminal network. [45] explores the relations between the PI and Lean concepts using the Value Stream Map (VSM) method for their integrated use. At a broader level, [55] outline the possibilities for the PI's logistics performance from the EU perspective. This article details the systematic EU planning for the PI definition and the initial projects, ventures, tools that underpin the PI. [4] examine the reasons for accessing a PI network and the effectiveness of continuous cooperation variables in a PI network.

Tactical level

The tactical level is mainly concerned with a medium-term strategy. Tactical planning extends from a week to few months. At this stage, vehicle fleet management in multiple modes ensures the movement of goods from their sources to their destination. The tactical level is the degree to which we calculate the transport movements over time between authorities and terminals, inter-terminals, and between terminals and destinations. Tactical planning creates a travel strategy to ensure customer service satisfaction and resource efficiency. Therefore, the tactical decisions are primarily concerned with the collection of services to be provided, their frequency (or schedule), the allocation of demands for specific services in transportation, and terminal level policies (consolidation).

Several articles discuss the inventory problem at the tactical level. For instance, [31] study inventory efficiency over the PI network. Later, [32] aim to identify the latest research topics surrounding inventory management in a PI Network and how the PI can impact conventional stock control policies. [28] and [29] introduce PI disruption mitigation strategies. They carry out their study using a simulation-based optimisation model. [27] suggest a mixed integrated linear programming (MILP) formulation to assess the benefits of the PI from the cost efficiency perspective and to resolve the issue that combines the integrated output inventory allocation judgment with PI. [43] and [44] discuss the production manufacturing scheduling problem at medium-term decision making. In the former, the authors analyse a PI-enabled manufacturing executive system (PIMES) for a machine workshop. For the processing and synchronisation of real-time data, PIMES depends heavily on RFID and wireless communication networks. In the latter, the authors implement a big data analysis framework for PI-based logistics data generated by installing RFID readers, tags, and wireless communications networks on manufacturing stores.

Other articles focus on the auction trading in the PI, such as [49], [50]. [49] provide combinatorial frameworks for the auction of transport services in the PI sense. [50] aim to take advantage of the PI's open and uncertain environments to distribute and even reconnect containers for each centre. According to the authors, the PI can improve container routing and the use of logistics network tools. However, to achieve these objectives, [64] emphasise the need for horizontal coordination in the PI and for more effective solutions to any potential technological difficulties. In the context of horizontal partnerships and the PI, the authors also explore the perspectives, perceptions, values, and motives of transit operators in Austria to contribute to the Austrian physical network's early moves positively. In addition, [61] make a simulation model showing that blockchain technology can support IoT automated data processing and facilitate the implementation of the PI vision by introducing the very first hyper-connected logistics application.

Operational level

The operational level of decisions concerns short-term management. This level includes diverse tasks such as assigning goods to liners and vehicles, container shipping, and truck routings.

Various researchers focus on container shipping issues. For example, [23] suggest a standard modular containers' production to minimise the number of boxes required for shipping goods using a decomposition-based solution methodology. [22] use a methodological engineering design process and explain how to create a modular and multifunctional loading unit by introducing a first real PI scenario in the fast-moving consumer goods industry and modelling the critical performance indicators (KPIs). Similarly, [8] propose a modular design consisting of three stages of characterisation, including the transport, storage, and packaging of PI containers. This paper outlines the proposed substance encapsulation transition and highlights significant opportunities and challenges for further research in industry and academia. [25] establish a new methodology for applying heuristic methods to address the liner shipping network design problem (LSNDP) using a mathematical model. Later, [26] concentrate on product development in the context of a global open logistics system, the PI, utilising a hyperconnected-mobile production mode. The study deals with the complex distribution of output units, allocation and sharing of resources.

Routing optimisation and transportation problems are the focus of several articles. For example, [39] highlight the efficient management of the PI for mobile resources, such as containers. In reality, management and storage of resources (containers) can cause imbalances in the logistic network, leading to starvation or unnecessary storage of logistic network nodes. For this reason, [36] present a π-container tracking device to create and hold the composite π-container practically 3D layout dynamically. [34] suggest a new method for routing based on the Internet BGP equivalent of BGP (the PI Border Gateway Protocol). To fix the issue of routing the PI-containers in the PI, the authors create this new protocol. Afterwards, [46], [47] present and explore a dynamic pricing model based on an auction mechanism for optimising transport service providers' bid prices in PI hubs. [41] define the best suitable hub for forwarding transport based on real-time traffic data. This article addresses the critical research questions of when and how to make routing decisions for road vehicles if the routing schemes within the PI take consider the data related to transport vacancies in real-time. [6] focus on integrated production-inventory-distribution planning in the PI for sustainability purposes. The key objective of their study is to thoroughly evaluate the efficiency of the PI-enabled integrated production-inventory-distribution model in the three sustainability dimensions.

Moreover, other studies discuss the interconnected city logistics (ICL) topic using simulation and mathematical models to solve some routing problems. To illustrate, [53] discuss PI-enabled urban transport and attempt to model essential organisational elements of the ICL initiative. [54] aim to explore an alternative approach to reverse e-commerce flows collection citywide, and [52] examine the potential for open mobility webs in an urban environment with e-commerce warehouses based on the PI concept.

Related to the theme Cloud logistics platform, [58] discuss crucial problems and difficulties with cloud logistics orders's deployment in China. We found two more articles that solve synchromodality issues with optimisation models. [59] examine how synchromodal transport can promote the modal transition to slower and more environmentally sustainable transport modes without sacrificing costs or responsiveness. [60] aim to extract rules for effective container allocation for inland services in real-time by analysing the solution structure of a centralised, offline system of optimisation used for historical results.

Tactical and operational level

Some other papers combine the tactical and organisational levels. For example, [30] solve a real problem in planning and distribution for multiple facilities that simultaneously maximise the effectiveness of decision-making in supply, inventory, demand, and distribution. [37] emphasise the difference between traditional logistics networks and the PI and characterised the PI by using a relatively more superficial P2P optimal dispatch MILP. [38] aim to increase the railroad PI-hub's performance by minimising the distance to the dock by each container and the number of used trucks. The research concerns a PI allocation issue. [48] propose a scheduling solution for trolley loading and sale in PI-Auction logistics centre (ALC) systematically. [28], [29] suggest a revolutionary vendor-managed inventory strategy that uses the PI's flexible, universal, interconnected logistics structure. [35] aim to compare PI's efficiency with the traditional logistics method to measure the truck and driver routing advantages and drawbacks of PI with an explicit limit to the driver's maximum return time. [33] propose an efficient and reactive multi-agent system-based model (MAS) for the resolution of trucks and PI-containers grouping.

All decisions level

Other articles englobe all levels of decisions, such as [21], who highlight the importance of semantics in distribution and logistical at multiple stages of judgment to achieve various optimisations such as bundling and durability, enforcement, and travel optimisation, thereby increasing sustainability through optimal use of capacities. Additionally, [51] attempt to establish the latest concept of an interconnected multilayer decision-making system and a PI methodology. [42] outline a theoretical framework that describes various levels of acceptance and trust as a key component of "Human-computer interaction" in technology innovation and point out a risk of an artificial divide both at the level of individuals and of companies.

C. Multiple freight transportation problems

The PI is a new concept made to serve supply chain sustainability throughout the world. [5] argue that three forms of efficiencies impact logistics and transport systems: social, economic, and environmental. First, the PI's societal goal is to improve the high quality of life of logistics personnel and the global population by improving physical items' timely accessibility and mobility. Second, the economic ambition is to sustainably reduce large-scale purchase strategies, international monetary burden, while freeing up enormous gains in companies' production. Third, the environmental intention is to reduce global greenhouse gas emissions caused by logistics, energy consumption, pollution, traffic, and waste materials.

The present state of the PI research focuses on studying the combination of economic and environmental goals. However, in the extant literature, the social element is less explored (**Fig. 4.**).

Fig. 4. shows that the PI research mainly centres on the economic dimension of sustainability. Of the 59 papers listed in **Table 1**, 17 papers highlight the financial targets of the PI. Seven papers analyse the PI's environmental and economic aspects, and 30 articles deal with the PI concept in the context of the three dimensions of sustainability. Finally, there is only one contribution to the PI concept from the perspective of social sustainability, and no papers deal with the PI and the convergence between the social and environmental and economic dimensions.

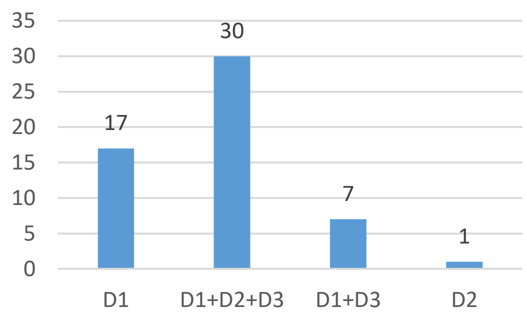


***Figure 4.** Count of dimensions*

The economic dimension

The most crucial economic issue is that the logistics and freight traffic prices rise at the global level. To solve this dilemma, numerous articles study the economics dispute on a tactical level. For example, [31] use objective-oriented modelling experiments to minimise the overall transportation and inventory costs. [50] utilise a simulation model for assigning and allocating containers at the hub to carriers while optimising container routing and using logistics network tools. [32] analyse an organisation's effect on inventory levels and costs using a multi-method simulation modelling software developed by XJ technologies (www.anylogic.com). [28], [29] examine how stock models using the PI react to intervention and durability in supply chain systems. The topic is explored through an advanced numerical analysis using the simulation-based optimisation model. [27] present a mathematical model to resolve the issue that incorporates an optimised production-inventory-distribution decision to measure the benefits of PI from the cost efficiency viewpoint. Likewise, plenty of researchers are interested in disbanding the operational issues. [39] conduct active research to optimise the freight by exposing the efficiency handling resources issues. [60] aim to develop guidelines for the efficient distribution of containers to inland installations in real-time by evaluating a critical offline optimisation approach used for historical outcomes. [34] focus on truck-transported PI-containers and suggest a new routing method to optimise road transportation. [46], [47] present and address a dynamic price model targeted at a service provider that offers a mechanism to optimise PI hubs' cost. The results and observations provide carriers with instructive guidelines on real-life pricing. [41] interview experts and make a qualitative analysis on artefact routing. More than that, some studies evaluate the economic issue at both the tactical and operational levels simultaneously. [30] use a mathematical model to minimise the total cost of production, inventory, and distribution. [37] suggest a cost-minimising mathematical model to minimise handling and delivery cost. [38] use a multi-agent system model (optimisation models) to reduce container travel distance to the dock. [48] conduct simulation experiments with a timely operation scheduling model to reduce logistics and labour costs. [28], [29] propose a retailer's creative inventory policy to minimise inventory levels, overall expense, and distances.

The environmental dimension

The harmful impacts of freight transport on the environment (greenhouse gases, energy consumption, noise) are growing amid the fact that nations and governments are seeking to enhance environmental quality, commit substantial budgets to environmental safety, and reduce the emission rate of air pollutants.

Several papers address economic and environmental issues, such as [49], who attempt to reduce the cost of transport services on the PI and greenhouse gas emissions by using a mathematical model. [43] suggest PIMES, an information system that can reduce paperwork and enhance information flow with an analytical data model. [43] propose a smart box compatible with PSS to provide revolutionary eco-logistics and safe packaging operation and minimise the logistics costs and emissions. [25] highlight the global CO2 emissions problem due to the large volumes transported and propose a heuristic for solving the liner shipping network design problem to maximise an organisation's profit. [59] research how synchro modal transport support a slower but eco-friendlier modal transition without compromising costs or responsiveness. [33] suggest models that discuss two of the three pillars: sustainable growth and the economic and environmental costs. The proposed models minimise the average expense from an economic point of view using a minimal number of cars and delay vehicles, leading to financial fines from the transport providers. From an environmental point of view, the two models decrease the number of used wagons, thus reducing CO2 emissions by the rail locomotive and the distance travelled for PI containers. [26] suggest a two-target optimisation model that minimises the cost and greenhouse gas of a producer to support its clients through the availability of open fabs.

The social dimension

The transportation of goods often causes social problems, such as road safety, lack of quick, efficient, accessible goods mobility. Furthermore, there is a shortage of stability, protection, and security in working conditions in the freight transport sector. We found that few studies address social issues facing logistics. For example, [42] discuss measures to improve human-computer interaction, especially the capacity of individuals and companies to implement state-of-the-art techniques and prevent an artificial divide, thus increasing social sustainability at an operational, tactical and strategic level. In addition, these articles attribute social problems along with financial and environmental issues [5], [9], [17], [18]. Further, [10], [19] evaluate and optimise certain aspects of the early conceptualisation of PI. [8], [21 – 24] develop the standardised modular containers to make the transport and handling of π-containers very easy. [6], [14], [15], [35], [36] seek to refine delivery network routing by reducing container journey distance that would positively affect carriers, lower congestion, and improve living conditions. [45] aims to reduce the number of trucks on roads and improve customers' quality of life and the logistics sector workers. [51 – 56] discuss the city logistics sector that offers less infrastructure usage and fewer jams, making consumers and community happier and distribution quicker. [7], [61 – 63] analyse how blockchain technology can improve visibility, security, and traceability of products through its decentralised and trusting mechanisms.

Researchers and professionals were prompted to carry out studies that aim to reduce various adverse effects of the lack of sustainability and unreliable traditional transport networks. Several perceptions have therefore arisen in recent years. Three creative theoretical frameworks are most effective in achieving sustainability, including collaborative logistics, city-logistics (CL), and the PI.

V. FUTURE RESEARCH

Our analysis recommends several topics for future studies. Our literature review urges PI researchers to focus more on theory and highlights the importance of a clear PI strategy in logistics. Addressing 'PI themes,' the 'decision stages', and the 'sustainability aspects' mentioned here will possibly provide a sound base for assessing the actual PI adoption and its impact on logistics sustainability. We took a closer look at critical and diverse topics regarding decision-making levels and sustainability aspects to provide recommendations for future research.

The PI literature addresses numerous strategy decisions ranging from strategic to operational. This demonstrates how the PI is relevant and beneficial to the logistics and transport sector. Researchers may expand future studies by evaluating the relationship between the logistical service providers and external factors, such as clients and competitors. In reality, consumer involvement through a collective strategy will improve business efficiency in general.

Environmental and social considerations have been relevant to broaden the economic implications and include other sustainable growth dimensions. This highlights the significance and value of incorporating facets of sustainable growth into planning decisions for logistics performance. However, social sustainability has been poorly considered compared to economic and environmental sustainability. This finding indicates that social dimensions must be studied to resolve all issues of sustainability holistically. Due to its qualitative nature, social sustainability is concerned with the employer's health and well-being issues. Therefore, future studies should also focus on quantifying this sustainability dimension. Integrating corporate social responsibility (CSR) in planning and policy-making helps researchers to properly determine the PI's effect on customers, workers, and the local population. It also allows identifying the detrimental influence of the PI on human lives to as well.

Usually, environmental sustainability is limited to measuring transportation CO2 emissions. The ecological evaluation revolves around assessing direct pollution from installations and inventory, waste volume and energy use.

Since the PI is still a fresh and novel concept, the lack of data is a barrier to further scientific advancement in this field. More studies are nevertheless necessary taking into account real-life cases. The collaboration between researchers and managers may result in insightful findings. Academics and researchers may study various real-life cases in the future. For example, this involves the validation of unimodal (road-based) and multimodal (road-to-rail) PI hubs, the identification of the next best appropriate hubs for onward transport based on real-time traffic data, the validation of the proposed design and types of PI containers, the confirmation of the proposed auction mechanisms, the optimisation of capacity utilisation for PI carriers in PI hubs, and validation of the proposed legal frameworks for horizontal collaboration. Most of the studies are concentrating on unimodal (road-based) and multimodal (road-to-rail) PI hubs. As a result, there is a lack of studies addressing maritime and air transportation and conceptualising multimodal road and rail to sea hubs. This will help to effectively connect multimodal road to air PI hubs.

[19] recommend a paradigm based exclusively on transportation issues, not those like network inventory management. Thus, future research may incorporate the PI's decisions, transport, and inventory. In addition, PI models may consider additional features of the PI. For example, cost and time may depend on other parameters such as shipping size, node driving times, capacity, speed limits, carrier limits. [28], [29] also affirm that more research is necessary to compare performance with optimal management decisions and procurement methods for distribution systems.

There is a need for future research to examine a network closer to a real-life case of connectable supply networks with more suppliers and employees. Some studies on how the inventory control models, such as operational models and decision-making tools, the analysis of management issues, and the adaptability of models to different systems are required. The development of awareness of business and organisational models used by the PI will significantly improve the network's vision and cooperation.

When several organisations, such as businesses, governments, and customers, are likely to use a common infrastructure, one of the future tasks will be to create specifications and architectures that will benefit each party in the logistics network. Therefore, we assume that criteria related to the conceptual architecture must be met to implement the PI, including the development of more operational processes, incentives, and reliable sharing of data.

The combination of blockchain and the PI will ensure trust, traceability, clarity, originality, and honesty. However, many studies are also required for the efficient integration of the technology in the PI environment. Scalability is one of the biggest challenges of blockchain. Scalability is the ability of a system to cope and perform well while increasing or expanding the network performance. The design of an overall logistics network necessitates high performance consistent with the large implementation of the PI. Perhaps the blockchain network's scalability and performance are unable to handle a sizeable interconnected PI network in its current situation.

Future-oriented logistical problems are highly prevalent. They include the excessive use of storage and transportation facilities, their costs and corresponding CO_2 reductions, the complexities of logistics activities, and increased process inefficiencies. Technologies like Big Data, the Internet of Things, 3D Printing, or other Industry 4.0 innovations have become part of logistics. Besides, to be competitive in the future market, organisations must deal with these developments.

VI. Conclusion

The PI has been a highly promising and evolving freight transport paradigm over the last few years. This paper systematically reviews the PI literature, which is published over the last decade. This review has shown that the PI is a global logistics and holistic concept that solves many problems associated with the current supply chain and logistics models. The growth in the PI literature and the capacity of this paradigm for disrupting supply chain and logistics techniques are examined, and several recommendations for future research are highlighted.

In this study, the latest state-of-the-art is reviewed and assessed. We categorise the literature according to three major classes: PI themes, decision levels, and sustainability dimensions. Next, we identify a range of research propositions based on the study's findings, which are worth exploring in future work to ensure logistics sustainability. The study findings show that the main decisions taken in the implementation of the PI are the tactical and operational decisions, with over 65% of selected publications addressing these two decision levels. Moreover, there is little emphasis on cooperation between external parties such as logistics service providers, competitors and customers. In addition, the emphasis is placed more on economic and environmental sustainability than on social sustainability.

This study proposes a novel coding framework to systematise the literature and outlines several future research directions for logistics and supply chain management scholars. Further PI research will likely lead to motivate the implementations of the PI and more development of technologies that will improve efficiency, performance and sustainable development of supply chains and logistics, thereby benefitting the standard of living of the society.

AUTHOR CONTRIBUTIONS

S. Ben Neila: Conceptualization, Experiments, Theoretical analysis.

A. Rejeb: Finite element modelling, Writing, Review and editing.

P. Németh: Supervision, Review and editing.

DISCLOSURE STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

A. Rejeb http://orcid.org/000-0002-2817-5377

P. Németh http://orcid.org/0000-0002-7825-6773

REFERENCES

[1] A. Scipioni, M. Mastrobuono, A. Mazzi, and A. Manzardo, 'Voluntary GHG management using a life cycle approach. A case study', J. Clean. Prod., vol. 18, no. 4, pp. 299–306, Mar. 2010. https://doi.org/10.1016/j.jclepro.2009.10.007

[2] J. Samuelson, The Six New Rules of Business: Creating Real Value in a Changing World. Berrett-Koehler Publishers, 2021.

[3] M. Henke and H. Kohl, Sustainability in Global Value Chains: Measures, Ethics and Best Practices for Responsible Businesses. Kogan Page Publishers, 2021.

[4] M. Plasch, S. Pfoser, M. Gerschberger, R. Gattringer, and O. Schauer, 'Why Collaborate in a Physical Internet Network?—Motives and Success Factors', J. Bus. Logist., vol. 42, no. 1, pp. 120–143, Mar. 2021. https://doi.org/10.1111/jbl.12260

[5] B. Montreuil, R. D. Meller, and E. Ballot, 'Physical Internet Foundations', IFAC Proc. Vol., vol. 45, no. 6, pp. 26–30, May 2012. https://doi.org/10.3182/20120523-3-RO-2023.00444

[6] X. Peng, S. Ji, and T. Ji, 'Promoting sustainability of the integrated production-inventory-distribution system through the Physical Internet', Int. J. Prod. Res., vol. 58, no. 22, pp. 6985–7004, Nov. 2020. https://doi.org/10.1080/00207543.2019.1687953

[7] H. R. Hasan, K. Salah, R. Jayaraman, I. Yaqoob, and M. Omar, 'Blockchain Architectures for Physical Internet: A Vision, Features, Requirements, and Applications', IEEE Netw., vol. 35, no. 2, pp. 174–181, Mar. 2021. https://doi.org/10.1109/MNET.021.2000442

[8] B. Montreuil, E. Ballot, and W. Tremblay, 'Modular Design of Physical Internet Transport, Handling and Packaging Containers', p. 30.

[9] B. Montreuil, 'Toward a Physical Internet: meeting the global logistics sustainability grand challenge', Logist. Res., vol. 3, no. 2, pp. 71–87, May 2011. https://doi.org/10.1007/s12159-011-0045-x

[10] H. Treiblmaier, K. Mirkovski, and P. B. Lowry, 'Conceptualizing the Physical Internet: Literature Review, Implications and Directions for Future Research', Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2861409, May 2016. Accessed: Jul. 07, 2021. https://papers.ssrn.com/abstract=2861409

[11] T. Ambra, A. Caris, and C. Macharis, 'Towards freight transport system unification: reviewing and combining the advancements in the physical internet and synchromodal transport research', Int. J. Prod. Res., vol. 57, no. 6, pp. 1606–1623, Mar. 2019. https://doi.org/10.1080/00207543.2018.1494392

[12] H. Treiblmaier, K. Mirkovski, P. B. Lowry, and Z. G. Zacharia, 'The physical internet as a new supply chain paradigm: a systematic literature review and a comprehensive framework', Int. J. Logist. Manag., vol. 31, no. 2, pp. 239–287, May 2020. https://doi.org/10.1108/IJLM-11-2018-0284

[13] K. Jharni, M. Hlyal, and J. El Alami, 'Physical Internet: Literature review and research opportunities', IOP Conf. Ser. Mater. Sci. Eng., vol. 827, p. 012044, Jun. 2020. https://doi.org/10.1088/1757-899X/827/1/012044

[14] B. Montreuil, R. D. Meller, and E. Ballot, 'Physical Internet Foundations', in Service Orientation in Holonic and Multi Agent Manufacturing and Robotics, T. Borangiu, A. Thomas, and D. Trentesaux, Eds. Berlin, Heidelberg: Springer, 2013, pp. 151–166. https://doi.org/10.1007/978-3-642-35852-4_10

[15] E. Ballot, B. Montreuil, and C. Thivierge, 'Functional Design of Physical Internet Facilities: A Road-Rail Hub', p. 36.

[16] D. Tranfield, D. Denyer, and P. Smart, 'Towards a Methodology for Developing Evidence-Informed Management Knowledge

by Means of Systematic Review', Br. J. Manag., vol. 14, no. 3, pp. 207–222, Sep. 2003. https://doi.org/10.1111/1467-8551.00375
[17] B. Montreuil, J.-F. Rougès, Y. Cimon, and D. Poulin, 'The Physical Internet and Business Model Innovation', Technol. Innov. Manag. Rev., p. 6, 2012.
[18] R. Sarraj, E. Ballot, S. Pan, and B. Montreuil, 'Analogies between Internet network and logistics service networks: challenges involved in the interconnection', J. Intell. Manuf., vol. 25, no. 6, pp. 1207–1219, Dec. 2014. https://doi.org/10.1007/s10845-012-0697-7
[19] C. Dong and R. Franklin, 'From the Digital Internet to the Physical Internet: A Conceptual Framework With a Stylized Network Model', J. Bus. Logist., vol. 42, no. 1, pp. 108–119, Mar. 2021. https://doi.org/10.1111/jbl.12253
[20] H. Sternberg and A. Norrman, 'The Physical Internet – review, analysis and future research agenda', Int. J. Phys. Distrib. Logist. Manag., vol. 47, no. 8, pp. 736–762, Sep. 2017. https://doi.org/10.1108/IJPDLM-12-2016-0353
[21] W. Hofman, M. Punter, H. Bastiaansen, E. Cornelisse, and S. Dalmolen, 'Semantic technology for enabling logistics innovations – towards Intelligent Cargo in the Physical Internet', Int. J. Adv. Logist., pp. 1–12, Apr. 2016. https://doi.org/10.1080/2287108X.2016.1144358
[22] C. Landschützer, F. Ehrentraut, and D. Jodin, 'Containers for the Physical Internet: requirements and engineering design related to FMCG logistics', Logist. Res., vol. 8, no. 1, p. 8, Dec. 2015. https://doi.org/10.1007/s12159-015-0126-3
[23] Y.-H. Lin, R. D. Meller, K. P. Ellis, L. M. Thomas, and B. J. Lombardi, 'A decomposition-based approach for the selection of standardized modular containers', Int. J. Prod. Res., vol. 52, no. 15, pp. 4660–4672, Aug. 2014. https://doi.org/10.1080/00207543.2014.883468
[24] H. S. Sternberg and M. Denizel, 'Toward the Physical Internet—Logistics Service Modularity and Design Implications', J. Bus. Logist., vol. 42, no. 1, pp. 144–166, Mar. 2021. https://doi.org/10.1111/jbl.12261
[25] A. Krogsgaard, D. Pisinger, and J. Thorsen, 'A flow-first route-next heuristic for liner shipping network design', Networks, vol. 72, no. 3, pp. 358–381, Oct. 2018. https://doi.org/10.1002/net.21819
[26] C. Fergani, A. El Bouzekri El Idrissi, S. Marcotte, and A. Hajjaji, 'Optimization of hyperconnected mobile modular production toward environmental and economic sustainability', Environ. Sci. Pollut. Res., vol. 27, no. 31, pp. 39241–39252, Nov. 2020. https://doi.org/10.1007/s11356-020-09966-9
[27] S. Ji, X. Peng, and R. Luo, 'An integrated model for the production-inventory-distribution problem in the Physical Internet', Int. J. Prod. Res., vol. 57, no. 4, pp. 1000–1017, Feb. 2019. https://doi.org/10.1080/00207543.2018.1497818
[28] Y. Yang, S. Pan, and E. Ballot, 'Innovative vendor-managed inventory strategy exploiting interconnected logistics services in the Physical Internet', Int. J. Prod. Res., vol. 55, no. 9, pp. 2685–2702, May 2017. https://doi.org/10.1080/00207543.2016.1275871.
[29] Y. Yang, S. Pan, and E. Ballot, 'Mitigating supply chain disruptions through interconnected logistics services in the Physical Internet', Int. J. Prod. Res., vol. 55, no. 14, pp. 3970–3983, Jul. 2017. https://doi.org/10.1080/00207543.2016.1223379
[30] M. Darvish, H. Larrain, and L. C. Coelho, 'A dynamic multi-plant lot-sizing and distribution problem', Int. J. Prod. Res., vol. 54, no. 22, pp. 6707–6717, Nov. 2016. https://doi.org/10.1080/00207543.2016.1154623
[31] S. Pan, M. Nigrelli, E. Ballot, and R. Sarraj, 'PERFORMANCE ASSESSMENT OF DISTRIBUTED INVENTORY IN PHYSICAL INTERNET', p. 17, 2013.
[32] S. Pan, M. Nigrelli, E. Ballot, R. Sarraj, and Y. Yang, 'Perspectives of inventory control models in the Physical Internet: A simulation study', Comput. Ind. Eng., vol. 84, pp. 122–132, Jun. 2015. https://doi.org/10.1016/j.cie.2014.11.027
[33] T. Chargui, A. Bekrar, M. Reghioui, and D. Trentesaux, 'Proposal of a multi-agent model for the sustainable truck scheduling and containers grouping problem in a Road-Rail physical internet hub', Int. J. Prod. Res., vol. 58, no. 18, pp. 5477–5501, Sep. 2020. https://doi.org/10.1080/00207543.2019.1660825
[34] S. Gontara, A. Boufaied, and O. Korbaa, 'Routing the Pi-Containers in the Physical Internet using the PI-BGP Protocol', in 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), Aqaba, Oct. 2018, pp. 1–8. https://doi.org/10.1109/AICCSA.2018.8612885
[35] M. Fazili, U. Venkatadri, P. Cyrus, and M. Tajbakhsh, 'Physical Internet, conventional and hybrid logistic systems: a routing

optimisation-based comparison using the Eastern Canada road network case study', Int. J. Prod. Res., vol. 55, no. 9, pp. 2703–2730, May 2017. https://doi.org/10.1080/00207543.2017.1285075

[36] H. Tran-Dang, N. Krommenacker, and P. Charpentier, 'Containers monitoring through the Physical Internet: a spatial 3D model based on wireless sensor networks', Int. J. Prod. Res., vol. 55, no. 9, pp. 2650–2663, May 2017. https://doi.org/10.1080/00207543.2016.1206220

[37] U. Venkatadri, K. S. Krishna, and M. A. Ulku, 'On Physical Internet Logistics: Modeling the Impact of Consolidation on Transportation and Inventory Costs', IEEE Trans. Autom. Sci. Eng., vol. 13, no. 4, pp. 1517–1527, Oct. 2016. https://doi.org/10.1109/TASE.2016.2590823

[38] F. Walha, A. Bekrar, S. Chaabane, and T. M. Loukil, 'A railroad PI-hub allocation problem: Active and reactive approaches', Comput. Ind., vol. 81, pp. 138–151, Sep. 2016. https://doi.org/10.1016/j.compind.2016.04.007

[39] J.-Y. Colin, M. Nakechbandi, and H. Mathieu, 'Management of mobile resources in Physical Internet logistic models', in 2015 4th International Conference on Advanced Logistics and Transport (ICALT), Valenciennes, France, May 2015, pp. 323–326. https://doi.org/10.1109/ICAdLT.2015.7136598

[40] B. Montreuil, R. D. Meller, C. Thivierge, and Z. Montreuil, 'Functional Design of Physical Internet Facilities: A Unimodal Road-Based Crossdocking Hub', p. 55.

[41] S. Kaup, A. Ludwig, and B. Franczyk, 'Design and Evaluation of Routing Artifacts as a Part of the Physical Internet Framework', p. 16.

[42] M. Klumpp and H. Zijm, 'Logistics Innovation and Social Sustainability: How to Prevent an Artificial Divide in Human–Computer Interaction', J. Bus. Logist., vol. 40, no. 3, pp. 265–278, Sep. 2019. https://doi.org/10.1111/jbl.12198

[43] R. Y. Zhong, H. Gong, C. Xu, and S. Lu, 'Physical Internet-Enabled Manufacturing Execution System for Intelligent Workshop Production', Int. J. Signal Process. Image Process. Pattern Recognit., vol. 9, no. 6, pp. 121–132, Jun. 2016. https://doi.org/10.14257/ijsip.2016.9.6.11

[44] R. Y. Zhong, C. Xu, C. Chen, and G. Q. Huang, 'Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors', Int. J. Prod. Res., vol. 55, no. 9, pp. 2610–2621, May 2017. https://doi.org/10.1080/00207543.2015.1086037

[45] V. Rodríguez Cornejo, Á. Cervera Paz, L. López Molina, and V. Pérez-Fernández, 'Lean Thinking to Foster the Transition from Traditional Logistics to the Physical Internet', Sustainability, vol. 12, no. 15, p. 6053, Jul. 2020. https://doi.org/10.3390/su12156053

[46] B. Qiao, S. Pan, and E. Ballot, 'Revenue optimization for less-than-truckload carriers in the Physical Internet: dynamic pricing and request selection', Comput. Ind. Eng., vol. 139, p. 105563, Jan. 2020. https://doi.org/10.1016/j.cie.2018.12.010

[47] B. Qiao, S. Pan, and E. Ballot, 'Dynamic pricing model for less-than-truckload carriers in the Physical Internet', J. Intell. Manuf., vol. 30, no. 7, pp. 2631–2643, Oct. 2019. https://doi.org/10.1007/s10845-016-1289-8

[48] X. T. R. Kong, J. Chen, H. Luo, and G. Q. Huang, 'Scheduling at an auction logistics centre with physical internet', Int. J. Prod. Res., vol. 54, no. 9, pp. 2670–2690, May 2016. https://doi.org/10.1080/00207543.2015.1117149

[49] I. B. Othmane, M. Rekik, and S. Mellouli, 'Impact of Shipper Collaboration on Carriers Selection in Reputation-Based Transportation Auctions', p. 9.

[50] S. Pan, X. Xu, and E. Ballot, 'Auction based transport services allocation in Physical Internet: a simulation framework', p. 13.

[51] D. Kubek and P. Więcek, 'An integrated multilayer decision-making framework in the Physical Internet concept for the City Logistics', Transp. Res. Procedia, vol. 39, pp. 221–230, 2019. https://doi.org/10.1016/j.trpro.2019.06.02

[52] L. Zheng, 'Assessment of the physical internet enabled urban logistics using agent-based simulation', p. 26.

[53] I. Ben Mohamed, W. Klibi, O. Labarthe, J.-C. Deschamps, and M. Z. Babai, 'Modelling and solution approaches for the interconnected city logistics', Int. J. Prod. Res., vol. 55, no. 9, pp. 2664–2684, May 2017. https://doi.org/10.1080/00207543.2016.1267412

[54] C. Chen, S. Pan, Z. Wang, and R. Y. Zhong, 'Using taxis to collect citywide E-commerce reverse flows: a crowdsourcing solution', Int. J. Prod. Res., vol. 55, no. 7, pp. 1833–1844, Apr. 2017. https://doi.org/10.1080/00207543.2016.1173258

[55] Matusiewicz, 'Logistics of the Future—Physical Internet and Its Practicality', Transp. J., vol. 59, no. 2, p. 200, 2020. https://doi.org/10.5325/transportationj.59.2.0200

[56] T. G. Crainic and B. Montreuil, 'Physical Internet Enabled Hyperconnected City Logistics', Transp. Res. Procedia, vol. 12, pp. 383–398, 2016. https://doi.org/10.1016/j.trpro.2016.02.074
[57] K. Chen, G. Xu, F. Xue, R. Y. Zhong, D. Liu, and W. Lu, 'A Physical Internet-enabled Building Information Modelling System for prefabricated construction', Int. J. Comput. Integr. Manuf., vol. 31, no. 4–5, pp. 349–361, Apr. 2018. https://doi.org/10.1080/0951192X.2017.1379095
[58] Y. Zhang, S. Liu, Y. Liu, and R. Li, 'Smart box-enabled product–service system for cloud logistics', Int. J. Prod. Res., vol. 54, no. 22, pp. 6693–6706, Nov. 2016. https://doi.org/10.1080/00207543.2015.1134840
[59] N. Lemmens, J. Gijsbrechts, and R. Boute, 'Synchromodality in the Physical Internet – dual sourcing and real-time switching between transport modes', Eur. Transp. Res. Rev., vol. 11, no. 1, p. 19, Dec. 2019. https://doi.org/10.1186/s12544-019-0357-5
[60] B. van Riessen, R. R. Negenborn, and R. Dekker, 'Real-time container transport planning with decision trees based on offline obtained optimal solutions', Decis. Support Syst., vol. 89, pp. 1–16, Sep. 2016. https://doi.org/10.1016/j.dss.2016.06.004
[61] Q. Betti, R. Khoury, S. Hallé, and B. Montreuil, 'Improving Hyperconnected Logistics with Blockchains and Smart Contracts', ArXiv190403633 Cs, Apr. 2019, Accessed: Jul. 07, 2021. http://arxiv.org/abs/1904.03633
[62] T. Meyer, M. Kuhn, and E. Hartmann, 'Blockchain technology enabling the Physical Internet: A synergetic application framework', Comput. Ind. Eng., vol. 136, pp. 5–17, Oct. 2019. https://doi.org/10.1016/j.cie.2019.07.006
[63] H. Treiblmaier, 'Combining Blockchain Technology and the Physical Internet to Achieve Triple Bottom Line Sustainability: A Comprehensive Research Agenda for Modern Logistics and Supply Chain Management', Logistics, vol. 3, no. 1, p. 10, Feb. 2019. https://doi.org/10.3390/logistics3010010
[64] L. Simmer, S. Pfoser, M. Grabner, O. Schauer, and L. M. Putz, 'From horizontal collaboration to the physical Internet – A case study from austria', Int. J. Transp. Dev. Integr., vol. 1, no. 2, pp. 129–136, Jan. 2017. https://doi.org/10.2495/TDI-V1-N2-129-136
[65] T. G. Crainic and G. Laporte, 'Planning models for freight transportation', Eur. J. Oper. Res., vol. 97, no. 3, pp. 409–438, Mar. 1997. https://doi.org/10.1016/S0377-2217(96)00298-6
[66] T. G. Crainic and K. H. Kim, 'Chapter 8 Intermodal Transportation', in Handbooks in Operations Research and Management Science, vol. 14, C. Barnhart and G. Laporte, Eds. Elsevier, 2007, pp. 467–537. https://doi.org/10.1016/S0927-0507(06)14008-6.bb

Acta Technica Jaurinensis
Vol. 15, No. 1, pp. 36-46, 2022
10.14513/actatechjaur.00642

Certification of Unmanned Aircraft Systems – from product safety to type certificate – a review about the operation of the EU safeguard processes

Zs. Sándor[1,*], M. Pusztai[1]

[1]KTI Institute for Transport Sciences Non-profit Ltd.
Than Károly street 3-5, H-1119 Budapest, Hungary
*e-mail: sandor.zsolt@kti.hu

Submitted: 09/08/2021 Accepted: 04/10/2021 Published online: 18/10/2021

Abstract: Significant changes are emerging in the market of unmanned aircraft systems since 2019 through the publication of two specific regulations that regulate all steps of the use of unmanned aerial vehicles in detail. With the implementation of the new EU drone regulations, the role of the notified bodies and the certification agencies will be more important from the viewpoint of product safety and the official certification required by the EU and national aviation authorities. The product safety chain consists of two major parts. One part belongs to the production phase, where the manufacturer has to prove the functionality (it is called the conformity assessment) and another part belongs to the distribution market, where the authorities assess the fulfilment of the conditions of the distribution (it is called the market surveillance). The first pillar concern to the design and manufacturing and the second to the distribution. Each segment is presented in this article and the authors introduce the different control approaches of these segments. It has to be taken into consideration that the drones are representing a special market with notable safety risks that have to be handled during the whole life path of the products from the design through the distribution until the aerial operations.

Keywords: certification; unmanned aircraft systems; product safety; safeguard; conformity assessment; market surveillance

I. Introduction

Unmanned Aerial Vehicles (UAVs) play a significant role in the industry of Europe. With this equipment, several human-related resources can be substituted. The global market has already identified it, thus the industry of drones is prospering. In 2020 the market of UAVs was valued at 30601.14 million USD, and researchers projected its compound annual growth rate to 8.5% for a 5 year forecast period between 2021 and 2026 [1].

Europe has a prominent position in this market. The European unmanned aerial vehicle market accounted for 2637.2 million USD in 2019 and will grow at ~10% annually between 2020 and 2026. Within 20 years, the European drone sector is expected to directly employ more than 100,000 people and it will have an economic impact exceeding 10 billion EUR per year, mainly in services [2].

These numbers are justifying the fact that rule makers and national decision-makers should handle the topic and the global European approach should be developed from the design through the manufacture until the use of these products.

According to the Treaty on the Functioning of the European Union, member states and EU institutions shall take all measures necessary for the creation of the single market where borders shall mean no obstacle to the free movement of products (i.e., goods) [3]. To this aim, it is vital, however, that products moving without barriers are safe.

Due to technical development, many products, that are easy to access but are very complicated and pose risks even at ordinary use, appeared in everyday lives. For an average person - called consumer (the term of the European law) - cannot be even legally expected the understanding of these products' technical background. The legislator shall be satisfied once the consumer undertakes the burden of learning the ways of the product's normal use and the inherent risks thereof (i.e., reading the user's manual). Every other task shall be fulfilled by the manufacturer or by other players in the supply chain (e.g., the importer and/or the distributor). This applies to all branches of the industry as well as to Unmanned Aircraft Systems (UAS) that constitute the topic of the article.

These circumstances called the product safety rules to life, which - through the (EU) 2019/945 Delegated Regulation (hereinafter DR) of the European Commission - created specific provisions in connection with UASs [4].

To provide fulfilling these tasks by the manufacturers properly and that level of product safety is sufficient in case of UAS intended to be operated in "open" category as provided for in Commission Implementing Regulation (EU) 2019/947 (hereinafter IR), the European legislator - simultaneously with the supply chain - created the product safety chain, including roles and competencies therein [5]. Players of the product safety chain are available in adequate phases of the supply chain and are capable of intervening in precisely defined stages. The two most important actors of the product safety chain are the conformity assessment bodies (i. e., NOtified BOdy, hereinafter NOBO) and authority for market surveillance.

The concept of product safety chain does not appear in literature and relevant laws. The other players of the supply chain (manufacturer, importer, distributor) have some obligations in connection with product safety however, these obligations are not characterizing. Meanwhile, the two pillars of the product safety chain (conformity assessment and market surveillance) deal only with product safety, and thanks to their strict regulations they are completely independent of every other actor of the supply chain. The concept of product safety chain has been elaborated to envisage the different roles and competencies and the legal, organizational, and economic independence and separation. It constitutes another ground for this separation that the two pillars of the product safety chain make an external audit of the supply chain in different stages of products therein.

In "specific" and "certified" operational categories as provided for in IR, the European legislator went further and stepped over the logic of product safety, and is approaching the regulation on certification of regular aircraft in terms of regulatory structure and mind-set.

In the present article, the authors are intending to show the legislative solutions of each operational category and the regulatory "arch" linking them together into a single system. The product safety chain consists of two pillars. The conformity assessment is presented in Section II and the market surveillance, which is the second pillar is presented in Section III. The operations of UASs are regularized in detail but until the product safety chain is not operating completely due to the lack of legislative gaps, some transitional provisions are in place. Section IV. deals with these provisions. Section V. presents the operational issues, while Section VI. gives a summary of the conventional certification process connected to UASs.

II. THE CONFORMITY ASSESSMENT - THE FIRST PILLAR OF THE PRODUCT SAFETY CHAIN

Air transport is a dangerous operation field where safety plays a key role. This safety might be achieved in different ways: training of experts in the industry, use of generally accepted directives/standards/regulatory means, and control of equipment throughout the whole cycle of their life from design to disassembly.

In case of UASs intended to be used in the "open" operational category as provided for in IR, the legislator warrants safety via conformity assessment of UAS, which is the first step of product safety and focuses on design, organization of production, and control of the manufacturer's activity thereby defending consumers from risks.

The product safety and market surveillance approach imply that the Delegated Act (DA) considers UAS with "CE marking" – Certificate of Compliance – and class identification label primarily as consumer goods. The "CE marking" is indeed a certificate of compliance, but it is not the abbreviation of "certificate of compliance". It is a unique marking which should proof that the product complies with defined standards of the European Economic Area. To summarize, as these products are aerial vehicles at the same time, it is possible to define them as consumer goods that are capable of flying.

It is important to set forth: this shall not mean that - in terms of flight safety - there would be any difference between drones operated in the "open" category or Standard Scenarios (hereinafter STS) as in IR and the unmanned aircraft operated in the stricter "specific" or "certified" operational categories [3]. On the contrary, the legislator achieves a sufficient level of air safety via product safety/consumer protection/market surveillance regulations in operational categories having lower risk (like the "open" category and STS).

There is another aspect behind choosing conformity assessment as a means of certification: the introduction of certification process applied for regular aircraft would result in disproportionately severe technical and financial burden on manufacturers' side in case of UAS having a Maximum Take-Off Weight (MTOW) of 25 kg and being used in a very limited operational environment. This could prevent many manufacturers from entering the market and could lead to a regression of the drone market in the long run.

Therefore, the DR and European Aviation Safety Agency (EASA) created a system, where UAS with MTOW no more than 25 kg - in case if they are targeted to be operated in the "open" category or STS as provided for in IR - are treated as products and the risk they pose by appearing in air traffic is handled with the introduction of UAS-specific product safety rules.

The first pillar is in this system is the conformity assessment, which as a regulatory technique is common since the 765/2008/EC Regulation that is completed by the 768/2008/EC Decision, that defines so-called conformity assessment modules. The conformity assessment has already been mandatory for drones before the DR, upon completely different standards and regulations, however [6], [7].

These modules are general and not limited to certain branches. When the European Union wishes to introduce conformity assessment on a certain new field of law as a regulatory means, however, it will define in the specific legal act - like DR - which conformity assessment modules it will make available in the specific branch. The novelty of the DR is that it constituted single and specifically UAS-tailored conformity assessment rules.

Concerning the topic of the present article, these modules shall be distinguished upon whether the NOBO is required during conformity assessment or the manufacturers themselves - undertaking all liability - can conduct the procedure, and the market surveillance authority shall supervise appropriateness of the assessment subsequently.

Conformity assessment might be performed according to the modules thereof, but these different modules are made available for actors of the supply chain in case of different sorts of products within the realm of UASs. The Annex of DR lists seven different Unmanned Aircraft Classes (UA Class) according to the technical requirements and capabilities thereof. The UA Class determines at the same time; which operational category the drone may fly. **Table 1** shows the conformity assessment modules applicable to different UA Classes. It is clear from the chart that Module A is not an option for the manufacturer in the case of UA Classes C1 to C3 and the NOBO shall be part of the process in these cases.

Table 1. *Modules in conformity assessment of unmanned aerial vehicles*

Conformity Assessment Module	***Content of the Module***	***UA Class***	***Is a Notified Body (NOBO) required?***
A	Internal production control	C0, C4-C6	none
B, C	EU-type examination, Conformity to type based on internal production control	C0-C6	yes
H	Conformity based on full quality assurance	C0-C6	yes

1. Conformity assessment upon harmonized standards

In the case of drones, Module A, the internal product control might be chosen by the manufacturers to conduct conformity assessments on their own. This module applies exclusively to UA Class C0 and C4 to C6 at the same time. It would be applicable: the harmonized standards required for the proper conduct of the procedure have not been completed yet. These new standards will constitute the so-called EN 4709 standards and **Table 2** lists the elements thereof [8].

It is clear from the naming that the DR introduced new technical requirements, so new drone types - sold to operate in the "open" category or STS - shall meet these new requirements before being launched on the market.

The new harmonized standards are so important because the application thereof fastens the conformity assessment procedure and market surveillance. The DR provides the presumption of conformity to these standards, which means that - if the manufacturer refers thereto - it shall not be investigated during the conformity assessment procedure whether the manufacturer chose the adequate technical process to achieve conformity, but only the result i.e., the fulfilment of technical requirement prescribed by law (e.g., DR) shall be verified and checked.

Table 2. *The elements of EN 4709 standards*

Name of the (sub)standard	***Content of the (sub)standard***
EN 4709-1	Harmonized standard on product and verification requirements for Unmanned Aircraft Systems intended to be operated in the “Open” category
EN 4709-2	Harmonized standard on direct remote identification for Unmanned Aircraft Systems
EN 4709-3	Harmonized standard on geo-awareness for Unmanned Aircraft Systems
EN 4709-4	Harmonized standard on lights for Unmanned Aircraft

2. Conformity assessment with NOBO

The DR provides exclusively B+C Modules for UA Class C1 to C3. In Module B (EU-type examination), the intervention of NOBO is mandatory. In the frame of this module, the NOBO shall issue an EU-type examination certificate following the evaluation of the manufacturer’s technical documentation and testing of already produced specimens. The manufacturer may only proceed with conformity assessment with Module C in possession of the EU-type examination certificate and later may commence the placing on the market of the drone.

It is the European Commission that grants the status of being notified body to NOBOs and records them, following the handover of documentation by the competent notifying authority (this is the Civil Aviation Authority (CAA) in Hungary). The EU-type examination certificate, issued by NOBO shall provide the opportunity of trading throughout the European Union.

The NOBOs may intervene in conformity assessment during the application of Module H too. In the case of this module, the NOBO performs full quality assurance and evaluates not the specific product but the production. Therefore, this is not a one-time but a regular inspection on the manufacturer’s site.

Until now, no conformity assessment body was recorded on the NANDO (New Approach Notified and Designated Organisations) list of the European Commission where these bodies are publicly available [9]. Regarding that notification shall be preceded by accreditation and the process itself before the European Commission takes at least two months - if no notifying authority from member states questions the notification -, the appearance of NOBOs is yet to come.

It is another important circumstance for drones with UA Class C5 or C6 that operational rules for STS shall enter into effect only in 2023. This means, that these kinds of operations are not permitted before that day, even if UAS with C5 or C6 class identification label existed on the market.

The EASA realized the indefensibility of this situation (i.e., there is no UAS equipped with class identification label) and created PDRA-S01 and PDRA-S02 within the framework of Pre-Defined Risk Assessment (PDRA)-s, where - apart from the obligation of using drones with C5 or C6 class identification label - every circumstance is identical with the ones in STS operations. This way the acquisition of operational authorization is still required however, the process has been facilitated as the EASA had defined the basis of risk assessment in advance.

3. Conformity assessment of privately built UAS

In the course of establishing the IR, the European Commission so took into consideration that there were many vehicles in the European sky that had been custom-designed and -built by their operators. This further means that the technical conformity and airworthiness of these vehicles have never been controlled by an independent entity. Accordingly, the privately built UAS has no manufacturer but a builder, which further implies that the latter shall not undertake the obligation that a manufacturer does. This alleviation has its price at the same time: the IR introduced several restrictions on the operation of privately built UAS in the “open” category. These restrictions are the following:

- privately built UAS may only be operated in A1 subcategory (MTOW < 0.25 kg) or in A3 subcategory (MTOW < 25 kg); and
- exclusively for the builder's use.

Especially, the latter provision that can be derived from the concept of privately built UAS is interesting and means that such UAS is not marketable, it may not be sold, otherwise the term “for the builder’s use” would lose its meaning.

These restrictions limit the operational opportunities of these privately built vehicles substantially, but at least the performance of certain operations remains feasible.

The abovementioned limitations apply to operations in the “open” category, but there is no legal obstacle to the use of privately built UAS in the "Specific" category. In the latter case, the process in Paragraph 4 of the present article is applicable, the technical conformity and airworthiness of these UASs shall be determined by the risk-mitigating measures defined in the operational authorization or Light UAS operator's Certificate (LUC) issued by the competent CAA. Presumably, competent CAAs will prescribe many tests and certification obligations before issuing operation authorization or LUC.

III. Market surveillance authority as the second pillar of the product safety chain

As the second pillar of product safety, the market surveillance is a bigger unit, which means a formal- and content-based assessment, similar to the conformity assessment (see especially Module B). In the literature of the European Union, the reference for market surveillance is indicated as a compliance assessment [10]. There is a difference between the compliance and the conformity assessment, but both pillars contain assessment procedures that indicate the close relationship between the two pillars.

According to the fact that the first unit of the product safety chain (the conformity assessment, as described in the second paragraph of this article) is not yet able to fulfill its function, the market surveillance authorities have even more significant roles until the appearance of the new standards and the NOBOs.

Although the IR applies only to drones belonging to the UA Class C0-C6 that already went through the conformity assessment, this does not mean that the market surveillance authorities would not have any duty until the appearance of new drones.

According to the fact described in Section II.2 of this article, currently, all drones, which are being placed on the market or made available thereon with a class identification label, got these identification labels illegally and they might mislead the consumers. Thus, it may purport that the given drone can be legally used for the UAS operations according to the class identification label.

However, it is not true, because the IR limits the use of drones that do not have a class identification label in the "open" category, and the A2 subcategory is excluded from the available possibilities. (Article 20 and 22 of the IR. It has to be noted that this limitation is not absolute. If the UAS operator has an operational authorization, it is possible to execute an operation that has the characteristics of the A2 subcategory operation with a UAS that does not have a class identification label.)

Thus, the market surveillance authority is entitled to conduct inspections and audits, not only at the manufacturers' but also at the importers', distributors', and the dealers' sites. In case of infringement, the market surveillance authority is entitled to apply the whole Hungarian and EU toolkit of market surveillance laws, which contains even the recall of the product from the entire EU market besides the infliction of fines and penalties.

The market surveillance will also have a significant role in the future because the manufacturers shall keep the technical and compliance assessment documents of the drone that already went through the conformity assessment. These documents should be presented upon request of the national market surveillance authority for 10 years after the product has been placed on the Union market.

It is another important provision that the manufacturer of a drone with a C5 or a C6 class identification label should notify the market surveillance authority about any new type in these UA Classes before placing the product on the market. The market surveillance authority is entitled to ask for documents from the NOBOs about the conformity assessments of drones done by NOBOs. According to relevant legal regulations, it can be stated that the NOBOs and the market surveillance authority constitute one single system. As part of the product safety chain, they are operating jointly, not isolated. This operating method validates the use of the pillar metaphor.

Let us take an exaggerated, but still very clear example: while the conformity assessment functions as a sort of type-certification, the market surveillance should monitor the continuous airworthiness by the use of test purchase and labour tests. These provide additional external control in the further phase of the product safety and supply chain during the conformity assessments.

This is especially important in the case of Module A, where the manufacturer conducts the assessment individually. The EU would execute these assessments in an institutional form in the latter, by the establishment of union test sites [11]. These assessments with these sites would strengthen the industry and raise it above the level of the Member States, thereby creating a single internal market.

IV. Transitional provisions of the Implementing Regulation

The legal and organizational (more correctly the standardization) background of the conformity assessment is not yet available. To handle these situations and to provide a sufficient preparation period, the IR has defined certain transitional provisions. It is important to emphasize, that these transitional provisions are concerning the operations of the „open" category only. In case of operational authorization or with a LUC the UAS operator may use its UAS differently from these rules.

According to Article 20 of the IR following its entry into force, but not later than 31st December 2022, it is possible to place UASs on the market that is intended to be operated in the "open" category and are not in compliance with all requirements of the DR - consequently do not have a Class Identification Label (CIL).

As provided for in the IR, the operational possibilities of these UASs are limited in the "open" category too. Without CIL, the A1 subcategory is available only to new UASs with MTOW less than 500 grams, provided that the member state has elaborated competence-requirements (other than UAS.OPEN.020) to these operations (As Hungary created no such requirements, only drones under 250 grams can be operated in A1 subcategory without CIL in Hungary.). UASs newly placed on the market with MTOW at least 500 grams but less than 2 kg can be used only in subcategory A2 without CIL, provided that the remote pilot has a competence equivalent to the one in UAS.OPEN.030 [5] and the drone maintains a minimum of 50-meter distance from other people.

UAS having MTOW of at least 2 kg but under 25 kg can be operated in the A3 subcategory without a class identification label. Once the above transitional period is over only those UAS can be operated without CIL in the "open" category (A1 with MTOW under 250 grams and A3 with MTOW under 25 kg) that has been placed on the market before 1st January 2023. UAS operators may differ from the provisions only in the possession of an operational authorization or a LUC.

UAS without a class identification label may be placed on the market after the 1st January 2023 as a new type too, but it can be operated only in the "special" category, thus an operational authorization or a LUC will be necessary for the fulfilment of the operations. **Fig. 1** will show the transitional period and its relevant deadlines [12].

It means that considering these dates the actors of the supply chain should inform the buyers about the operational limits regarding the operations. The fulfillment of the information shall be inspected according to Article 39 of the DR by the market surveillance authorities of the Member States.

The situation is complicated because the IR defines deadlines to the act of placing on the market, not to act of making available on the market (Placing on the market means the first making available of a product on the Union market – according to the IR). If placing on the market is performed until the indicated deadline, the product can be distributed in the EU, until the manufactured stock runs out, without any obligation to acquire an operational authorization or a LUC for the operation of the given products.

Based on the prognosis of the EASA the potential buyers will be able to buy UASs for "open" category operations without class identification labels due to the fact these kinds of products will be available until 2026 at the merchants [13].

The manufacturers should not have to start the planning of the products from the beginning to

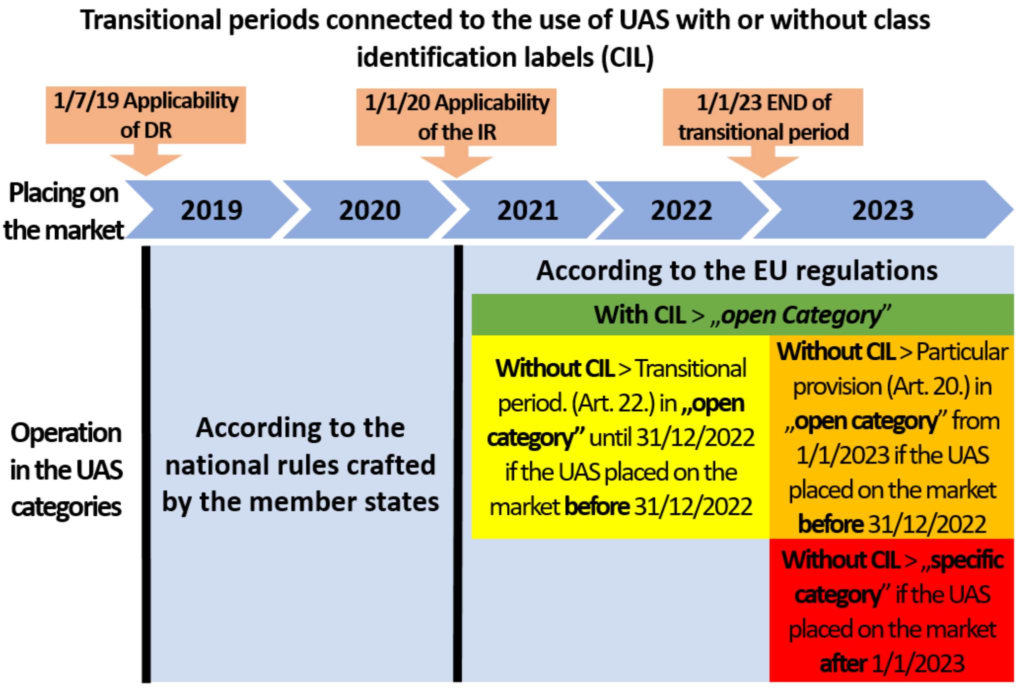


Figure 1. *Transitional period for the use of UAS without CIL*

comply with the rules and regulations. If a model is fully compliant with the new rules and regulations after an upgrade or a retrofit, then the retrofit or the upgrade will be the basis of the conformity assessment. According to the regulations, the upgraded or retrofitted models are new models, thus they will be granted with new serial numbers - irrespective of the scale of the modification.

Market surveillance authorities will have a significant role in the additional inspections of the new products that appear in the market, because according to the plans of the EASA, the operator may place the new identification labels on the UAS following the manufacturer's instructions after the controlled upgrade [13]. The administrational background of these inspections could be facilitated if the registration of the UASs were compulsory. This process can track the lifecycle of each UAS. Almost only Hungary has this kind of registration in the EU (the registration has a wider scope in France than the minimum requirement of the EU).

It is not obvious at the same time, why the European legislator considers drones without class identification label as meaning lower hazard to flight safety during the transitional period than after it. This applies especially to transitional limitations in subcategory A2, where the limitations (in terms of MTOW and distance) are supposed to balance the risk posed by the lack of conformity to DR and CIL. This provision outlined in Article 20 and 22 of the IR might serve as a legislative means to force manufacturers to provide upgrade or retrofit opportunities to operators of elder UASs - or to make UAS-operators buy new drones.

V. The authorization of the operations in the "specific" category - risk assessment and design verification report

1. Risk assessment

The UAS operator shall acquire an operational authorization if the intended operation may not be executed in the "open" category or according to STS. The UAS operator shall submit a detailed risk assessment to the competent authority (except if the operator executes the operations based on a PDRA) [5]. It is the requirement of the acquisition of the operational authorization.

In the course of risk assessment, the Specific Assurance and Integrity Level (SAIL) is determined based on the environmental characteristics of the operational area (characteristics of the UAS, location of the operational area, airspace characteristics above the operational area, etc.). SAIL comprises levels from I to VI and a given level defines the degree of the risk of a certain operation in a complex way. The SAIL levels identify to which level the robustness of the given Operational Safety Objectives (OSOs) shall be assured in a specific operation [14].

The OSOs are the sum of all risk mitigation activities and safety barriers, which provide the mitigation of the effects of possible risks that might influence the operation. So far, 24 OSOs have been defined by the EASA. They concern the specific areas of the UAS operations (design of the UAS, maintenance, operation, competencies of the remote pilots, human errors and factors, operations under adverse weather conditions, etc.) and determine the related criteria and/or requirements. Based on the risk characteristics of the given operation, the SAIL levels determine which robustness level (low, medium, high) has to be achieved connected to the operation. In the case of low risk, the fulfilment of some OSOs is optional [15].

The robustness is the feature of the risk mitigation measures, which consists of two factors:

- level of integrity and
- level of assurance.

Both factors have low, medium, and high levels. The combination of these factors results in robustness. The level of robustness will always be equal to the lowest level of either the integrity or the assurance.

The EASA defined certain OSOs that can be validated by an independent third party and also defined those, which can be validated only by them. In the case of some OSOs, the medium or high robustness levels require the design verification report by the EASA.

2. The design of verification report

EASA decided to divide operations into three groups based on their SAIL. For operations in SAIL I and II that are considered as low-risk operations, the EASA accepts if the UAS operator provides conformity declaration of the manufacturer (that is available in conformity assessment) and any further step (i.e., higher level of verification) to prove conformity is optional.

But in the case of SAIL III-IV operations, it is up to the National Aviation Authorities (NAAs) to decide whether design-related OSOs shall be proven by a process called design verification. This process is conducted by EASA that will - similarly to regular aircraft - investigate technical issues of the UAS, but only those, and the OSOs in connection with the production and every other OSO remain in the competence of NAA to evaluate the robustness of documentation provided by the UAS operator [14], [15], [16], [17].

The EASA will, therefore, investigate the compliance of the following elements:

- the full design of the UAS;
- the mitigation means linked with the design; and
- the enhanced containment function.

As said the decision whether design verification is required or not the NAA shall bear the liability for not prescribing the need for the above process of EASA. Therefore, the EASA recommends to NAAs to choose this option, especially in case of operations over populated areas (The concept of populated area is yet to be clearly defined by the European legislator). Another important feature of the design verification report is that it is strictly linked to the concept of operations provided by the UAS operator (this can be derived from the investigated mitigation measures that depend on the concrete operation).

This means that this report has no general effectiveness covering all possible future operations and shall not result in the acquisition of the operational authorization or LUC automatically. It neither means that all UAS, that possesses a design verification report can be used in all operations of the "specific" category without further action required.

The rest of OSOs, including the production-related ones, will remain in national competence that further questions the effectiveness of this procedure - contrary to conformity assessment where design and production form one single unit. If it is considered, that the validity of the design verification report - unlike type certificate - will be limited to the European Union, the sense of introducing a third model between conformity assessment and type certification becomes doubtful.

Not to mention, that Article 13 of the IR already provides mutual recognition for operational authorizations and LUCs that means that these documents can be acquired from one NAA with validity to the entire European Union, - if the NAA so decides - without design verification report of the EASA.

The fact, that the NAAs still have the right to decide about design-related OSOs, hinders the creation of uniform application of EU law and consequently the single European market in the field of unmanned aircraft systems, which runs counter to aims defined in the preamble of the IR. The following **Fig. 2** shows the concept of EASA to the system of conformity, verification, and certification of UASs [16].

It is visible, that there is a shift in terms of regulatory mindset in the "specific" category from product safety towards the transport/flight safety approach. The essence of this change is the handling of the proven methods: the operation and the management of detected risks during the operation are in the focus of the design verification report.

Contrary to the UASs used in the "open" category, the UAS operator bears more responsibility for the compliance of the UAS as the determiner of the characteristics of the operations in the "specific" category, where the conformity means the capability of maintaining the high level of flight safety by mitigating the risks, especially in the field of design.

In the end, the operational authorization or LUC (either with or without design verification report) will - according to Article 40 par. (3) of the DR - serve as a document to declare the conformity of the UAS with the operation it is intended to be used for.

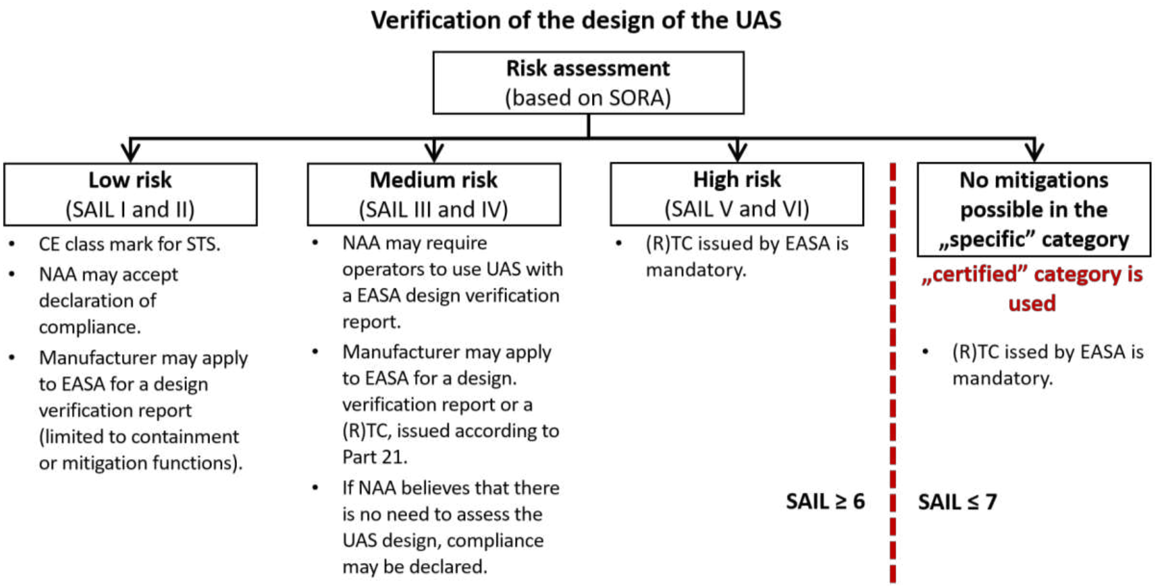


Figure 2. *Regulation system of the UAS certification according to the EASA recommendation*

VI. Type certificate - the "certified" category of UAS operations

The "specific" category acts as a "bridge" between the rules of the "open" and the "certified" category. A glaring example of this is the list of UAs operation-types that are subject to certification indicated in the DR. According to Article 40 (1) point (d) of the DR, if the risk mitigation is not possible without a certification - based on the risk assessment conducted by the aviation authority -, a "type" certification fulfilled by the EASA shall be compulsory for the given UAS, that is operated in the "specific" category.

The ex-lege "certified" category UAS operations are subject to EASA certification, although the certification regulations for UASs - optimized versions of the conventional aircraft certification rules - are currently not available. Only ideas are existing about these kinds of processes [18].

This is the reason, why the electric Vertical Take-Off and Landing (eVTOL) drone "air-taxi" - which is suitable for passenger transport and the best-known form of a drone to be operated in the "classified" category – has not appeared and spread yet in daily traffic, although many developments are underway.

Fig. 2 shows, however, a very interesting novelty in the interpretation of Article 40 (1) d) of the DR.

According to the above **Fig. 2**, the EASA provides a greater legem interpretation to SAIL V and VI operations as making them falling under the above provision of the DR. By the acceptance of the above, far more "specific" category operations shall require type certificate as provided for in "Part 21", similarly to regular aircraft and this might lead to disproportionate burden within the "specific" category and vanish the border between "specific" and "certified" category (Part 21 of Annex I to Commission Regulation (EU) No 748/2012 of 3 August 2012 laying down implementing rules for the airworthiness and environmental certification of aircraft and related products, parts and appliances, as well as for the certification of design and production organizations) [19].

VII. Outlook

In the future until 2023 there will be limited number of UAS with CIL. It is limiting the legal possibilities of the flight missions and require more authorizations which need more human effort (due to the single operational authorization). In this period the transitional provisions of the IR ensure the possible operations with limitations In the next few years NOBOs will emerge in the field of UASs but the spread of these organizations will require years. New standards will be published and the EASA has started programs for the establishment of the technical regulations concerned to UAS.

The future works in the field of the topic will concern to the standards and the new technical functions and their assessment. UASs have a lot of new technical functions that should be correctly evaluate and the new standards will contain the methods. Research of the functional appropriateness of the new technical possibilities will be appreciated in the future. Manufacturers should use new evaluation methods with the involvement of NOBOs, and the assessment procedure will cover more functions that nowadays.

The emergence of CIL will facilitate the use of UAVs both in the industrial and the recreational segment because it will clearly define the usage possibilities and the necessary legal and technical requirements.

VIII. Conclusions

In the article, the authors have presented the regulatory system of the UAS certification and its connections. This system - according to the level of risk - starts with conformity assessment and - through design verification report - leads towards "Part 21" certification. Special attention was dedicated to identifying the significant difference in the three operational categories defined by the IR. They are different but require and follow an interdependent regulatory logic and system to execute the certification and assessment of the conformity of the unmanned aerial systems. These different legal philosophies tried to make it possible to establish a uniform set of rules for the whole of the European Union, which is proportionate to the risks arising from their use and size, and which is capable of handling their differences with conventional aircraft. Further clarification is required, however, in the interpretation of rules and the application thereof: the system will not work if EASA does not provide proper guidance on problematic issues to avoid fragmentation of the single market of professional drones.

Acknowledgement

The work was realized based on Agreement No. NKFIH-866-5/2020, with the support of the National Research, Development and Innovation Office and the Ministry of Innovation and Technology, financed by the Security Technologies National Laboratory.

AUTHOR CONTRIBUTIONS

Zs. Sándor: Conceptualization, Writing, Review and editing.

M. Pusztai: Conceptualization, Writing, Review and editing.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Unmanned aerial vehicles market - Growth, trends, COVID-19 impact, and forecasts (2021 - 2026) 2021, [cited 05-07-2021]. https://www.mordorintelligence.com/industry-reports/uav-market

[2] Europe Unmanned Aerial Vehicle (UAV) market 2020-2026 by offering, frame, MTOW, technology, range, class, end user, and country: Trend forecast and growth opportunity. Report, no. 4901755 2020, [cited 05-07-2021]. https://www.researchandmarkets.com/reports/4901755/europe-unmanned-aerial-vehicle-uav-market-2020

[3] "Consolidated version of the treaty on the functioning of the European Union," Official Journal of the European Union, col. 47, no. C 326, pp. 1-390, 2012.

[4] "Commission delegated regulation (EU) 2019/945 of 12 March 2019 on unmanned aircraft systems and on third-country operators of unmanned aircraft systems," Official Journal of the European Union, vol. 62, no. L 152, pp. 1-40, 2019.

[5] "Commission implementing regulation (EU) 2019/947 of 24 May 2019 on the rules and procedures for the operation of unmanned aircraft", Official Journal of the European Union, vol. 62, no. L 152, pp. 45-71, 2019.

[6] "Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products," Official Journal of the European Union, vol. 51, no. L 218, pp. 30-47, 2008.

[7] "Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC," Official Journal of the European Union, vol. 51, no. L 218, pp. 82-128, 2008.

[8] ASD-STAN Standardization: Direct Remote ID – Introduction to the European UAS digital remote technical standard, 2021, [cited 05-07-2021]. https://asd-stan.org/wp-content/uploads/ASD-STAN_DRI_Introduction_to_the_European_digital_RID_UAS_Standard.pdf

[9] European Commission, Notified Bodies NANDO, [cited 05-07-2021]. https://ec.europa.eu/growth/tools-databases/nando/index.cfm?fuseaction=directive.main

[10] Good practice for market surveillance. [cited 05-07-2021]. https://ec.europa.eu/docsroom/documents/23041

[11] "Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/201," Official Journal of the European Union, vol. 62, no. L 169, pp. 1-44, 2019.

[12] EASA Flying a Drone for fun - Get to know how to become a good drone pilot. Video presentation: Are you flying a #drone for fun and want to become a good pilot? [cited 05-07-2021]. https://www.youtube.com/watch?v=9aXcUMygkAY

[13] EASA Operations in the medium risk of the specific category Video presentation: The new design verification process for authorising drone operations in the "specific" category – SA, [cited 05-07-2021]. https://www.youtube.com/watch?v=69EUfd4snzY

[14] JARUS guidelines on Specific Operations Risk Assessment (SORA), Joint Authorities for Rulemaking of Unmanned Systems. JAR-DEL-WG6-D.04, 2019.

[15] Acceptable Means of Compliance (AMC) and Guidance Material (GM) to Commission Implementing Regulation (EU) 2019/947. European Union Aviation Safety Agency, no. 1, pp. 11-113, 2019.

[16] EASA Webinar on Design Verification and update of EASA activities: Drone design verification in the medium risk specific category operations. EASA Drones team 16 April 2021, [cited 05-07-2021]. https://www.enac.gov.it/sites/default/files/allegati/2021-Apr/2021_04%20Webinar_design_verification_and_update_of_EASA_activities.pdf

[17] Guidelines on Design verification of UAS operated in the 'specific' category and classified in SAIL III and IV, European Union Aviation Safety Agency, no. 1, pp. 1-7, 2021.

[18] Special condition light USA, European Union Aviation Safety Agency, no. 1, pp. 1-25, 2020.

ORCID

Zs. Sándor http://orcid.org/0000-0001-7117-9069

M. Pusztai http://orcid.org/0000-0001-6983-4554

[19] "Commission implementing regulation (EU) 2012/748 of 3 August 2012 laying down implementing rules for the airworthiness and environmental certification of aircraft and related products, parts and appliances, as well as for the certification of design and production organizations", Official Journal of the European Union, vol. 55, no. L 224, pp. 1-92, 2012.

ACTA TECHNICA JAURINENSIS
Vol. 15, No. 1, pp. 47-57, 2022
10.14513/actatechjaur.00644

Sensitivity of CNOSSOS-EU sound propagation model to digital surface components

Edina Anna Balogh[1,*], Tamás Schmelz[1], László Orosz[1]

[1] KTI Institute for Transport Sciences, Department for Transport Acoustics,
Than Károly utca 3–5, H 1119 Budapest, Hungary
*e-mail: balogh.edina@kti.hu

Submitted: 22/10/2021 Accepted: 08/01/2022 Published online: 23/02/2022

Abstract: The accuracy of input data is a key issue in sound propagation model calculations. This paper aimed to assess the effect of building and land cover input data accuracy on CNOSSOS-EU sound propagation model outputs. Calculations were performed for a study site with a quite diverse land cover structure, located along a major road in Monor, Hungary. Nine test cases were defined based on building and land cover datasets with different accuracy. Comparing the results of the test cases to each other, it was found that in residential areas, the sound propagation model is more sensitive to the building data than to the land cover data. Therefore, it is recommended to use more detailed building input data in those areas, while using a land cover database with higher resolution than CORINE land cover data does not provide significantly better results. In non-residential areas, the influence of land cover input data on model results increases significantly.

Keywords: strategic noise mapping; CNOSSOS-EU; environmental noise propagation; input data

I. INTRODUCTION

According to Environmental Noise Directive 2002/49/EC (END), the first European Directive about strategic noise mapping, EU Member States are required to prepare strategic noise maps and action plans for agglomerations (urbanized areas with more than 100 000 inhabitants), and also for major roads, major railways and major airports every five years. Strategic noise maps are required for each transport sector, applying the day-evening-night noise indicator (Lden [dB]: day-evening-night sound level, defined by the formula (1), according to the Annex I of END) and the night-time noise indicator (Lnight [dB]: night sound level, defined by the description below the formula (1)). Based on the emission values calculated from annual average daily traffic data, the immission values at the receivers are determined by a model calculation describing the sound propagation. To prepare the strategic noise maps, Member States are required to use the new common assessment methods (CNOSSOS-EU method, a harmonised methodological framework for noise assessment, developed through the project "Common NOise aSSessment methOdS in the EU" lead by the Joint Research Centre of the European Commission between 2008-2014 [1]) published in the Commission Directive (EU) 2015/996 from 1 January 2019.

$$L_{den} = 10lg\frac{1}{24}\left(12 \cdot 10^{\frac{L_{day}}{10}} + 4 \cdot 10^{\frac{L_{evening}+5}{10}} + 8 \cdot 10^{\frac{L_{night}+10}{10}}\right) \quad (1)$$

where: Lday, Levening and Lnight are the A-weighted long-term average sound levels [dB] as defined in ISO 1996-2: 1987 [2], determined over all the day, evening and night periods of a year successively.

To develop the sound propagation model of the CNOSSOS-EU method, three existing sound propagation models (NMPB 2008: "Nouvelle Méthode de Prévision du Bruit des Routes", the French method for road traffic noise prediction [3], HARMONOISE: developed within the frame of the European project "Harmonoise", 2001-2004 [4] and ISO 9613 1 [5]) were evaluated on the basis of several criteria (namely precision, accuracy, computational speed, flexibility, simplicity, and number of parameters) [6]. The results of the investigation showed that the most complex HARMONOISE model, requires too high calculation time, which makes the model unsuitable for large-scale calculations that are needed to prepare strategic noise maps. Furthermore, the flexibility of the HARMONOISE model is presumably not relevant for strategic noise mapping, which is based on yearly-averaged noise indicators. Consequently, the NMPB 2008 model was suggested to be applied in the CNOSSOS-EU method. Examining test cases, another comparative analysis of the above three models gave the result that the HARMONOISE model does not perform better than the two simpler ones [7]. The reason for this may be that the HARMONOISE model is more sensitive to the input data. It was pointed out that it is not expected that the data to be used for strategic noise mapping would be more detailed than now.

The CNOSSOS-EU sound propagation model based on NMPB, is a point-to-point sound propagation method. From the sound emission level at the source, the sound level at the receiver is calculated by subtracting attenuation terms, which represent geometrical attenuation, air absorption, ground attenuation and screening attenuation [8]. Two particular types of atmospheric conditions (favourable and homogeneous) are taken into account. The long-term occurrence of favourable conditions is described by parameter p with a value of between 0 and 1 [1]. The acoustic absorption of the ground is represented by a dimensionless coefficient G with a value of between 0 and 1, which is independent of frequency [1].

A number of problems with the CNOSSOS-EU sound propagation model has been identified (e.g. concerning the Rayleigh-criterion and the retrodiffraction [8]) since it was published in Directive 2015/996. In 2018–19 a study was prepared by an EU working group with a view to establish the refining of the COSSOS-EU method, including its sound propagation parts [9] [10]. Based on that study, the amendments were published in Commission Delegated Directive (EU) 2021/1226.

The quality of input data is a key issue for preparing strategic noise maps with sufficient accuracy. As each EU member state can use different sources for input levels, it is difficult to obtain comparable results in spite of the common calculation method [11]. A data guideline would be needed to solve that problem [11], but no instructions have been published so far. Therefore, each member state is responsible for the selection of input data sources.

When selecting the input datasets to be used, availability, cost and required data processing have to be considered. Calculation time is another important issue to be examined for making a decision about the number and the accuracy of parameters taken into account in the model. The required accuracy of the input data should be determined considering the above aspects, besides the strategic level and the scale of the task.

The impact of input data on the outputs of model calculations has been discussed in several articles. Morley et al. assessed the performance of the CNOSSOS-EU road traffic noise prediction model using input data with different resolutions [12]. Six test scenarios were examined from the highest to the lowest resolution dataset. Both input data determining the emission (e.g. traffic flow, speed limits) and those influencing the sound propagation (e.g. land cover, building heights and topography) were analysed. The effect of traffic flow data proved to be the most significant factor in model calculations.

Aballéa et al. investigated the sensitivity of outdoor sound propagation predictions to environmental parameters to suggest different simplified approaches which can be used to reduce the calculation time, providing results that still meet the accuracy requirements [13]. Among other parameters, the ground effect was investigated, comparing cases with various heterogeneity values of ground surface absorption to a case assuming a homogeneous equivalent ground surface. It was stated that the difference in the A-weighted global excess attenuations calculated for the real case and the average ground surface does not exceed 0.3 dB.

Concerning the environmental noise propagation model calculations, the main influencing factors to be taken into account are the following:

- atmospheric absorption (which depends on the distance from the source and the atmospheric conditions);
- topography;
- foliage;
- acoustic characterisation of ground;
- obstacles (e.g. buildings and barriers);
- meteorological conditions (e.g. temperature gradient, wind direction).

The required accuracy of building data is an important issue, because the cost of such databases is quite significant in the implementation of strategic noise mapping tasks. Land cover data which can be used to describe the acoustic absorption properties of the ground is important to be examined, as assigning G values used in CNOSSOS-EU model to various land cover classes provided by different land cover data sources is not obvious. A sensitivity analysis could provide information on the importance of the accuracy of land cover data in sound propagation model calculation.

Several data sources for buildings used in sound propagation model calculations are discussed in the relevant literature, some of which provide data at the international level, while others are only available in a given country. The data model for noise simulation studied by Kumar et al. contains the following data for buildings: class of building, function (residential, public or industrial), usage, year of construction, year of demolition, type of the roof, measured height, number of the stories above and below ground and their heights [11]. Morley et al. examined two versions concerning the building height: in the higher resolution version, individual building heights from LiDAR measurements were applied, while in the lower resolution case generalized surface provided by averaging the LiDAR building heights over a 50 m grid or the constant value of 9.5 m were used [12]. The Irish guide for CNOSSOS-EU adaptation denotes OSiPRIME 2 product as the possible source of 2D building data, while building heights are recommended to be derived from LiDAR datasets or based on site surveys on the number of stories or estimated height values [14]. In the examination of the spatial relationship between air pollution and noise caused by traffic in two Danish cities, polygon shapefile of building footprints based on a national dataset (Kort10DK) was used with estimated building heights applying a national elevation model having 1 m × 1 m resolution [15].

Concerning the acoustic absorption of the ground, Kumar et al. used an attribute to model the degree of noise absorption by the land area with the values of 0 (hard), 0.5 (medium/middle) and 1 (soft) [11]. In the sensitivity analysis of Aballéa et al., two types of ground (reflective and absorbing) modelled by two impedances were taken into account [13]. In the examination performed by Morley et al., the extra detailed OS MasterMap Topography Layer® (1 m precision) was used in the higher resolution test cases, while the lower resolution versions were based on the less detailed CORINE 2006v16 database (~100 m precision) [12]. (The CORINE land cover dataset was developed in the framework of the CORINE - CO-oRdination of INformation on the Environment programme, accepted by the European Commission in 1985.) In Ireland, the following datasets are available: OSiPRIME2 (with a resolution of 1:1 000, 1:2 500, 1:5 000), OSi DigiCity (1:15 000), OSi LiDAR (may not be available in all areas of the country), CORINE (1:100 000) [14]. In the study of Khan et al. on the spatial relationship between traffic-related air pollution and noise, CORINE land cover data (version 2012) with a spatial resolution of 100 m was applied [15].

The aim of the present study is to examine the effect of the accuracy of building and land cover data on CNOSSOS-EU sound propagation model outputs, comparing results based on different data sources to each other. Road traffic was taken into account as a source, and calculations were performed on datasets from a study site located along a major road. The aim was to provide recommendations on how detailed building and land cover input data are worth using in strategic noise mapping tasks. It was expected that the accuracy of building data is particularly significant in residential areas, while the influence of land cover data may increase in areas with fewer buildings and a more diverse land cover characteristic.

II. Data acquisition and methods

1. Study area

The study site shown in **Fig. 1** is located in the Municipality of Monor, which is a town in Pest

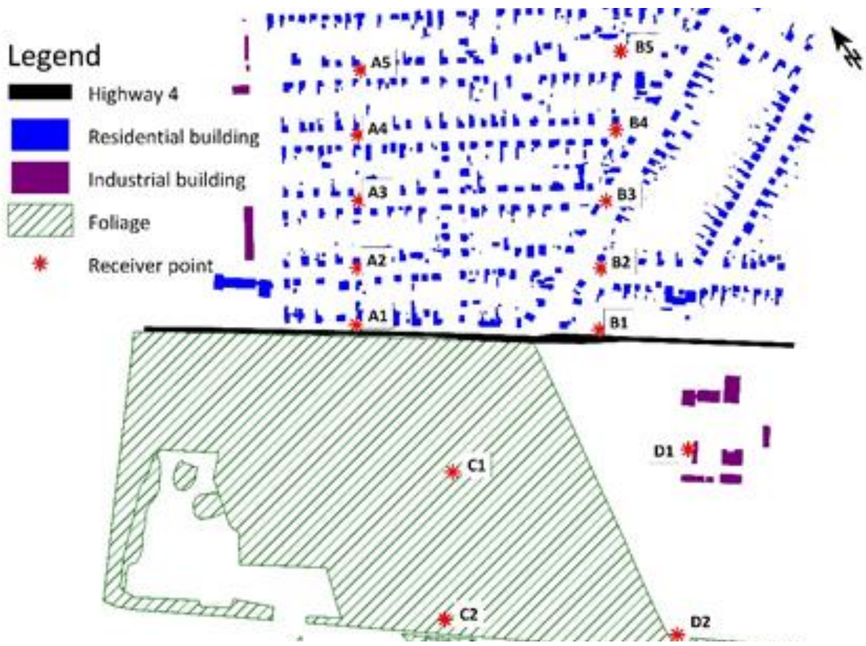


Figure 1. *Study site /Sources of data: buildings: Regulatory Plan of Municipality of Monor, www.monor.hu; road geometry and foliage: 2019 orthophotos, Lechner Knowledge Center, www. geoshop.hu/*

County with a population of 17 626 [16]. The study site is crossed by Highway 4, which is a major road according to the 49/2002 EU Directive, as it has more than three million vehicle passages a year on the examined section [17]. The total area of the study site is 0.53 km2. In site selection, it was a main aspect that the land cover is quite different on the two sides of the area. On the left side of Highway 4, there is a residential area consisting mainly of detached houses, with some shops and restaurants, while on the other side of the road, the area is mainly covered by forest, and there is a small industrial area as well. Another important factor was that several different building data sources and both CORINE and Urban Atlas as land cover databases were available for that area.

2. Input data

Some of the input data used in the sound propagation model were available directly, while other data were generated by digitizing different maps or based on on-site surveys.

A. Emission

Road traffic noise emission was determined applying the CNOSSOS-EU calculation method, based on the 2020 annual average daily traffic flow data [17]. Vehicle categories on which the Hungarian traffic count is based were assigned to the CNOSSOS-EU vehicle classes. The speed limit for all vehicle categories was considered as 50 km/h according to the traffic regulation for that given road section.

B. Buildings

Different data sources were used for the purpose of sensitivity analysis. The data sources and the input data provided by them were the following:

- Regulatory Plan of Municipality of Monor (2010, freely available on the official website of the municipality, www.monor.hu): building footprints (see Fig. 1);
- orthophotos (2019, provided by the Lechner Knowledge Center, www.geoshop.hu): building footprints;
- site survey (July 2021): building heights (reference building heights were measured and the height of the other buildings were established based on the number of levels, roof type and other characteristics e.g. mezzanine-floor), function of buildings (noise sensitive buildings: residential, educational; auxiliary buildings: outbuildings, garages, industrial buildings, shops, restaurants etc.), refined building footprint datasets (e.g. by deleting buildings present in the 2019 orthophotos or the 2010 Regulatory Plan but that no longer exist, recording roofed terraces and carports, which are considered as "floating screens" in model calculation).

Based on these available data sources, three input datasets with different accuracy were composed:

- building footprints based on the Regulatory Plan and estimated building heights, assigning a constant value of 6 m for each building uniformly (Regulatory Plan, version "a");
- building footprints based on the Regulatory Plan and building heights based on the results of the site survey (Regulatory Plan, version "b");
- building footprints based on 2019 orthophotos and building heights based on the results of the site survey (orthophotos).

C. Land cover

The different data sources and the provided input data were the following:

- CORINE (2020, version 2020 20u1): land cover classes;
- Urban Atlas (2020, version UA2018_v012): land cover classes;
- orthophotos (2019, provided by the Lechner Knowledge Center, www.geoshop.hu): extra absorbing areas identified in the residential areas (e. g. gravel or asphalt roads and car parks in the yards), where acoustic absorption properties are different from the ones of gardens.

Based on these three data sources, the following input datasets with different accuracy were created:

- CORINE land cover classes to which the suitable G values were assigned, e.g. G=1 for classes "Broad-leaved forest" and "Pastures"; G=0 for class "Industrial or commercial units" (CORINE);
- Urban Atlas land cover classes to which the suitable G values were assigned, e.g. G=1 for classes "Forests" and "Pastures"; G=0.7 for class "Discontinuous dense urban fabric"; G=0 for class "Other roads and associated land" (Urban Atlas);
- Urban Atlas land cover classes, improved with the extra reflective areas (G=0), based on the orthophotos (Urban Atlas+).

Land cover classes on the study site according to the different datasets are shown in **Fig. 2-3**. Note that in **Fig. 2-3** orthophotos from 2010 are shown, which are free to publish (source: fentrol.hu, Lechner Knowledge Center), in the model calculation building data based on 2019 orthophotos were used, nevertheless

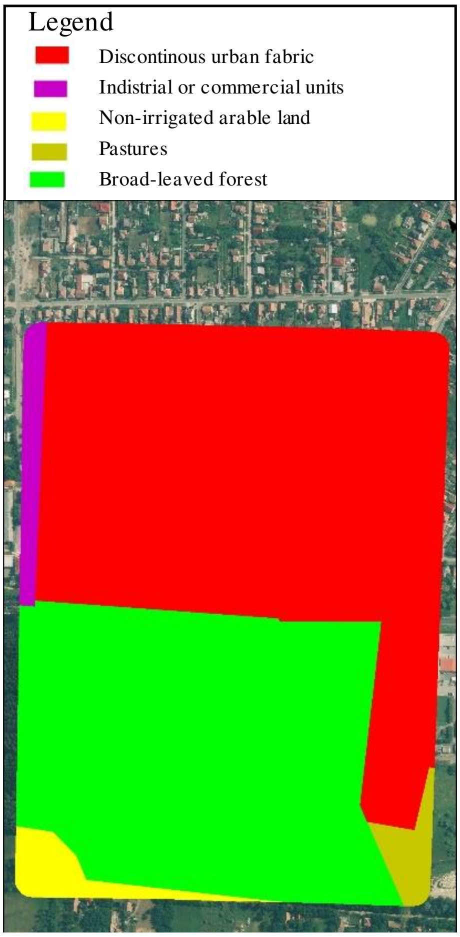


***Figure 2.** Land cover classes on the study area (CORINE)*

D. Topography

The digital terrain model (DDM5) is based on contour lines of topographic maps at the scale of 1:10 000, corrected by stereophotogrammetric evaluation. The dataset was provided by the Lechner Knowledge Center, www.geoshop.hu (downloaded in June 2021).

E. Foliage

Foliage dataset was generated from 2019 orthophotos (provided by the Lechner Knowledge Center, www.geoshop.hu) with height data evaluated based on site survey (July 2021).

F. Road geometry

Road geometry data was generated from 2019 orthophotos (provided by the Lechner Knowledge Center, www.geoshop.hu).

G. Meteorological conditions

As the input data for meteorological conditions were not the subject of the investigation, there was no need for data that describes real conditions accurately. As there is no data available yet in Hungary for parameter p, a value of 0.5 were considered in the model calculation for each part of the day (namely for day, evening and night periods).

3. Test cases

Nine test cases were defined based on the building and land cover datasets with different accuracy. The test cases and their abbreviated names are shown in **Table 1**.

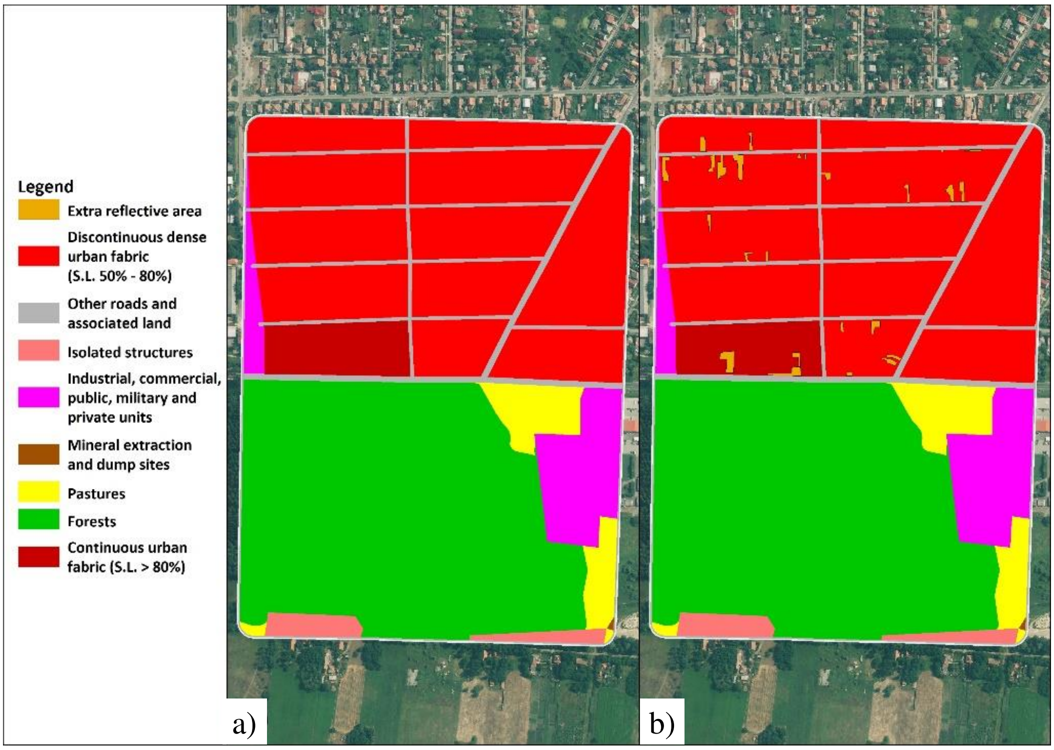


***Figure 3.** Land cover classes on the study area a) Urban Atlas b) Urban Atlas +*

Table 1. *Test cases*

Building data source	***Land cover data source***		
	CORINE	***Urban Atlas***	***Urban Atlas+***
Regulatory Plan – version "a"	RPa-CORINE	RPa-UA	RPa-UA+
Regulatory Plan – version "b"	RPb-CORINE	RPb-UA	RPb - UA+
Orthophotos	ORTO-CORINE	ORTO-UA	ORTO-UA+

For the comparison of the performance of each test case, results provided by the ORTO-UA+ test case was considered as the reference value, since it was based on the most detailed building and land cover datasets.

For the same reason, to assess the influence of building input data on the model results, the test cases using building data based on orthophotos were considered as "building reference values". Referring to land cover input data, test cases based on the Urban Atlas+ dataset were regarded as "land cover reference values".

4. Sound propagation calculation

Sound propagation calculation was performed using the SoundPLAN software (version 8.2), based on the CNOSSOS-EU Road: 2015 calculation method. 14 receiver points were designated: points A1–A5 and B1–B5 are located in the residential area on the left side of Highway 4, while on the other side, points C1–C2 are found in the forest and points D1–D2 were designated in and behind the industrial area (see Fig. 1). At each receiver point, the day-night-evening sound levels (Lden [dB]) were calculated at a height of 4 m.

III. Results and Discussions

Day-evening-night sound levels (Lden) at each receiver point for the nine test cases are given in **Table 2**.

At each receiver point, the maximum difference amongst the results for the different test cases (namely the range of Lden values) was calculated (see Fig. 3). In the residential area, the values of maximum differences are quite small at the receiver points close to the source (less than 1 dB at points A1 and B1). Away from the source, while Lden values decrease, the ranges increase significantly (it exceeds 6 dB at point B4), as differences between the different data sources exercise their effect more strongly on the longer propagation path. At the furthest receiver points, there are smaller ranges again, as out of the study site there are no buildings and ground effect given in the model; therefore, the propagation conditions are the same for each test case. In the woodland, the differences amongst the test cases are negligible (the range is 0.2 dB at point C1 and 0.5 dB at point C2), as there are no buildings in that area and the land cover input data according to the different sources is quite similar (see Fig. 2). In the industrial area, there is a more significant range in the middle of the propagation path (3.5 dB at point D1). This may be partly due to the differences in the input data for buildings. Moreover, the boundaries of land cover classes according to the CORINE and the Urban Atlas databases do not correspond completely either in that area (**Fig. 2-3**).

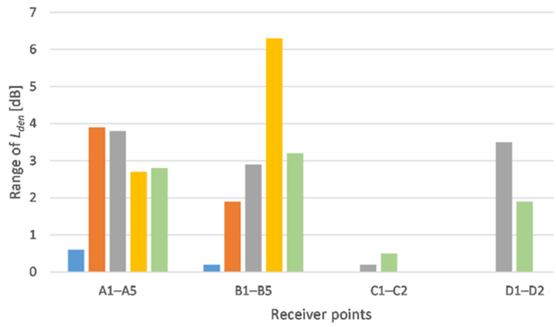


Figure 4. *Maximum differences in day-evening-night sound levels (Lden) in dB for the test cases at each receiver point*

In the next step, the magnitude of the influence on the results was assessed referring to the building and

Table 2. *Test cases*

Receiver point	***RPa-CORINE***	***RPa-UA***	***RPa-UA+***	***RPb-CORINE***	***RPb-UA***	***RPb-UA+***	***ORTHO-CORINE***	***ORTHO-UA***	***ORTHO-UA+***
A1	67.4	67.6	67.7	67.1	67.3	67.3	67.1	67.3	67.3
A2	49.3	49.9	50.5	50.3	50.9	51.4	47.5	48.0	48.6
A3	41.9	42.3	42.4	45.2	45.5	45.7	44.8	45.2	45.3
A4	39.8	40.1	40.1	42.1	42.5	42.5	40.7	41.2	41.3
A5	38.6	38.9	39.1	40.8	41.2	41.4	40.3	40.7	40.9
B1	66.8	66.9	66.9	66.8	66.9	66.9	66.9	67.0	67.0
B2	52.0	52.3	52.3	53.0	53.3	53.3	51.4	51.8	51.9
B3	43.8	45.2	45.3	45.7	46.7	46.7	44.6	45.6	45.6
B4	38.4	38.8	38.8	44.0	44.5	44.5	44.3	44.7	44.7
B5	39.3	39.7	39.8	42.0	42.4	42.5	41.2	41.6	41.6
C1	48.2	48.2	48.2	48.1	48.2	48.2	48.0	48.1	48.1
C2	41.9	42.1	42.1	41.9	42.1	42.1	41.6	41.8	41.9
D1	48.8	51.1	51.1	47.6	49.9	49.9	47.7	51.0	51.0
D2	40.8	42.0	42.0	40.6	41.7	41.7	40.1	41.0	41.0

land cover data. To investigate the importance of input data for buildings, the results of test cases using the same land cover data sources were compared to the corresponding building reference values (results for ORTO-CORINE, ORTO-UA and ORTO-UA+ test cases). Similarly, the analysis of the influence of input data for land cover was based on comparing the results of test cases applying the same building data sources to the suitable land cover reference values (results for RPa-UA+, RPb-UA+ and ORTO-UA+ test cases). To compare the results of test cases to the reference values, two indicators were used at each receiver point: maximum differences from the building/land cover reference values (**Table 3**) and mean of the absolute values of differences from the building/land cover reference values (**Table 3** and **Fig. 5**). Concerning the maximum differences, the signed values are presented, which enables to evaluate the direction of the differences (i. e. whether it is an overestimation or underestimation of the reference value) in addition to their magnitude. On the other hand, when calculating the mean differences, the absolute values of differences were used to represent the average magnitude of the differences at the receiver points in a more appropriate way.

Table 3. *Maximum differences and mean of the absolute values of differences from building/land cover reference values in day-evening-night sound levels (L_{den}) in dB at each receiver point*

Receiver point	*Max. diff. from building reference value [dB]*	*Max. diff. from land cover reference value [dB]*	*Mean of the absolute values of diff. from building reference values [dB]*	*Mean of the absolute values of diff. from land cover reference values [dB]*
A1	0.4	-0.3	0.2	0.1
A2	2.9	-1.2	2.4	0.9
A3	-2.9	-0.5	1.6	0.3
A4	1.4	-0.6	1.2	0.2
A5	-1.8	-0.6	1.1	0.4
B1	-0.1	-0.1	0.1	0.1
B2	1.6	-0.5	1.0	0.2
B3	1.1	-1.5	0.8	0.6
B4	-5.9	-0.5	3.1	0.2
B5	-1.9	-0.5	1.4	0.3
C1	0.2	-0.1	0.1	0.0
C2	0.3	-0.3	0.1	0.1
D1	1.1	-3.3	0.6	1.3
D2	1.0	-1.2	0.8	0.5

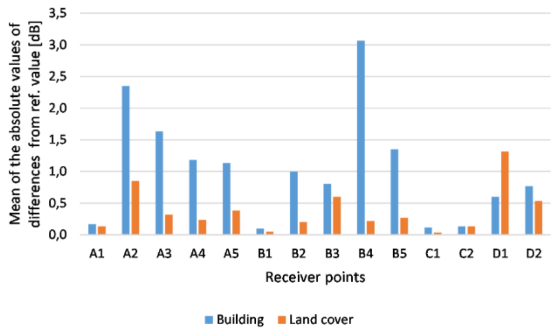


Figure 5. *Mean of the absolute values of differences from building/land cover reference values in day-evening-night sound levels (Lden) in dB at each receiver point*

Concerning the building data, in the residential area there are significant differences at the receiver points located in the middle range of the propagation path. The highest values of maximum differences and mean differences were detected at point A3 (-2.9 dB as maximum difference, building data source: RPa, and 1,6 dB as mean difference) and at point B4 (-5.9 dB, building data source: RPa, and 3,1 dB). Smaller L_{den} values than the building reference values are likely to be caused by assuming a constant value of 6 m for building height in test cases based on RPa building data source. This is an overestimation of the real building heights and the higher obstacles in the model result in lower sound levels. In the woodland, the differences are negligible, which follows from the lack of the buildings. In the industrial area, quite small differences were detected as well. This may be explained by that there are only a few buildings in that area; therefore, the differences in the building data sources do not influence the results significantly.

With respect to the land cover data, in the residential area much smaller differences can be noticed. The highest values of maximum differences and mean differences occur at point A2 (-1.2 dB, land cover data source: CORINE, and 0,9 dB) and at point B3 (-1.5 dB, land cover data source: CORINE, and 0,6 dB). In the woodland the differences may be ignored (maximum differences are -0.1 dB and -0.3 dB, land cover data source: CORINE, mean differences are 0,0 dB and 0,1 dB), which can be explained by the quite similar land cover input data according to the different sources. The small negative deviation is likely to be caused by that the surface of the highway is considered as a separate land cover category according to the Urban Atlas ("Other roads and associated lands", **Fig. 3a-b**), which were taken into account as a reflective surface in the model. On the other hand, in the CORINE database, the area of the highway is covered by the class "Broad-leaved forest" (**Fig. 2**), which was counted as an absorbing surface in the model calculation. In the industrial area, there is a significant difference between the land cover data according to both the CORINE and the Urban Atlas (**Fig. 2-3**) databases. The latter shows a broader industrial area, which means a reflective surface with smaller *G* value, while according to the CORINE database, a part of that area is defined as broad-leaved forest, which is a more absorbing surface with higher *G* value. This dissimilarity can cause that there is a considerable value of maximum difference at point D1 (-3.3 dB, land cover source: CORINE) and also the mean difference is higher than it is at any other receiver points (1,3 dB).

According to the results shown in **Table 3** and **Fig. 5** the model calculation is more sensitive to the building data in the residential area. However, land cover data may have an important role in the industrial area.

Finally, the performance of each test case was estimated comparing its result to the reference value (result provided by the test case ORTO-UA+). The absolute value of the deviation of the results provided by each test case (L_{den} [dB]) from the reference value at each receiver point is presented in **Fig. 6-7**. **Fig. 6 and Fig. 7a** show that in the residential area, the test cases using PRa as building data source provided the largest deviations. Within these test cases, the version using CORINE as the land cover data source performed the weakest (3.4 dB difference at point A3 and 6.3 dB at point B4). We note that the results provided by the test case ORTO-UA are almost equal to the reference value. This can be explained by the fact that in this part of the residential area hardly any extra reflective area has been identified; therefore, UA and UA+ data sources did not differ significantly (**Fig. 3a-b**). The deviations in the woodland are negligible (0-0.3 dB), as shown in **Fig. 7b**. As described above, this can be caused by the fact that there are no buildings in that area, and the land cover input data according to the different data sources are quite similar. In the industrial area, the highest deviation values at point D1 were provided by the test cases using CORINE land cover data (**Fig. 7c**). This can be explained by the different boundaries of land cover categories according to the CORINE database and the Urban Atlas in that area, as explained above.

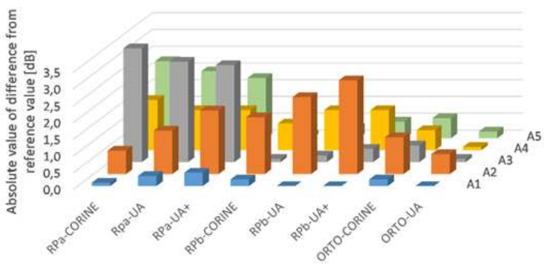
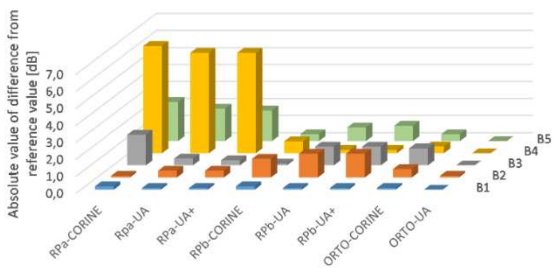
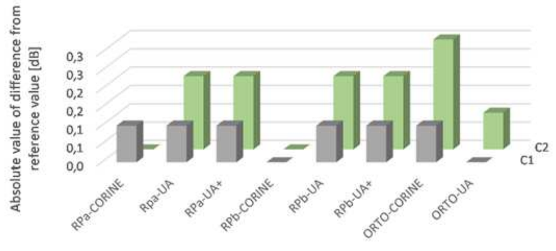


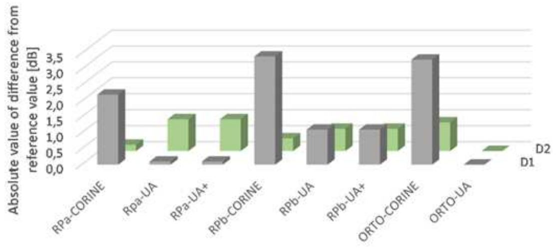
Figure 6. *Absolute values of differences from reference values in day-evening-night sound levels (Lden) in dB receiver points A1–A5*



a)



b)



c)

Figure 7. *Absolute values of differences from reference values in day-evening-night sound levels (Lden) in dB a) receiver points B1–B5 b) receiver points C1–C2 c) receiver points D1–D2*

IV. CONCLUSIONS

The accuracy of input data is a key issue in sound propagation model calculations. To perform the mandatory strategic noise mapping tasks, input data for quite large areas have been provided. As the financial sources and manpower required for this task are significant, it is essential to get to know the importance of input data accuracy for each influencing factor. The present study aimed to assess the effect of building and land cover input data accuracy on CNOSSOS-EU sound propagation model results.

The analysis shown in this article has limitations. It focused only on two influencing parameters, while the effects of other important factors (e. g. meteorological conditions) were not examined. Because of the lack of data for parameter *p* in Hungary, default values were used in the model. As the same modelling conditions were used to determine the noise levels for each test cases to compare, it was assumed that this simplification did not influence the conclusions can be drawn from the comparison. However, the values of parameter *p* for each part of the day at a regional scale in Hungary are expected to be determined in the framework of the CNOSSOS-EU adaptation project. Based on the result data, it could be expedient to perform further examinations on the effect of input data describing meteorological conditions.

An additional limitation is that the investigation was performed only in a given study site. Although an area with diverse land cover was selected, which includes residential, wooded and industrial areas, the conclusions drawn from the results of the analysis might be too specific. As in the residential area examined in the study there are only detached houses, the effect of other types of residential buildings (e. g. blocks of flats of different heights) cannot be assessed based on present analysis. However, horizontal and vertical expanse of the buildings can influence the propagation conditions in a significant way, mainly through multiple diffractions and reflections of the acoustic waves. Furthermore, the examined industrial area consists of only a few buildings of similar heights, but a more diverse complex of buildings may be more appropriate to assess the importance of input data for buildings. In the context of the study area, traffic flow data can be another influencing factor which could not been examined in the present study, as particular traffic flow values of a given road section were taken into account. Larger noise load caused by higher traffic flow (e. g. on a motorway) would result in a larger affected area with more buildings influencing the sound propagation. When evaluating the results, it should also be taken into account that the selected input data may reflect some typical Hungarian characteristics, which can be different from the conditions in other countries (e. g. in Western Europe). Therefore, further investigations need to be done in the future in various study sites, even in other countries. When selecting the study areas, it is recommended to select different types of residential and industrial areas, along with roads of different volume of traffic (e. g. highly urbanized, densely built-up areas with tall buildings and even a more complex road network; sparsely built-up areas along motorways; industrial areas with buildings of various sizes).

It is also important to highlight that the reference values are not considered as absolutely accurate results. They were regarded as the basis of the comparison, as they were the results of the test cases using the most detailed input data, which described the real conditions in the possibly most accurate way. In further investigations, reference values can be improved by using more accurate input data for buildings where it is available.

The results show that in the residential area, the range of L_{den} values (the maximum difference amongst the results for the different test cases at each receiver point) can be quite large, especially in the middle of the propagation path, where differences between the different data sources exercise their effect strongly. Similarly, there is a significant range in the middle of the propagation path in the industrial area, where the boundaries of land cover classes according to the CORINE and the Urban Atlas databases do not correspond completely, besides the differences in the building databases. In the woodland, the differences amongst the test cases are negligible, as there are no buildings in that area and the land cover input data according to the different sources is quite similar. The quite large deviations amongst the results for the different test cases in the residential and industrial areas show that the accuracy of input data fundamentally influences the model results.

Taking into account all the limitations described above, it can be stated that as expected, in residential areas, the sound propagation model is more sensitive to the building data than to the land cover data. Therefore, it is recommended to use more detailed building input data (building footprints based on data sources at least at the accuracy level of orthophotos and building heights individually determined for each building instead of generalized building height levels) in those areas, while using a land cover database with higher resolution than CORINE land cover data does not provide significantly better results. Improving the land cover dataset by identifying extra absorptive areas does not have a significant effect on the results either. On the other hand, the influence of land cover input data on model results increases significantly in non-residential areas. To specify and extend the results on this issue, further investigations are recommended in areas with diverse land cover, considering that the Urban Atlas data source is only available for urban areas.

AUTHOR CONTRIBUTIONS

E. Balogh: Conceptualization, Experiments, Writing.

T. Schmelz: Theoretical analysis, Review and editing.

L. Orosz: Supervision, Review and editing.

DISCLOSURE STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

E. Balogh http://orcid.org/0000-0002-9006-0351

References

[1] S. Kephalopoulos, M. Paviotti, F. Anfosso-Lédée, Common Noise Assessment Methods in Europe (CNOSSOS-EU). EUR 25379 EN. JRC72550. Luxembourg: Publications Office of the European Union (2012) https://doi.org/10.2788/31776

[2] Acoustics — Description and measurement of environmental noise — Part 2: Acquisition of data pertinent to land use, ISO 1996-2:1987 (1987). https://www.iso.org/standard/6749.html

[3] Acoustics — outdoor noise — calculation of sound levels, NF S 31–133 (2011). https://infostore.saiglobal.com/en-us/standards/nfs-31-133-2011-60554_saig_afnor_afnor_129788/

[4] E. Salomons, D. Van Maercke, et al., The Harmonoise sound propagation model, Acta Acustica United with Acustica 97 (2011) pp. 62–74. https://doi.org/10.3813/AAA.918387

[5] Acoustics — Attenuation of sound during propagation outdoors – Part 1: Calculation of the absorption of sound by the atmosphere, ISO 9613-1:1993(E) (1993). https://www.iso.org/standard/17426.html

[6] S. Kephalopoulos, M. Paviotti et al., Advances in the development of common noise assessment methods in Europe: The CNOSSOS-EU framework for strategic environmental noise mapping, Science of the Total Environment 482–483 (2014) pp. 400–-410. https://doi.org/10.1016/j.scitotenv.2014.02.031

[7] C. Foy, D. Ecotière, G. Dutilleux, Comparison of 3 engineering methods for outdoor sound propagation, in: 40. Jahrestagung für Akustik: DAGA, Oldenburg, 2014, pp. 828–829. https://pub.dega-akustik.de/DAGA_2014/data/articles/000456.pdf

[8] E. Salomons, A. Eisses, Investigations of the Cnossos sound propagation model, in: 11th European Congress and Exposition on Noise Control Engineering: Euronoise 2018, European Acoustics Association, Crete, 2018, pp. 1253–1260. https://www.euronoise2018.eu/docs/papers/212_Euronoise2018.pdf

[9] A. Kok, Refining the CNOSSOS-EU calculation method for environmental noise, in: 48th International Congress and Exhibition on Noise Control Engineering: INTERNOISE 2019, International Institute of Noise Control Engineering (I-INCE), Madrid, 2019, 1418. http://www.sea-acustica.es/fileadmin/INTERNOISE_2019/Fchrs/Proceedings/1418.pdf

[10] A. Kok, A. van Beek, Amendments for CNOSSOS-EU: Description of issues and proposed solutions, RIVM Letter report 2019-0023, National Institute for Public Health and the Environment, Bilthoven, The Netherlands (2019). https://doi.org/10.21945/RIVM-2019-0023

[11] K. Kumar, H. Ledoux et al., A Harmonized Data Model for Noise Simulation in the EU, ISPRS International Journal of Geo-Information 9 (2) (2020) 121. https://doi.org/10.3390/ijgi9020121

[12] D. W. Morley, K. de Hoogh et al., International scale implementation of the CNOSSOS-EU road traffic noise prediction model for epidemiological studies, Environmental Pollution 206 (2015) pp. 332–341. https://doi.org/10.1016/j.envpol.2015.07.031

[13] F-E. Aballéa, J. Defrance et al., Sensitivities of outdoor sound propagation predictions to environmental input parameters, Noise Control Engineering Journal 55 (1) (2007) pp. 38–49. https://doi.org/10.3397/1.2402310

[14] J-P. Faulkner, E. Murphy et al., Towards a Good Practice Guide for Implementing CNOSSOS-EU in Ireland, EPA Research Report 383. Environmental Protection Agency, Wexford, Ireland (2021). https://www.epa.ie/publications/research/environment--health/Research_Report_383.pdf

[15] J. Khan, K. Kakosimos et al., The spatial relationship between traffic-related air pollution and noise in two Danish cities: Implications for health-related studies, Science of the Total Environment 726 (2020) 138577 https://doi.org/10.1016/j.scitotenv.2020.138577

[16] KSH, 2011. ÉVI NÉPSZÁMLÁLÁS, 3. Területi adatok, 3.14 Pest megye [CENSUS 2011, 3. Spatial Data, 3. 14 Pest County], Hungarian Central Statistical Office, Budapest, Hungary (2013), in Hungarian. URL https://www.ksh.hu/docs/hun/xftp/idoszaki/nepsz2011/nepsz_03_14_2011.pdf

[17] MK, Az országos közutak 2020. évre vonatkozó keresztmetszeti forgalma [Cross-sectional flow data of national roads in 2020], Hungarian Public Roads Ltd., Budapest, Hungary (2021), in Hungarian https://internet.kozut.hu/kozerdeku-adatok/orszagos-kozuti-adatbank/forgalomszamlalas/