Editorial


Christian Müller


Konrad Marfurt


Norbert Ketterer


Frank Herrmann

Liebe Leserinnen und Leser,

vor Ihnen liegt nunmehr die bereits zwanzigste Ausgabe des E-Journals **Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI)** – wir hoffen, dass wir Ihnen wieder eine Reihe von spannenden Artikeln aus dem Umfeld der Wirtschaftsinformatik zusammenstellen konnten. Wir möchten auch noch einmal darauf hinweisen, dass die regulären Artikel alle durch einen komplett anonymisierten Review-Prozess laufen, in dem zwei Gutachter und ein Redakteur/ Herausgeber den Artikel begleiten.

Im Rahmen dieser Ausgabe möchten wir herzlich unserem Kollegen Konrad Marfurt danken, der das Journal seit seiner Gründung begleitet und dabei stark geprägt hat. Kollege Marfurt wurde vor einiger Zeit an der Hochschule Luzern pensioniert und verfolgt nun eigene Projekte u.a. in Ruanda. Während seiner Tätigkeit als Mitherausgeber hat er sich sehr stark um den Betrieb des Journals gekümmert, welches zuerst auf eigener Hardware betrieben wurde. In dieser Zeit musste er sich um mehrere herausfordernde Upgrades und Umzüge des Journals kümmern, bis es schließlich dem aktuellen Provider SOAP2, der von Swiss Universities unterstützt wird, übergeben werden konnte. Die übrigen Mitherausgeber möchten sich noch einmal sehr herzlich bei ihm für den jahrelangen Einsatz bedanken und wünschen ihm eine spannende Zeit in seinen aktuellen Projekten.

Diese Ausgabe enthält wieder eine Reihe von Zweitveröffentlichungen der European Conference on Modelling and Simulation (ECMS), welche im Rahmen dieses Journals zulässig sind und sich selbstverständlich mit dem Thema der Simulation befassen. Wesentlich zu bemerken sind hier Artikel zum Energiemanagement auf Basis eines Prognosemodells, der Simulation von Flüssigkeiten und der Belegung von Montagelinien sowie der Simulation von Beständen in der Fertigung, um Bestellzeitpunkte und Mengen zu optimieren.

Die Kernartikel dieser Ausgabe lassen sich grob in die Bereiche Informationsmanagement, Anwendungssysteme, Geschäftsprozessmanagement und BI einordnen.

Der Bereich des Prozessmanagements wird durch einen Artikel abgedeckt, der sich mit der noch recht neuen Thematik des Process Minings befasst. Hier wird speziell das traditionelle fallbasierte Mining mit dem neueren objektzentrierten Mining in einer Studie verglichen und unter Verwendung von Celonis an Hand einer Fallstudie unter anderem die Frage behandelt, welche Vorteile der neuere Ansatz birgt.

Der Bereich des Informationsmanagements wird durch zwei Artikel adressiert: Ein Artikel behandelt ein Support-Ticketsystem, ein weiterer den Einsatz von Gamification in Anwendungen. In ersterem Artikel wird die Einführung des Ticketsystems für ein Unternehmen der Automatisierungstechnik, hier speziell für die IT beschrieben, welches in die vorhandenen Prozesse des Geschäftsbereichs eingefügt werden kann. Der Artikel vergleicht mehrere Open-Source Projekte miteinander und prüft, ob diese die benötigte Grundfunktionalität zur Verfügung stellen. Der Artikel zur Gamification behandelt den Anwendungsbereich digitaler Gesundheitsanwendungen, wobei der Fokus auf die Untersuchung des Einflusses von Gamification auf die Nutzerakzeptanz gelegt wird.

Im Bereich der Anwendungssysteme wird die Implementierung eines Treibers zur Anbindung von mikrocontrollerbasierten Maschinen über OPC-UA an das Konfigurations- und Inbetriebnahme-Tool für einen Industrieausrüster beschrieben. Konkret befasst sich der Artikel mit der Implementierung eines Gerätetreibers zur Anbindung von Maschinen mit einer neuen, eigenentwickelten SPS an das Konfigurations- und Inbetriebnahme-Tool des Unternehmens.

Im Bereich der BI behandelt ein Artikel das Filter Pruning für Mask R-CNN, bei welchem das Neuronale Netz optimiert wird. Der Anwendungsfall stammt zwar aus der Bildverarbeitung, aber auch für Applikationen der WI werden diese Technologien immer relevanter werden.

Wir haben diesmal wieder vier Kurzdarstellungen von Abschlussarbeiten aufgenommen, welche den Bereich des Informationsmanagements und der Business Intelligence abdecken. Zwei Arbeiten befassen sich mit KI-basierten Chatbots. Eine Arbeit adressiert über einen KI-Screenreader das Thema der Barrierefreiheit und die vierte Arbeit untersucht die Fragestellung, in welchem Masse BI Prozesse durch KI erweitert und effizienter gestaltet werden können.

Über Ihr Interesse an der Zeitschrift freuen wir uns und wünschen Ihnen Freude bei der Lektüre.

Regensburg, Fulda, Luzern und Wildau, im Dezember 2024.

Frank Herrmann, Norbert Ketterer, Konrad Marfurt und Christian Müller

Structured Filter Pruning applied to Mask R-CNN: A path to efficient image segmentation

Yannik Frühwirth
Business Information Science
Baden-Wuerttemberg
Cooperative State University
(DHBW) Stuttgart
Rotebühlstr. 133
70197 Stuttgart
yannik.fruehwirth@web.de

ABSTRACT

This study explores the optimization of two-stage object recognition systems, which are integral to numerous applications, by leveraging advanced machine learning techniques. While such systems, including Mask R-CNN, achieve high recognition accuracy, they are often hindered by over-parameterization, excessive computational demands, and significant storagere quirements. To address these challenges, this research introduces a pruning method specifically designed for complex architectures like Mask R-CNN, aimed at reducing computing time, simplifying model complexity, and optimizing storage, all while maintaining detection accuracy.

The proposed method employs a Global Kernel Level Filter Pruning strategy, guided by the *L1-Norm*, to strategically remove non-essential parameters post-training. Experimental results demonstrate that this approach preserves recognition accuracy up to 50% pruning while achieving an 11.6% improvement in computing time on Graphics Processing Units and an 8% improvement on Central Processing Units. Furthermore, the method achieves a compression ratio of 1.47, reducing memory requirements by 33.5%, without compromising Average Precision, which remained at 0.32, equal to the unpruned model at this level.

These findings provide valuable insights into the efficiency optimization of Neural Networks, offering a practical and scalable solution for balancing accuracy, speed, and resource usage in complex architectures. This work contributes to advancing the state-of-the-art in Artificial Intelligence and opens new pathways for integrating complementary techniques such as quantization for further enhancements.

Keywords

Pruning, Filter Pruning, Mask R-CNN, Image segmentation, Two-stage detection

INTRODUCTION

Machine learning, driven by advanced computational power and Graphics Processing Units, has recently gained immense interest (Alpaydin 2016, (Shinde and Shah 2018), particularly in natural language processing, predictive analytics, and image processing (Shinde and Shah 2018). A key focus is object recognition, crucial for applications like autonomous vehicles (D. Feng et al. 2021). This technology's significance is evident in both academic research and public discourse, highlighting its increasing impact on daily life.

However, a significant challenge within this domain is addressing the complexity and computational demands of two-stage object recognition processes (Zou et al. 2019). These processes are characterized by high accuracy but suffer from issues such as over-parametrization (Canziani 2016), extended inference times (Chen et al. 2021), and substantial model storage requirements (Basili 2002). A notable gap in current research is the lack of insights regarding pruning strategies, specifically for complex, two-stage object recognition systems like Mask-RCNN, as evidenced by the scarcity of literature on this topic (Tzelepis et al. 2019, Aguiar Salvi and Barros 2021).

The primary objective of this research is to improve computing time for two-stage object recognition systems through the design and implementation of a tailored pruning approach. While computing time, measured as the change in inference time, is the main focus, the method also aims to address over-parameterization (quantified by the compression ratio) and reduce model memory size (measured as the change in memory size in megabytes). This is achieved by strategically manipulating the model parameters post-training to optimize performance across these metrics.

RELATED WORK

The research contributes primarily to Filter Pruning in the broader context of model compression. Scientific findings on model compression can be clustered into three categories:

Connection pruning

Connection pruning introduces sparsity in deep Neural Networks by eliminating redundant connections (Blalock et al. 2020), a vital aspect of model optimization. Pioneering techniques, such as those in (LeCun 1989) and (Hassibi and Stork 1992), used Taylor expansion for parameter significance assessment. However, these methods often require specialized hardware to leverage the resulting sparsity effectively. Recent advancements in connection pruning have focused on unstructured approaches, like the iterative pruning method in (Han 2016), which removes weights below a certain threshold. Although beneficial in fully connected

layers, these methods do not typically lead to significant reductions in computational load in Convolutional Layers (Anwar 2017).

Filter pruning

This method involves evaluating the importance of each filter within the network and pruning accordingly, followed by a crucial retraining phase to recuperate any loss in accuracy (Li et al. 2017).

In determining filter importance, various methodologies are employed. For instance, (Abbasi-Asl and Yu 2017) assesses filter significance by monitoring the impact of its removal on the model's accuracy. Similarly, (Li et al. 2017) utilizes the *L1-Norm* to determine filter importance, while (Hu et al. 2016) bases its evaluation on the activation of output Feature Maps from a subset of training data. These methods largely rely on hand-crafted heuristics.

Advancing beyond these, data-driven approaches for ranking filters have been proposed. One such method, as seen in (Liu et al. 2017), involves Channel Level Pruning, where a learnable scaling factor is attached to each channel. This factor is governed by the *L1-Norm* during training. Group sparsity has also emerged as a promising direction for Filter Pruning, with works like employing group lasso for this purpose (H. Zhou 2016, Wen et al. 2016). However, these techniques sometimes necessitate specialized hardware to optimize SpeedUp during inference (Anwar 2017).

Particularly relevant to the research are the approaches in (Molchanov et al. 2017), (Tzelepis et al. 2019) and (Singh et al. 2019), which, among other things, represents an innovation in filter classification through the use of absolute gradient values. This methods have demonstrated competitive results compared to more traditional brute-force methods, such as checking loss deviation for each filter.

Quantization

Quantization complements the pruning methods by reducing the memory and computational demands of a network. It involves converting network weights into a lower bit configuration (Cosman et al. 1993). Techniques like binarization (Rastegari et al. 2016) and ternary quantization (H. Zhou 2016) have been pivotal in this area. However, these methods sometimes necessitate special hardware for optimal implementation (Gholami et al. 2015). The combination of pruning and quantization, as seen in (Huang et al. 2017), highlights the potential for achieving significant model compression while maintaining performance.

RESEARCH METHODOLOGY

The methodology follows the Design Science Research approach (Peffers et al. 2007), which is known for its iterative nature, to experimentally evaluate the proposed compression method on various modern Neural Network architectures. This process is focused on the goal of optimizing compression, memory requirements and inference time for networks of different depths and widths in different domains. The project was divided into two distinct iterations, each of which produced a prototype that embodied the dynamic and adaptive progression that characterizes Design Science Research. The iterative process facilitates the refinement of the compression technique and allows to respond effectively to the insights gained at each stage.

Dataset and Algorithm

The widely used Microsoft COCO dataset (Lin et al. 2014) was selected for image classification and segmentation. The full scope of the subdataset train (80k images) (Lin et al. 2014) and subdataset validate (35k images) (Lin et al. 2014) are applied to the Mask R-CNN algorithm. Mask R-CNN is based on a Resnet50 backbone and is already pre-trained (Facebook 2020). The pruning is applied post-training. Recognition accuracy and computing time are measured using the minival sub-dataset (5k images) (Lin et al. 2014).

All implementations of the compression method are executed in PyTorch and CUDA, ensuring high performance and compatibility with modern graphics hardware. The inference time is tested on both graphics and central processing hardware to evaluate performance across diverse systems. Specifically, testing on the graphics hardware was conducted using the Nvidia H100 NVL, with a maximum memory capacity of 93 GB, while central processing hardware testing was performed on an Apple M1 chip with 16 GB of unified memory. Furthermore, to foster reproducibility and community engagement, the implementations are publicly available via GitLab.

Evaluation

The evaluation primarily focuses on computing time, with inference time serving as the key metric. Inference time is measured under consistent and controlled conditions, timing the model's forward pass during inference on both graphics and central processing hardware.

Additionally, the compression ratio is calculated, which is a direct measure of the reduction in network size:

$$\text{compression ratio} = \frac{\text{base model size}}{\text{pruned model size}}$$

Finally, the model's detection performance is evaluated using Average Precision for Intersection over Union thresholds ≥ 0.5. For the remainder of this paper, Average Precision refers to Average Precision computed at Intersection over Union 0.5. This metric is used to validate that the pruning method preserves detection accuracy.

These metrics are critical in gauging the trade-off between model efficiency and performance, ensuring that the compression method achieves the optimal balance for practical deployment.

EXPERIMENTS

Filter Pruning at Channel Level

Approach

The initial approach involved systematically deactivating filters by nullifying channels. This was based on the calculation of the *L0-Norm* for each channel, which counts non-zero values and returns the result (M. Feng et al. 2013). Channels were then sorted ascendingly based on their *L0-Norm*, and

a predetermined percentage of the least important channels were pruned.

Referring to Table 1, the results demonstrate that the pruning approach becomes ineffective for pruning proportions exceeding 10%, as the Average Precision (AP) drops to 0. While pruning leads to a reduction in inference time for both the Graphics Processing Unit (GPU) and the Central Processing Unit (CPU), the loss in detection accuracy outweighs the computational benefits. This sharp decline underscores the limitations of the current pruning strategy, suggesting it is unsuitable for practical deployment.

The failure of the method is likely due to the collapse of output feature maps, where essential information required for accurate detection is eliminated during the pruning process.

Table 1: Filter Pruning at Channel Level

Base Model	**Pruning Proportion [%]**	**Inference Time on GPU [ms]**	**Inference Time on CPU [ms]**	**AP**
Mask R-CNN	*0*	22.80	2128.03	0.31
Mask R-CNN	*10*	27.36	2236.58	0.18
Mask R-CNN	*20*	24.47	2165.98	0.03
Mask R-CNN	*30*	16.30	1963.02	0.00
Mask R-CNN	*40*	13.28	1837.71	0.00
Mask R-CNN	*50*	13.28	1801.18	0.00
Mask R-CNN	*60*	13.33	1836.08	0.00
Mask R-CNN	*70*	13.21	1746.75	0.00
Mask R-CNN	*80*	13.06	1793.32	0.00

The findings highlight the inadequacy of this pruning approach for complex architectures, as the severe reduction in Average Precision negates the benefits of reduced model size and computational efficiency. These results emphasize the need for alternative pruning strategies that can achieve model compression while preserving detection accuracy.

Filter Pruning on Global Kernel Level

Deficit analysis

The rapid decline in recognition accuracy was attributed to the collapse of subsequent layers caused by structured Filter Pruning at the Channel Level. It was clear that the method should not completely nullify the Feature Maps of filter outputs, essential for subsequent calculations.

Approach

The second iteration involved atomic-level manipulation. Similiar to the researches by (Li et al. 2017) and (Kumar et al. 2021), every parameter within the convolution kernels was considered for pruning, with the least important parameters identified using the *L1-Norm* method, which calculates the sum of vector sizes (Kumar et al. 2021). This allowed for a more nuanced pruning approach where fewer values might be pruned in some kernels compared to others.

Results

The pruning approach demonstrated improved computational efficiency, particularly in terms of inference time, with some trade-offs in detection accuracy as reflected in Average Precision (AP). Table 2 provides a comprehensive overview of these results.

The inference time on the Graphics Processing Unit showed a consistent reduction as pruning proportions increased. At 50% pruning, the inference time decreased from 22.8 milliseconds for the base model to 20.15 milliseconds, representing an 11.6% improvement in computational efficiency. The reduction continued with further pruning, reaching 17.31 milliseconds at 80% pruning, marking a total 24.1% improvement compared to the base model. Similarly, the inference time on the Central Processing Unit exhibited minor improvements. At 50% pruning, the inference time decreased from 2128.03 milliseconds for the base model to 1958.89 milliseconds, yielding an 8.0% improvement. Further pruning resulted in a consistent reduction, with the inference time reaching 1830.48 milliseconds at 80% pruning, amounting to a total 14.0% improvement compared to the base model.

For the Central Processing Unit, a similar trend was observed, though the reductions were less pronounced. At 50% pruning, the inference time decreased from 2731 milliseconds to 2428 milliseconds, corresponding to an 11.1% improvement. At 80% pruning, the inference time further decreased to 2386 milliseconds, resulting in a total 12.6% improvement compared to the base model.

The detection accuracy, as measured by Average Precision (AP), remained stable up to 50% pruning, maintaining values between 0.31 and 0.34. Beyond this point, the Average Precision began to decline significantly, dropping to 0.22 at 60% pruning and experiencing a sharp fall to 0.05 at 70% pruning. At 80% pruning, the Average Precision reached 0.00, indicating a complete loss of detection capability. This decline suggests that high pruning rates lead to a collapse of individual filter kernels, disrupting subsequent computations and feature map generation.

Table 2: Filter Pruning on Global Kernel Level

Base Model	**Pruning Proportion [%]**	**Inference Time on GPU [ms]**	**Inference Time on CPU [ms]**	**AP**
Mask R-CNN	*0*	22.80	2128.03	0.31
Mask R-CNN	*10*	22.74	2263.18	0.31
Mask R-CNN	*20*	23.67	2263.18	0.31
Mask R-CNN	*30*	24.10	1982.79	0.34
Mask R-CNN	*40*	21.73	2103.54	0.33
Mask R-CNN	*50*	20.15	1958.89	0.32
Mask R-CNN	*60*	19.97	1873.57	0.22
Mask R-CNN	*70*	18.45	1853.44	0.05
Mask R-CNN	*80*	17.31	1830.48	0.00

Filter pruning at the channel level proved to be ineffective. The collapse of filters resulted in unusable detections, rendering a comparison of computing time irrelevant. In con-

trast, the proposed approach, filter pruning on the global kernel level, demonstrated its effectiveness by maintaining consistent detection accuracy at a pruning rate of 50% of all parameters in the convolutional layers, along with a 11.6% improvement in computing time.

Furthermore, the base model requires 179 megabytes of memory, while the pruned model reduces this requirement to 119 megabytes. This corresponds to a 33.5% reduction in memory usage, making the pruned model significantly more efficient in terms of storage while preserving detection performance.

CONCLUSION

Objective

The study embarked on addressing a critical issue in the realm of Neural Networks, particularly focusing on the complexity and efficiency of two-stage object recognition methods like Mask R-CNN. The challenge lays in reducing computing time, over-parametrization and decreasing model storage size without compromising the high recognition accuracy inherent to these methods. The existing corpus of literature demonstrates a conspicuous paucity of insights into the application of pruning techniques within the ambit of intricate, two-stage object recognition frameworks. This research endeavor specifically targets this lacuna, with a focus on elucidating the implications of such techniques when applied to the Mask R-CNN architecture. The objective is to enrich the academic discourse by providing a comprehensive analysis of pruning strategies in complex Neural Networks, thereby bridging the identified knowledge gap.

Results

The research introduced a novel concept of Filter Pruning at the Global Kernel Level. This approach strategically identifies and eliminates the least significant parameters within the convolutional kernels of Mask R-CNN using the *L1-Norm*. This method represents a significant advancement in network optimization, effectively reducing the network's complexity and computational time while preserving the crucial accuracy required for object recognition tasks. The findings highlight the potential of precise, kernel-focused pruning as a powerful strategy to enhance the efficiency of complex Convolutional Neural Network architectures.

Key results of this study include maintaining high recognition accuracy up to 50% pruning, achieving an 11.6% improvement in computational time on the Graphics Processing Unit and an 8% improvement on the Central Processing Unit, while delivering a total compression ratio of 1.47. At this pruning level, the Average Precision remained at 0.32, equivalent to that of the unpruned model, demonstrating the effectiveness of this approach in preserving detection performance.

Implications

This study provides key insights for optimizing Neural Networks, introducing a post-training compression technique for Mask R-CNN that enhances algorithm refinement and efficiency. The findings reveal that pruning Feature Map output channels offers limited benefits, whereas fine-tuning filter kernels at a granular level is more effective and adaptable for similar two-stage recognition methods. This approach not only reduces computing time, model size and complexity but also maintains high recognition accuracy, ensuring its practicality for real-world applications.

Furthermore, the proposed Filter Pruning strategy significantly enhances the model's suitability for complementary compression techniques, particularly quantization. By eliminating redundant parameters and structuring the model more efficiently, this method opens the door to substantial improvements in computing time when combined with quantization. Together, these methods create a synergistic pathway for optimizing resource usage while maintaining detection performance, making this approach highly relevant for deployment in resource-constrained environments, such as edge devices and real-time systems.

FUTURE WORK

The field of object recognition, especially with complex architectures like Mask R-CNN, is on a trajectory of continuous evolution and expansion. This growth trajectory underscores the pressing need for effective compression methods that can adeptly manage the intricacy and expansiveness of these systems. As the utilization of object recognition methodologies escalates, it's anticipated that the significance of algorithms such as Mask R-CNN will correspondingly rise, marking a pivotal juncture in the field's advancement.

Future research in this domain is poised to traverse several critical paths. Firstly, there is a compelling need to pioneer new compression approaches. These novel strategies should ideally harness the latest developments in machine learning and Artificial Intelligence, specifically tailored to address the unique challenges posed by intricate Neural Network architectures. The creation of these innovative methods is paramount to keeping pace with the escalating complexity and capabilities of these systems.

Furthermore, there is a significant opportunity to refine and optimize existing compression procedures. This optimization could focus on multiple fronts, including enhancing efficiency, minimizing computational demands, and striking a more effective balance between model size, processing speed, and accuracy. Fine-tuning these elements is crucial to ensure that compression techniques maintain their relevance and efficacy, especially as technology rapidly advances and network architectures grow increasingly complex.

Another vital area of focus is the application of compression techniques to a diverse array of algorithms within object detection. Moving beyond Mask R-CNN, this research would extend the reach and applicability of these compression methods, making them useful across a broader spectrum of object detection applications. By encompassing a variety of algorithms, this research endeavor can significantly contribute to the overall functionality and utility of object detectors.

Finally, this work demonstrates that pruning not only reduces over-parameterization but also prepares the model for quantization, a complementary compression technique with significant potential for further reducing memory and com-

putational requirements. By removing redundant parameters through pruning, the model structure becomes better suited for lower-precision quantization, enabling even more substantial gains in efficiency. Future studies should explore the combined impact of pruning and quantization, particularly in resource-constrained environments such as edge devices, where lightweight models are paramount.

REFERENCES

Abbasi-Asl, R. and B. Yu 2017. "Structural Compression of Convolutional Neural Networks Based on Greedy Filter Pruning". In: *arXiv preprint arXiv:1705.07356*.

Aguiar Salvi, A. de and R. C. Barros 2021. "Model Compression in Object Detection". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Alpaydin, E. 2016. *Machine learning, The new AI*. Cambridge, MA: MIT Press.

Anwar, S. et al. 2017. "Structured Pruning of Deep Convolutional Neural Networks". In: *ACM Journal on Emerging Technologies in Computing Systems* 13.3, pp. 1–18.

Basili, V. R. 2002. "The role of experimentation in software engineering: past, current, and future". In: *Proceedings of IEEE 18th International Conference on Software Engineering*. 1, pp. 1–8.

Blalock, D. et al. 2020. "What is the State of Neural Network Pruning?" In: *Proceedings of Machine Learning and Systems 2020* 1.1, pp. 1–18.

Canziani, A. et al. 2016. *An Analysis of Deep Neural Network Models for Practical Applications*. arXiv.

Chen, L. et al. 2021. "Knowledge from the original network: restore a better pruned network with knowledge distillation". In: *Complex Intelligent Systems* 8.1, pp. 1–10.

Cosman, P. C. et al. 1993. "Using vector quantization for image processing". In: *Proceedings of the IEEE* 81.9, pp. 1326–1341.

Facebook 2020. *From Research to Production*. `https://pytorch.org/`. Retrieved: 05.05.2020.

Feng, Di et al. 2021. "A Review and Comparative Study on Probabilistic Object Detection in Autonomous Driving". In: *IEEE Transactions on Intelligent Transportation Systems* 1, pp. 1–20.

Feng, M. et al. 2013. *Complementarity formulations of l0-norm optimization problems*. Industrial Engineering and Management Sciences. Technical Report.

Gholami, A. et al. 2015. "A Survey of Quantization Methods for Efficient Neural Network Inference". In: *Proceedings of the IEEE International Conference on Computer Vision*. Santiago, pp. 1440–1448.

H. Zhou, et. al 2016. "Less Is More: Towards Compact CNNs". In: *ECCV*. Springer, pp. 662–677.

Han, S. et. al 2016. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding". In: *International Conference on Learning Representations*. San Juan, pp. 1–14.

Hassibi, B. and D. Stork 1992. "Second Order Derivatives for Network Pruning: Optimal Brain Surgeon". In: *Neural Information Processing Systems 1992*, pp. 164–171.

Hu, H. et al. 2016. "Network Trimming: A Data-Driven Neuron Pruning Approach Towards Efficient Deep Architectures". In: *arXiv preprint arXiv:1607.03250*.

Huang, J. et al. 2017. "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors". In: *IEEE Conference on Computer Vision and Pattern Recognition 2017*, pp. 1–21.

Kumar, A. et al. 2021. "Pruning filters with L1-norm and capped L1-norm for CNN compression". In: *Appl Intell* 51, pp. 1152–1160. DOI: 10.1007/s10489-020-01894-y.

LeCun, Y. et al. 1989. "Optimal Brain Damage". In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Morgan-Kaufmann, pp. 1–8.

Li, H. et al. 2017. "Pruning Filters for Efficient ConvNets". In: *International Conference on Learning Representations*. Toulon, pp. 1–13.

Lin, T.-Y. et al. 2014. "Microsoft COCO: Common Objects in Context". In: *European Conference on Computer Vision*. Springer, pp. 740–755.

Liu, Z. et al. 2017. "Learning Efficient Convolutional Networks Through Network Slimming". In: *ICCV*. IEEE, pp. 2755–2763.

Molchanov, P. et al. 2017. "Pruning Convolutional Neural Networks for Resource Efficient Inference". In: *ICLR*.

Peffers, K. et al. 2007. "A Design Science Research Methodology for Information Systems Research". In: *Journal of Management Information Systems* 24.3, pp. 45–77.

Rastegari, M. et al. 2016. "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks". In: *ECCV*. Springer, pp. 525–542.

Shinde, P. P. and S. Shah 2018. "A Review of Machine Learning and Deep Learning Applications". In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, pp. 1–6.

Singh, P. et al. 2019. "Stability Based Filter Pruning for Accelerating Deep CNNs". In: *Winter Conference on Applications of Computer Vision 2019*. IEEE. Waikoloa Village, pp. 1–9.

Tzelepis, G. et al. 2019. *Deep Neural Network Compression for Image Classification and Object Detection*. n.d.: n.p..

Wen, W. et al. 2016. "Learning Structured Sparsity in Deep Neural Networks". In: *NIPS*, pp. 2074–2082.

Zou, Z. et al. 2019. *Object Detection in 20 Years: A Survey*. n.d.: n.p.

Contact

Mail: yannik.fruehwirth@web.de

Code: `https://gitlab.com/yannik_2f/structuredFilterPruningForMaskRCNN`

PHYSICS-BASED MODELLING OF A MILK COOLING SYSTEM FOR INTELLIGENT ENERGY MANAGEMENT

Lars Kappertz[1] and Christof Büskens[1]

[1]Center for Industrial Mathematics, Universität Bremen, e-mail: kappertz@uni-bremen.de

KEYWORDS

Modelling, Optimization, Parameter Identification, Demand Side Management, Renewable Energy

ABSTRACT

Forecast-based energy management can play a large role in a smarter and more efficient use of renewable energies based on demand side management. Using approaches such as model predictive control, individual consumption devices can be shifted within operation constraints so that their electricity consumption optimally matches generation. In agriculture, large thermal storages make up a sizeable part of electricity consumption, and offer a potential use in the short term shifting of demand. Necessary for this are accurate models to forecast behaviour of such dynamic systems, so that minimal power demand and fulfilment of operation constraints can be ensured when computing optimal controls. This work focuses on the physics-based modelling of a milk cooling storage through parameter identification on real measurement data. Emphasized are the derivation of a suitable model ODE with regards to available data, and evaluation of the model on a rolling horizon. All major features of the measurement data can be recreated by the model forecasts, and model performance values show errors of around 30% relative to mean temperature. Model performance is considered suitable for use in energy management at least on short forecast horizons, while practicability on longer horizons is subject to further research.

INTRODUCTION

One of the main difficulties in the large scale integration of renewable energy sources like solar and wind plants is their varying and uncontrollable generation, dependent on the weather. Sufficient storage solutions are very expensive, and at least currently and in the near future not available on a large enough scale (Pickard et al., 2009). This leads to situations of either over- or underproduction of renewable energies, which on the grid level causes a need for (usually carbon-based) buffer generation, and in some situations also the temporary shut down of plants. Analogue situations also affect individuals generating their own energy, who face expensive import or unprofitable export of energy when self-consumption is not possible. One strategy to mitigate this problem is demand side management (DSM), which comprises strategies of influencing the energy consumption side of the situation, in order to have overall energy demand match the available generation as closely as possible at all times. On the scale of an individual household or enterprise, this can be achieved through an energy management system (EMS). Economic incentives making this worthwhile for individual households can derive from the difference between electricity price and feed-in tariff when producing own renewable energy, or from variable electricity pricing schemes, which are becoming an option with increasing relevance in Germany. Intelligent EMS approaches (sometimes also smart EMS) can include the forecast-based consideration and control over individual devices and storages. Where possible under operation constraints, device start times can be shifted to optimal times, continuous control values adapted or the operation of storage systems optimized, often in a model predictive control (MPC) scheme. Especially the autonomous and optimized management of larger consumers has gained increasing interest in research, including investigation of mathematical solution strategies for the ensuing problems in the fields of optimization, control and modelling. Burda et al. (2023) published approaches for the optimized control of thermal and electrical building energy supply with a Mixed-Integer Non-linear MPC. An agricultural setting stood in the centre of research projects *SmartFarm* and *SmartFarm2*, where a focus was put on the development of lightweight optimization algorithms for data-based modelling, scheduling and optimal control in a forecast-based EMS (Lachmann et al., 2020).

This work examines the modelling of thermal storages, needed for their use in an intelligent, forecast-based EMS. Application example is a milk cooling system, this being a large consumer of electricity on dairy farms. In order to fulfil operational constraints, the temperature behaviour of such a thermal storage needs to be forecast based on the applied controls. Lachmann and Büskens (2021) presented the use of data-based modelling for this and other storage devices, emphasizing the need for a sufficient basis of data when using purely data-based model approaches. Afram et al. (2014) reported good model performance for data-based, first order linear models for thermal storage tank data developed for a similar use case. Other

publications on the modelling of thermal storages in this context often base their models on consideration of the relevant physical processes, e.g. for a combined heat and power unit and heat storage tank (Bitner et al., 2021), or for a residential system including heat pump and thermal storage system Schütz et al. (2015). Such models can usually simulate the considered system well, but their application to specific real systems is dependent on expert knowledge, and usually not transferable. This work aims to bridge this gap between the purely data-based modelling of devices based on real-world measurements and the physical modelling of such systems in the context of energy management by employing widely used methods of parameter identification (PI) a for parametrized physical model of the dynamic system. Using the milk cooling system as an application example, approaches outlined in previous work (Kappertz and Büskens, 2023) are investigated and tested for real data. A complete workflow from the derivation of a phyics-based dynamic model, the fitting of its parameters in a PI problem and the evaluation of this model is presented. An emphasis lies on the difficulties of modelling and parameter identification with limited real-life data, and an evaluation of model performance through a simulation scheme motivated by the designated application in a forecast-based EMS.

THEORY AND METHODS

Parameter Identification of Dynamical Systems

This work focusses on the modelling of dynamical systems, whose behaviour is assumed to be representable through an ordinary differential equation (ODE) for the state derivative $\dot{x}(t) = f(x(t), u(t), t, p)$ as a function of current state of the system $x(t) \in \mathbb{R}^{n_x}$, any applied controls (external influences) $u(t) \in \mathbb{R}^{n_u}$, and a set of parameters $p \in \mathbb{R}^{n_p}$ describing the specific properties of the modelled system. An explicit time dependency may be accounted for in this ODE, but is not considered in the following.

Any future state of the system at time $t > t_0$ can then be computed through integration of the ODE from $x(t_0)$ for given controls $u(\tau), t_0 \leq \tau \leq t$, as

$$x(t) = \int_{t_0}^{t} f(x(\tau), u(\tau), p) \mathrm{d}\tau.$$

In order to do this, a reasonably accurate estimate of the 'true' parameters is needed, which is achieved by PI on measured data. Measured data $\{t_i, \bar{u}_i, \bar{x}_i \mid i = 0, 1, ..., n_t - 1\}$ contain control and state values measured at n_t time points. These are used to identify the optimal parameters p^* that minimize the error between measured and predicted state values in a (direct) PI problem of the general form

$$\begin{aligned} p^* = \underset{p \in \mathbb{R}^{n_p}}{\arg\min} \quad & \sum_{i=0}^{n_t-1} \sum_{j=0}^{n_x-1} (\bar{x}_{i,j} - x_j(t_i))^2 \\ \text{s.t.} \quad & \dot{x}(t) = f(x(t), \hat{u}(t), p), \quad t_0 \leq t \leq t_{n_t-1}, \end{aligned}$$

where for the error measure usually the quadratic norm is used, based on an assumed normal distribution of measurement errors. The dynamic model and the necessary integration can make this a non-linear and numerically intensive problem, for which different numerical solution schemes exist (Schittkowski, 2002). In the following, the approach of *full discretization* is used, in which the numerical integration of the ODE is included in the optimization problem as additional equality constraints at a specified number of discretization points (Schäfer et al., 2018). While this approach increases the overall dimension of the optimization problem, it eliminates the need for costly iterative integration steps. The model ODE not having to be fulfilled at all intermediate steps can also make it possible to avoid local minima in search of a better local or even global minimum (Wiesner and Büskens, 2023).

Milk Cooling System

A milk cooling system is a large agricultural thermal storage used on dairy farms to cool and store milk at conditions constrained by sanitary regulations. Its relevance for energy management stems from its large electricity consumption, and the fact that a margin in the temperature constraints allows shifting of this consumption in time. Generally, milk cooling systems have an internal on-off controller set to cool the fresh milk to around $5°\mathrm{C}$ until it is emptied. In this work, real measurement data previously discussed by Lachmann and Büskens (2021) are used, gathered on a dairy farm in Northern Germany. The milk cooling system is operated such that new milk is input twice per day (around 3500 and 2000 l), and emptied every two days. The data contain measurements of the temperature within the tank, as well as the electrical power used for cooling the milk, both available in minutely resolution. In the following, eight days of data from February and April are used. As visible in the measured power data (Fig. 2), the cooling aggregate kicks in twice per day when warmer fresh milk is entered, then averaging at $11.9\,\mathrm{kWh}$. An average daily consumption of $55\,\mathrm{kWh}$ is observed. The different phases of operation have clear influences on the temperature data, where the small, twice daily peaks of simultaneous milk inflow and cooling are alternated with long periods of no activity, during which ambient warming of the system seems small enough to not warrant additional cooling. Every two days, the cleaning period after the tank is emptied expresses itself in temperature peaks of more than $50°\mathrm{C}$, followed by a cooling-down period until the next milking in the morning.

Not all of these processes are covered satisfactorily in the available data. With the two measurements available, many aspects of the state behaviour, i.e. temperature, are not linked to control input, i.e. cooling power. Important external influences onto the system, like the cleaning process, or the adding of milk, are not available as control data. Following (Lachmann and Büskens, 2021) they are therefore substituted by 'auxiliary' controls derived in a preprocessing step from the

available data. Since the influence of these processes onto the temperature measurement is straightforward, simple conditions on the available data allow for the generation of additional boolean variables marking e.g. the external influence of milk inflow (defined by a rise in temperature when cooling power is active). Similarly, the filling level of the tank can be estimated based on the number of milking processes since last pickup. In this work, every process phase (like cleaning or milk inflow) is only marked by a single value at the beginning of the process. This assumption of instantaneous processes is made to keep preprocessing and model formulations simple. Overall, four substitute variables are generated, resulting in an augmented dataset of one state and five control measurements.

For external control by an intelligent EMS, a temperature margin for safe operation of the milk cooling system of $1°\mathrm{C}$ is assumed as a conservative estimate of what food safety permits. Based on a superficial physical consideration of the system, the $1°\mathrm{C}$ temperature margin for a tank filled for example with a volume of $V = 9000\,\mathrm{l}$ milk corresponds to a margin of approximately $\Delta Q = c \cdot V \cdot \rho \cdot \Delta T \approx 10\,\mathrm{kWh}$ in terms of heat energy Q, where c is the specific heat content of milk and ρ its density. The amount of electrical energy that could on short timescales be shifted to more optimal times depends on the coefficient of performance (COP) of the cooling aggregate used, defined as the fraction of heat removed per applied amount of energy. Generally, COP values can vary between up to 4 down to 1 (Mhundwa et al., 2017), which leaves a potential for short term load-shifting in the order of 2.5 to $10\,\mathrm{kWh}$.

In order to intelligently shift the electrical consumption of the milk cooling system, its behaviour needs to be predictable, to comply with the temperature constraints, and to assess future power demand accurately, as not to waste energy on over-cooling. A physical description of the relevant processes, and thus a basis for a model ODE can be derived from considering energy conservation of the relevant heat flows at time t as

$$\dot{Q}_\mathrm{m}(t) + \dot{Q}_\mathrm{a}(t) + \dot{Q}_\mathrm{w}(t) - \dot{Q}_\mathrm{c}(t) = \dot{Q}_\mathrm{internal}(t), \quad (1)$$

where $\dot{Q}_\mathrm{m}$ denotes the heat inflow when fresh milk is added, $\dot{Q}_\mathrm{a}$ the heat exchange with the environment, $\dot{Q}_\mathrm{w}$ the heat inflow of the hot water during the cleaning phase, and $\dot{Q}_\mathrm{c}$ the outflow of heat achieved by the cooling aggregate. Any accumulated heat flow affects the state of the system, and is 'stored' in form of a change in its internal heat content as

$$\dot{Q}_\mathrm{internal}(t) = C(t) \cdot \dot{T}(t) + T(t) \cdot \dot{C}(t),$$

dependent on system heat capacity C and temperature T. Eq. 1 can then be reformulated to yield an ODE for the temperature of the system as

$$\dot{T}(t) = \frac{\dot{Q}_\mathrm{m}(t) + \dot{Q}_\mathrm{a}(t) + \dot{Q}_\mathrm{w}(t) - \dot{Q}_\mathrm{c}(t) - T(t) \cdot \dot{C}(t)}{C(t)}. \quad (2)$$

The individual terms of this ODE are approximated to provide a model function to predict temperature behaviour based on the available augmented dataset. Time-dependent physical properties are thus approximated through time series data, but since the further use of the model function only involves numerical operations in a discretized setting, this is not an issue. Heat capacity of the system is approximated as

$$C(t) \approx C_\mathrm{t} + c_\mathrm{m} \cdot i_\mathrm{m}(t) \cdot V_\mathrm{m} \cdot \rho_\mathrm{m},$$

consisting of a constant term C_t for the heat capacity of the tank itself and a second term, varying with the amount of milk in the tank. Since no data on milk level or in- and outflow is available, this is expressed in terms of a 'counter' variable $i_\mathrm{m}(t)$, defining how many milking processes (i.e. from zero to four) have taken place since last emptying. For each milking process, a constant volume of added milk V_m is assumed, with constant material properties for specific heat content of milk c_m and milk density ρ_m. The heat flow from adding or removing milk is approximated as

$$\dot{Q}_\mathrm{m}(t) \approx (b_\mathrm{o}(t) \cdot T(t) + b_\mathrm{i}(t) \cdot T_\mathrm{m}) \cdot \dot{C}(t)/\Delta t_\mathrm{m},$$

relying on boolean variables $b_\mathrm{o}(t)$ and $b_\mathrm{i}(t)$ to mark times of milk out- and inflow, as well as assumptions of constant values for (fresh) milk temperature T_m and milking duration Δt_m, and the implicit assumption of perfect mixing. Again the above assumption of the varying milk content being the only time-dependent component of the overall systems heat capacity is used. Ambient heat exchange with the environment is approximated using Newton's law of cooling as

$$\dot{Q}_\mathrm{a}(t) \approx (h_\mathrm{c} + b_\mathrm{d}(t) \cdot \Delta h_\mathrm{d}) \cdot A \cdot (T_\mathrm{a} - T(t)).$$

Another boolean variable $b_\mathrm{d}(t)$ is used to describe times where the tank door is open, leading to an increase in heat exchange coefficient h, whose values in the two situations themselves are assumed to be constant. Also assumed constant is the tank surface area A, and – less likely to match reality – ambient temperature T_a. A more realistic approach including either explicit or implicit time dependency is neglected in favour of a simpler model function. Also the heat inflow from hot cleaning water is approximated only in a simple manner as $\dot{Q}_\mathrm{w}(t) \approx \frac{Q_\mathrm{w}}{\Delta t_\mathrm{w}} \cdot b_\mathrm{w}(t)$ by using binary variable $b_\mathrm{w}(t)$, describing times of active cooling, together with an assumedly constant amount of heat energy Q_w of said water delivered within cleaning duration Δt_w, also assumed constant. Finally, as described above, the heat extracted by cooling is approximated as $\dot{Q}_\mathrm{c}(t) \approx \mathrm{COP} \cdot P_\mathrm{c}(t)$, where, analogous to the assumption of constant external temperature, a constant COP is assumed. Using all approximations listed above in Eq. 2, a parametrized model ODE can now be based on the physical relationships as

$$\begin{aligned}\dot{x}_0(t) = \frac{1}{p_3 + u_4(t)} \cdot \Big(& p_4 \cdot u_2(t) - p_5 \cdot u_0(t) \\ & + p_1 \cdot u_3(t) \cdot (p_0 - x_0(t)) \\ & + (p_2 - x_0(t)) \cdot u_1(t) \Big),\end{aligned}$$

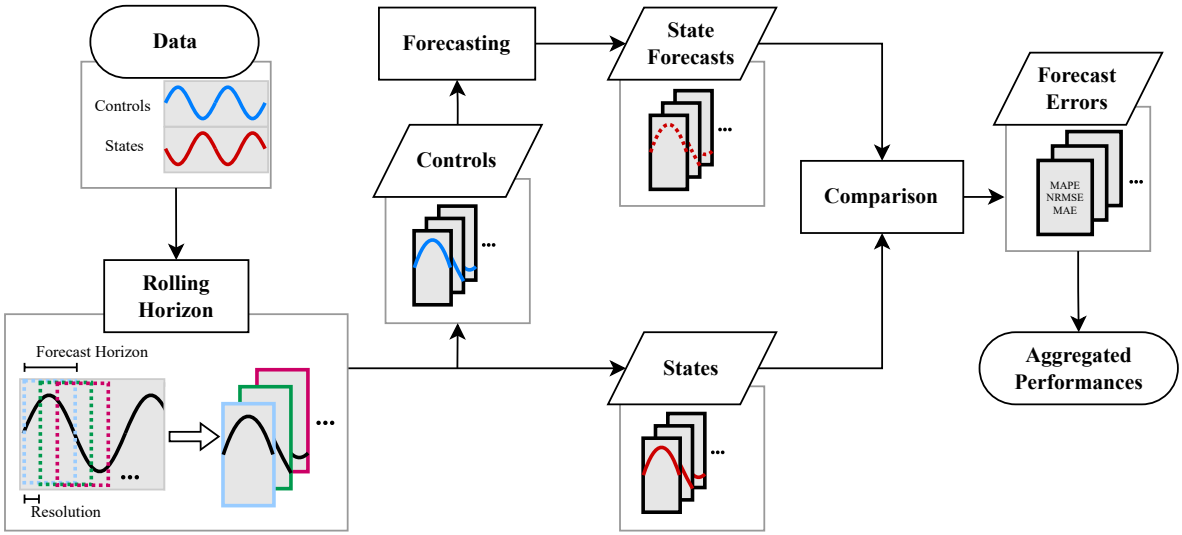


Figure 1: Overview of the Rolling Horizon Scheme for Model Evaluation

where the state $x_0(t)$ of the system is the temperature, and the controls of the system are the inputs of

$$u_0(t) = P_c(t), \quad u_1(t) = b_i(t)/\Delta t_m, \quad u_2(t) = b_w(t),$$
$$u_3(t) = b_d(t), \quad u_4(t) = i_m(t).$$

All physical constants (or variables assumed constant) are covered by six parameters, formulated so that structural identifiability is possible, i.e. such that there are not multiple combinations of parameters leading to the same result. For a complete overview of physical equivalents to each parameter, see Tab. 1.

Methodology

The augmented dataset is used to both fit and validate the presented physics-based model. Since the milk cooling system runs on a periodic cycle of two days, no large variance of behaviour is expected over longer durations. Therefore, only four days of measurement data are used for model training in the parameter identification step, and an equal length dataset is reserved for test purposes. In accordance to the designated use in an EMS, the training set is the data right up to the test set. In application, this should enable the model to account also for slowly varying external influences.

To solve the PI problem, initial parameter guesses are provided based on estimates of the physical properties they are based on (Tab. 1). Arbitrary initial guesses are possible – and would be more practical for real-life application of the forecast-based EMS approach – but can lead to longer computation times and higher risk of getting stuck in local minima. The PI problem is solved using the full discretization scheme implemented in the *Topas Model Fitting* tool (Wiesner et al., 2021). It transposes the problem into a high-dimensional, but sparse non-linear program (NLP), which the local solver *WORHP* solves with a Penalty-Interior-Point algorithm (Büskens and Wassel, 2013).

To use and evaluate the model, simple Euler integration is deemed accurate enough, since the model ODE is linear in its single state. Model evaluation is performed by comparing model forecast and measured data for the test set, but because of the dynamical nature of the model, and mirroring its intended online application in an EMS, this is performed on a rolling horizon. The model is intended for short term forecasts, and evaluation over the longer test set is carried out by iterative forecasting, as shown in Fig. 1. Given a fixed forecast horizon Δt_{fh} and a resolution of the rolling horizon of Δt_{RH}, the test set of length Δt_t is split into $n_{RH} = (\Delta t_t - \Delta t_{fh})/\Delta t_{RH}$ windows. For each, a forecast is generated by iterative integration of the model ODE using initial value and control values from measurements over the full forecast horizon. Predicted and measured state values are compared by computing mean average percentage error MAPE, and normalized root mean squared error NRMSE (normalized by mean measured value of the respective period) for the window. An additional error measure more specific to the use case in an EMS is computed as the mean average error of all those durations of interest for demand shifting, denoted as MAE_{EMS}. The durations relevant for energy management are those of normal operation outside of the cleaning and empty phases, and are here approximated as all times where $T \leq 10.0°C$. Repeating the procedure with the next window from time $t_{i+1} = t_i + \Delta t_{RH}$, the whole test set is covered, and all n_{RH} performance measures can be averaged into an aggregate performance measure.

For energy management, forecast horizons of 24 h are of interest, since relevant cycles in consumption and (solar) generation usually occur daily. Nevertheless, a running EMS would frequently re-evaluate the situation and therefore only the first time steps of generated control output would actually be applied before calculating new ones. Therefore model performance of the first few minutes and hours (depending also on the frequency with which the EMS updates) is much more relevant than model performance over the full 24 h. To account for this, different forecast horizons

Table 1: Model Parameters with Physical Equivalents and Derived Initial Guesses compared to the Identified Values

	Physical Equivalent	Unit	Initial Guess	Expected Range	Identified Parameter
p_0	T_a	°C	10.0	-10.0 to 30.0	19.1
p_1	$\frac{\Delta h_\mathrm{d} \cdot A}{c_\mathrm{m} \cdot V_\mathrm{m} \cdot \rho_\mathrm{m}}$	$1/\mathrm{s}$	$2.71 \cdot 10^{-5}$	0.0 to $1.81 \cdot 10^{-4}$	$9.60 \cdot 10^{-2}$
p_2	T_m	°C	20.0	10.0 to 30.0	18.8
p_3	$\frac{C_\mathrm{t}}{c_\mathrm{m} \cdot V_\mathrm{m} \cdot \rho_\mathrm{m}}$		$7.34 \cdot 10^{-5}$	0.0 to $1.31 \cdot 10^{-1}$	$2.93 \cdot 10^{-2}$
p_4	$\frac{\dot{Q}_\mathrm{w}}{\Delta t_\mathrm{w} \cdot c_\mathrm{m} \cdot V_\mathrm{m} \cdot \rho_\mathrm{m}}$	°C/s	$1.03 \cdot 10^{-5}$	$0.26 \cdot 10^{-5}$ to $3.42 \cdot 10^{-5}$	$4.33 \cdot 10^{-4}$
p_5	$\frac{\mathrm{COP}}{c_\mathrm{m} \cdot V_\mathrm{m} \cdot \rho_\mathrm{m}}$	°C/J	$2.05 \cdot 10^{-7}$	$0.52 \cdot 10^{-7}$ to $6.84 \cdot 10^{-7}$	$1.35 \cdot 10^{-7}$

from 24 h down to 1 h are evaluated, with a resolution of the rolling horizon of $\Delta t_\mathrm{RH} = 600\,\mathrm{s}$.

NUMERICAL RESULTS

The parameter identification problem is thus solved with 5760 points of measured data (4 days in minutely resolution). With a number of discretization points chosen as $n_\mathrm{dis} = 0.7 \cdot n_\mathrm{t}$, the NLP resulting from the full discretization scheme consists of 4037 variables, for which an optimal solution is found within $554\,s$. The identified parameters are displayed as well in Tab. 1, together with the initial guesses and expected range. Major displacement from the initial guesses by more than one order of magnitude is present only for p_1 and p_3. This is to be expected, since these parameters account for uncertain and very device-specific properties like heat capacity and heat transfer coefficients of the tank. Only in the cases of p_1 and p_4 do the identified parameters fall outside of the bounds of expected values. Whether this is due to the found local solution differing from a possible 'true' global solution, due to processes not considered in the formulation of the physical model ODE, or simply due to operation of the milk cooling system different to the expectations used in the initial guesses, is unclear.

In general, while the model function was derived using physical descriptions of the relevant processes, the overall model trained on data does not necessarily match these expectations of physical meaning. Processes not considered or unduly simplified in the model formulation, as well as simple measurement noise and other influences, are all included in the fitting of the parameters on the measured data, and can have a large influence on the identified values. The fit model should be checked and evaluated with the same scrutiny as any data-based model. A visual comparison of individual 24 h-forecasts against the measured data is presented in Fig. 2. For both training and test set, out of the $n_\mathrm{RH} = 432$ available forecasts, only those for $i = 0, 144, 288, 432$ are shown, leaving out all overlapping ones in between. The data show an accordance of modelled and observed temperature for all important operation phases.

Main discrepancies between model forecasts and measured temperatures occur during the cooling down phase after cleaning (which would not matter in a real EMS system), during the milking phases, where the simplifying assumption of instantaneous inflow is clearly visible (but this offset is limited to a short duration), and - especially on the test set - as constant offsets after the individual cooling phases. The latter observation is most relevant as a source of error in the overall model performance and especially for use in an EMS, since these are the periods where cooling activity could be shifted, and where an accurate estimate of temperature behaviour is needed. The reason for these offsets lies in the fact that no information about future (or even past) heat inflow from milking is available to the model. The consideration of this inflow with only a constant parameter obviously cannot account for the fluctuations present in the real operation. The different individual model forecasts shown for this period also display the dynamic nature of the model. Since future behaviour of the system depends on its current state, any error within a forecast propagates to all further values. Overall forecast error therefore can also depend on the starting point of the integration, as visible in the second and fourth day of the test period, which produce larger deviations than the first and third.

Tab. 2 shows mean and standard deviation of the performance measures for both training and test set of all individual forecasts within the rolling horizon, for forecast horizons of 24, 12, 6 and 1 h. As expected, model performance on the data in the training set is notably better than for the unseen data of the test set for all error measures. The standard deviations are used as secondary metrics for the distribution of performances, although interpretation is difficult since these are not necessarily normal distributions, but increasingly skewed when average model performance is better. Nonetheless, they indicate that within the set of forecasts of the rolling horizon, there are large differences in performance. This matches the observation of starting time (rather, initial state) having a large influence onto the forecast. Overall, the percentage model performances for full day forecasts of around 30 % on the test set confirm the visual impression of an adequate forecasting ability with room for improvement. This does however not provide any clear inferences for its designated application in an EMS. Values of the use case-specific $\mathrm{MAE}_\mathrm{EMS}$ provide a more tangible estimate here. Assuming a temperature margin of 1°C for EMS operation, forecast errors should be well below this bound, although there is no definite threshold. This is the case for 1 h forecasts, but the $\mathrm{MAE}_\mathrm{EMS}$ of 1.7°C on the test set observed for 24 h forecasts is very high for beneficial use in an EMS. Notably, the shorter

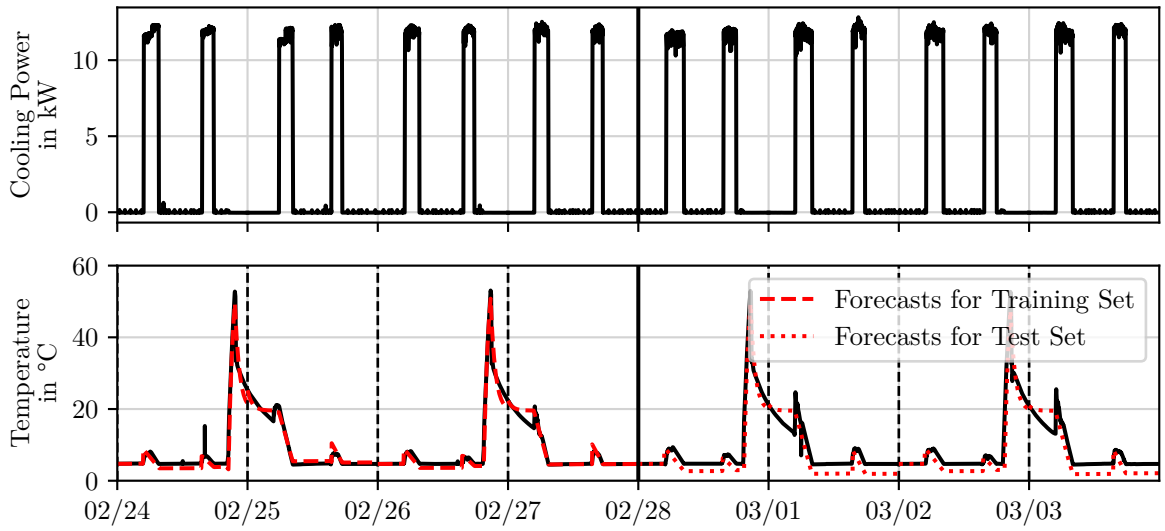


Figure 2: Individual 24 h-Forecasts of the Milk Cooling Model against Measured Data of Training and Test Set

Table 2: Mean Performance Measures of the Milk Cooling Model for Different Forecast Horizons

Δt_{fh}	Set	$\mu_{\text{MAPE}}[\%]$ in %	σ_{MAPE} in %	μ_{NRMSE} in %	σ_{NRMSE} in %	$\mu_{\text{MAE}_{\text{EMS}}}$ in °C	$\sigma_{\text{MAE}_{\text{EMS}}}$ in °C
24 h	Training	11.55	6.57	15.74	3.92	0.67	0.47
	Test	29.41	13.39	28.86	9.74	1.74	0.91
12 h	Training	10.46	9.07	15.30	7.19	0.54	0.50
	Test	22.44	18.52	25.83	13.58	1.32	1.20
6 h	Training	8.68	8.99	12.45	8.30	0.46	0.55
	Test	15.33	16.99	18.70	14.50	0.98	1.37
1 h	Training	4.28	6.22	5.55	8.51	0.24	0.46
	Test	5.95	8.94	7.59	12.07	0.38	1.08

forecasts generated in the rolling horizon evaluation seem to behave even more diversely than the full-day forecasts. The standard deviations of the performance measures increase with decreasing forecast horizons. This again shows the strong dependency on the starting point of the forecast, which is exacerbated by the additionally generated auxiliary controls being only instantaneous influences, even though in reality these processes last longer. The MAE_{EMS} decreases as expected with decreasing forecast horizon, and for Δt_{fh} shorter than six hours, the error is below the discussed bound of 1°C also for the test set. This does not necessarily mean that the model already allows for optimal energy management, but it shows the possible use of local models with realistic handicaps in an EMS.

CONCLUSION AND OUTLOOK

Accurate modelling of dynamical systems is a necessary but complex step in the development of a forecast-based EMS. To develop a model usable and transferable in practice, inclusion of device-specific measurement data in the form of PI or similar fitting approaches is unavoidable. The presented work shows that a physics-based approach to this can embed important information in the model to help performance even with lacking data. However, this also means a stronger reliance on expert knowledge, and makes the approach less transferable to other types of devices compared to purely data-based approaches. For the considered milk cooling system, forecasting quality is at least comparable to results from fully data-based methods published previously for the same dataset (Lachmann and Büskens, 2021), where even more information was provided as model inputs from preprocessing. A higher forecast performance through improvements in the model, intermediate model updates, or improvements in data preprocessing may be possible. The main obstacle of the data not containing all needed information about future heat inflow however remains.

The question of what level of model performance is needed for successful energy management cannot be answered definitely, but the constraints to be fulfilled during the operation of the milk cooling system provide some guidance. Assuming a conservative 1°C margin for EMS operation, the MAE_{EMS} computed for 24 h forecasts is quite large. For shorter forecast horizons, errors well below this goalpost are reached. In a frequently updating EMS, these short term forecasts would be most relevant and flow directly into control

values that are actually applied before the next iteration of the system. The question of whether optimizing energy use over a short horizon with good model forecasts, or over a longer horizon with less accurate model forecasts leads to better outcomes provides an interesting setup for further studies. Important future steps also include the direct comparison of models developed without any assumptions on physical behaviour of the system, and the transfer of both approaches to other relevant thermal storages. The development and test of an EMS using these methods hold their own challenges; Already now, however, the practicability and the possible benefits of intelligent energy management of large thermal storages in the agricultural domain can be shown, motivating further research and development.

ACKNOWLEDGEMENTS

This research is funded by the Federal Ministry for Economics Affairs and Climate Action of Germany within the project *SmartFarm2 - Autonomes EMS für den ländlichen Raum* (Ref.: 03EI6046B). Special thanks goes to all who made this work possible, including the technical staff of the University of Bremen.

References

A. Afram, F. Janabi-Sharifi, and G. Giorgio. Data-driven modeling of thermal energy storage tank. In *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5, Toronto, ON, Canada, May 2014. IEEE.

D. Bitner, A. Burda, and M. Grotjahn. Optimized supervisory control of a combined heat and power plant by mixed-integer nonlinear model predictive control. In *NEIS 2021; Conference on Sustainable Energy Supply and Energy Storage Systems*, pages 1–7, September 2021.

A. Burda, D. Bitner, F. Bestehorn, C. Kirches, and M. Grotjahn. Mixed-Integer Real-Time Control of a Building Energy Supply System. *IEEE Control Systems Letters*, 7:907–912, 2023.

C. Büskens and D. Wassel. The ESA NLP Solver WORHP. In G. Fasano and J.D. Pintér, editors, *Modeling and Optimization in Space Engineering*, Optimization and Its Applications, pages 85–110. Springer, New York, NY, 2013.

L. Kappertz and C. Büskens. Towards modelling of energy storages for use in an intelligent energy management system. *PAMM*, 22(1): e202200257, 2023.

M. Lachmann and C. Büskens. A Hybrid Approach for Data-Based Models Using a Least-Squares Regression. In *Optimization and Learning*, Communications in Computer and Information Science, page 62–73, Cham, 2021. Springer International Publishing.

M. Lachmann, J. Maldonado, W. Bergmann, F. Jung, M. Weber, and C. Büskens. Self-Learning Data-Based Models as Basis of a Universally Applicable Energy Management System. *Energies*, 13(8):2084, April 2020.

R. Mhundwa, M. Simon, and S. Tangwe. Comparative analysis of the coefficient of performance of an on-farm direct expansion bulk milk cooler. pages 1–7, August 2017.

W.F. Pickard, A.Q. Shen, and N.J. Hansing. Parking the power: Strategies and physical limitations for bulk energy storage in supply–demand matching on a grid whose input power is provided by intermittent sources. *Renewable and Sustainable Energy Reviews*, 13(8):1934–1945, October 2009.

K. Schittkowski. *Numerical Data Fitting in Dynamical Systems*, volume 77 of *Applied Optimization*. Springer US, Boston, MA, 2002.

K. Schäfer, M. Runge, K. Flaßkamp, and C. Büskens. Parameter Identification for Dynamical Systems Using Optimal Control Techniques. In *2018 European Control Conference (ECC)*, pages 137–142, June 2018.

T. Schütz, R. Streblow, and D. Müller. A comparison of thermal energy storage models for building energy system optimization. *Energy and Buildings*, 93:23–31, 2015.

M. Wiesner and C. Büskens. Benchmarking solution methods for parameter identification in dynamical systems. *PAMM*, 23(2):e202300134, 2023.

M. Wiesner, K. Schäfer, W. Bergmann, A. Berger, P. Shulpyakov, C. Dittert, and C. Büskens. Analyzing the Influence of Measurements in Dynamical Parameter Identification Using Parametric Sensitivities. *IFAC-PapersOnLine*, 54(14):7–12, 2021.

AUTHOR BIOGRAPHIES

LARS KAPPERTZ studied Physics at the University of Bremen. After obtaining his M.Sc. in Physics in 2020, he joined the working group Optimization and Optimal Control under Prof. Büskens, where he works on applications of industrial mathematics in the field of energy and environment.

CHRISTOF BÜSKENS gained his PhD at the Institute for Numerical Mathematics of the University of Münster, from where his career led him to the University of Bayreuth and finally to the University of Bremen. Here, since 2004, he holds a full professorship in applied mathematics and heads the working group of Optimization and Optimal Control within the Center for Industrial Mathematics.

DIGITAL TWIN DRIVEN ASSEMBLY LINE RE-BALANCING AND DECISION SUPPORT

Giovanni Lugaresi[1], Kovacs Laszlo[2], and Kornel Tamas[3]

[1]Department of Mechanical Engineering, KU Leuven, Leuven, Belgium, e-mail: giovanni.lugaresi@kuleuven.be
[2]Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary, e-mail: kovacs.laszlo@vik.bme.hu
[3]Department of Machine and Industrial Product Design, Budapest University of Technology and Economics, Budapest, Hungary, e-mail: tamas.kornel@gt3.bme.hu

KEYWORDS

Digital twins; assembly lines; production control; Industry 5.0; work assignment; re-balancing.

ABSTRACT

Recent investments in industrial digitization together with the concrete need for short-term planning capabilities mean digital twins can effectively aid enterprises in the management of their production systems and value chains. This paper introduces a conceptual framework for assembly line re-balancing in the context of Industry 5.0, focusing on manual assembly processes. The framework aims to leverage a digital twin for obtaining a synchronized representation of the current task allocations in the assembly line, and uses data-driven scenario generation methods for investigating alternative balancing solutions that are proposed to operators in real time. A proof-of-concept platform is implemented in a laboratory environment, utilizing an assembly line with industrial components. Preliminary results demonstrate the compatibility of the proposed components within the digital twin framework. The potential applicability to various manual assembly scenarios is discussed, along with considerations for incorporating additional constraints in the evaluation process.

INTRODUCTION

Manufacturing companies are recognized for delivering high-quality products tailored to customer specifications. Achieving stringent standards while managing product and process variability is challenging, also considering that companies face with elevated costs of labor. This is particularly crucial in production sectors with substantial labor involvement and operating under takt time, such as assembly lines. Meanwhile, resilience, sustainability, and human-centric production are the cornerstones of the new Industry 5.0 paradigm (Ivanov, 2023), which focuses to improve production processes involving interactions with human workers.

Assembly is the pivotal production process for generating products with numerous variations, relying on a relatively small set of standardized modules. Assembly processes heavily rely on specialized human labor, with nearly a third of employees in the European metal and electronic industry engaged in assembly work (Monika, 2021). As product intricacy continues to grow, accompanied by a shrinking pool of human resources and escalating labor costs, the efficient utilization of human labor in assembly processes emerges as a critical factor for European industry and its global competitiveness. Given the high cost of labor, a typical goal is to minimize the under-utilization of the human workforce, which is challenging given the diversity of assembly tasks and their resulting processing time variability from one order to another. Line balancing involves grouping assembly processes, assigning them to workstations and workers, defining workstation borders, and establishing the takt time. It is among the planning steps with the most significant impact on worker efficiency (Becker and Scholl, 2006; Battini et al., 2020).

Despite advances brought by recent approaches exploiting hardware reconfigurations and flexible layouts (Hottenrott and Grunow, 2019), modern assembly systems face new challenges. First, the push toward high customization means a high variability of product variants. Several companies aim for one-piece-flow capable systems. However, the production control algorithms are typically not designed to be dynamically changed, potentially even after each work-piece. Also, the ever-increasing diversity in the workforce (ageing, re-training, and re-skilling are some of the main factors) results in high variability of worker's productivity levels, which is visible even within a single day (Lassila et al., 2004; Silva et al., 2013). The result is that an optimal solution with a certain workers' setup may result sub-optimal with another. Smart system and production assets reconfigurations are needed to face this challenge, while monitoring and supervision

systems alone are not enough. There is the need to provide a proactive digital-to-physical interaction. In order to correctly evaluate alternative production scenarios, an up-to-date model of the real system must be readily available whenever needed. Also, a smart decision support system must be able to quickly learn the system's features and to generate alternative scenarios autonomously, and the insights learned from the digital world should be readily applicable in the physical system. This paper presents a research project aiming to overcome the aforementioned limitations with the proposal of a digital twin (DT) framework for manual assembly line re-balancing and online reconfiguration.

PROBLEM STATEMENT

As manufacturers face an aging workforce and seek to implement more flexible work shift models in assembly, recent approaches propose fixed shift schemes but in more adaptable settings, where there is no fixed start and end of shift and workers are allowed to engage at various times, receiving real-time workload assignments (Boysen et al., 2022). However, these approaches remain at a conceptual level: to be effectively applied they necessitate the capability to dynamic rebalance and to re-allocate tasks between available workers, while considering factors such as worker qualifications, learning curves, and the current system configuration to search for an optimal one (Cimen et al., 2022). The goal is to ensure a smooth adaptation process without overburdening the existing workforce, allowing for flexible cycle time adjustments based on the current system state.

A correct and optimal (re)-balancing and task allocation solution can only be obtained and verified if a valid model of the system is available and synchronized with the system state. These are the characteristics of simulation-based DTs (Tao et al., 2018). Indeed, by simulating production runs, planners can forecast the impact of production modifications beforehand (Monostori et al., 2016). Virtualization also aids in tracking parts and products in real time throughout the production process (Uhlemann et al., 2017). Recent approaches demonstrated the capabilities of generating accurate digital models starting from available data in business and production processes (Van Der Aalst and van der Aalst, 2016). Process mining emerges as a valuable tool in the realm of digital twinning, proving its efficacy in several tasks such as model generation, trace profiling, and performance evaluation. The capability for an automated generation of a simulation model paves the way to achieving a DT that is able to provide in real-time actionable insights to a manufacturing system (Lugaresi and Matta, 2021; Pourbafrani and van der Aalst, 2023). Complex manufacturing processes such as assembly are characterized by multiple interacting objects, which necessitates the application of novel approaches based on object-centric process mining (Lugaresi and Matta, 2023). However, the application of such model generation techniques to assembly system remains limited to synthetic data (Lugaresi and Matta, 2023). Besides, to the best of our knowledge there have been no implementations of complete DT architectures (i.e. from shop floor data to actions) in real environments for production planning and control applications. The current methods often involve the presence of managers/operators to define the search space and the desired performance levels. This slows down the decision-making process, together with the risk of obtaining sub-optimal solutions. Currently, the maturity/automation levels of a DT have been defined (Uhlenkamp et al., 2022) but not implemented, nor benchmarked with existing case studies. Current scenario generation approaches are limited to optimization approaches (Bounitsis et al., 2022) while there are no clear contributions that integrate with simulation-based methods and keep into account the real-time constraints for short-term decision making. Hence, there is the need to provide insights from test cases with real industrial equipment.

The research project presented in this paper aims to design and implement within a proof-of-concept platform a flexible DT of supervised assembly processes. The project takes as reference a setup available at the Industry 4.0 laboratory of Budapest University of Technology and Economics. The setup is a replica of a real assembly process in the automotive sector. Within this setting, the project aims at (1) testing a state of the art model generation technique with a real system dataset (i.e. industrial logic controllers and Manufacturing Execution System) and gather insights and data requirements from the behavior of model generation techniques facing unpredictable changes in a flexible assembly system; (2) designing and testing a method for feature's extraction and automated scenario generation, to proactively investigate if better solutions can be achieved on the line in real time; (3) designing and testing a method for DT-driven decision support for flexible assembly systems. The goal is to continuously compare the proposed solutions from the DT and provide actionable insights online to reconfigure the production. This paper contributes in (1) outlining the research project with respect to the literature, (2) presenting the main steps of the proposed methodology, and (3) presenting preliminary results obtained in the available setup.

RELATED WORK

In order to select relevant papers and worthy a comparison with this work, the following query has been done on Scopus on 9-Feb-2024: *"digital twin"* AND *"assembly"* AND (*"re-balancing"* OR *"balancing"*). The search results in 13 papers[1]. Among them, 8 papers are considered based on these criteria: (1) the paper regards production systems (2) the DT is used during the production operations, (3) the DT purpose is faithful to the definition (i.e., feedback loop of information). Also, one paper has been added following a

[1]List of papers: `http://tinyurl.com/sota-dt-assembly`

forward-backward citation analysis. Table 1 collects the selected contributions for comparison.

Selected Works

Zhang et al. (2022) introduced a reconfiguration framework for assembly lines with frequent changeovers, exploiting DTs based on an open architecture for both equipment and the assembly line, facilitating swift physical reconfiguration. The proposed technique is based on the idea of using DT-based predictive simulation for testing possible changeover schemes to reduce changeover times in real cases. An optimal reconfiguration algorithm is presented which employs analytical target cascading to address the joint optimization of order scheduling, line balancing, and buffer allocation.

Pabolu and Shrivastava (2021) proposed a dynamic solution to the worker assignment assembly problem. The fatigue-inducing factors are detected from the workers and classified, then forwarded to a worker-job rotation search algorithm. The algorithm generates recommendations for the production supervisor and suggests optimal dynamic solutions for worker job rotation or reallocation based on fatigue details.

Pabolu et al. (2022) leveraged DT for predicting the task execution time of an assembly line and integrating the forecasted values in an assembly line work assignment framework. The authors propose a human-in-the-loop decision making cycle, in which task execution times are estimated online and used to update alternative assignments evaluated by an heuristic algorithm. The results are then proposed to a line supervisor for reconfiguration decisions.

Yang et al. (2021) addressed the Assembly Line Worker Assignment and Balancing Problem (ALWABP) in which task processing times are influenced by the skill levels of the workers. The authors introduced positional constraints in the ALWABP and presented two mathematical programming models to allocate workers and tasks that address cases in which either new products are introduced or a worker is temporarily absent or leaves a position. The proposed approach is also integrated with a real-time dashboard.

Zhang et al. (2023) explored an adaptable scheduling technology employed in a manual assembly workshop. An analysis is conducted on the DT environment model designed for scheduling: a hierarchical reinforcement learning algorithm is developed to enhance the dynamic adjustment efficiency of the assembly workshop system.

Ragazzini et al. (2021) presented a preliminary DT to solve a a real-time balancing problem in a learning factory based on triggers coming from machine error states. The application of the DT reportedly improved lead time and utilization performance measures when facing unpredicted disruptions.

Xu et al. (2023) established a DT model for a robotic mixed-model assembly line. A mathematical model is developed to minimize reconfiguration costs and optimal load balancing. Authors employ an adaptive neighborhood search algorithm to solve the reconfiguration problem.

Santos et al. (2023) presented a strategy integrating Deep Reinforcement Learning into a Digital Twin framework to establish a training environment closely resembling real-world conditions and employing Discrete Event Simulation to replicate the dynamic aspects of the production system. The experiments demonstrated the effectiveness of dynamic resource allocation to tasks and workers.

Aslan et al. (2023) introduced an approach that utilizes ultra-wideband data for the discovery of process models in manufacturing activities through process mining techniques. The methodology was implemented on an actual assembly line and used to reveal deviations from the prescribed process steps and to identify bottlenecks.

Contribution

Based on the selected papers, we can conclude that: (1) most papers focus solely on either assembly line balancing or workers assignment separately, while the dynamic joint reconfiguration is not addressed, (2) all the works assume the structure of the system and production steps are fixed, and (3) none of the existing papers proposes a specific method to quickly generate alternative scenarios. In this work, we propose a framework that includes the combined assembly line balancing and worker assignment, in which the assembly steps are not preliminary known or may change dynamically during the production. Also, the processing times may be subject to variations (e.g., due to worker's fatigue). The dynamically changing conditions result in the need for model generation capabilities to grasp the dynamics of the system and achieve a higher physical-to-digital fidelity and synchronization.

PROPOSED METHODOLOGY

This research project proposes a methodology for enabling and enhancing the DT-based decision support system on a flexible assembly line. It takes as reference a multi-stage assembly line supervised by a camera for parts recognition. Namely, the operators are guided in picking components from the components kits and a camera-based supervisory system controls the correct assembly sequence and tracks the assembly steps in a time-series database. The system is flexible: both the assembly kits and the operators can be easily moved and interchanged along the line. For instance, the operator of the first station can either assemble all the parts in the kit or a portion of them, leaving the rest for the next operator. The choice depends on the current line balancing policy, the operators' availability and skill-set, together with other contingent factors. As a result, the line can potentially adopt several configurations with a different grouping of assembly steps. Such system can be considered as representative of a large set of other processes that can flexibly change between production resources and operators (Colledani

Table 1: Related works.

Reference	Scope	Method	Input Data	Model Generation	Scenario Generation	Decision Support
Zhang et al. (2022)	Reconfigure System	Simulation	Start-End Timestamps	-	-	-
Pabolu and Shrivastava (2021)	Line Re-Balancing	Heuristic Algorithm	Fatigue	-	-	•
Pabolu et al. (2022)	Worker Assignment	Statistical learning	IoT Sensor Data	-	-	•
Yang et al. (2021)	Worker Assignment & Balancing Problem	Mathematical Programming	Workers' presence & position	-	-	•
Zhang et al. (2023)	Rescheduling	Machine Learning	Multiset	-	-	•
Ragazzini et al. (2021)	Line Re-Balancing	Mathematical Programming	Error States	-	-	•
Xu et al. (2023)	Reconfigure System	Heuristic Algorithm	Multiset	-	-	•
Santos et al. (2023)	Line Re-Balancing	Reinforcement Learning	System State	-	-	-
Aslan et al. (2023)	Process Discovery	Process Mining	Workers Position	-	-	-
This work	Line Re-Balancing & Worker Assignment	Process Mining, Machine Learning	Video Sequences	•	•	•

et al., 2018). The methodology proposed within this research project is summarized in Figure 1.

Data-Driven Model Generation

The information system (i.e., MES, ERP) is responsible to provide tracking data of the assembly operations recorded by the supervision system. First, in order to grasp the system state and the current configuration of the line, the data-driven model generation technique proposed by Lugaresi and Matta (2021) is used to exploit shop floor data for generating a graph-based model of the system. The graph model effectively represents the skeleton of a DT, as it can be converted in a forward-looking model such as discrete-event simulation (Pourbafrani and van der Aalst, 2023). The system parameters are also reflected in the simulation model from the available datasets (e.g., statistical process time distributions). Starting from the as-is model, a continuous performance evaluation process is activated. The simulation model is used to estimate future system performance (e.g., throughput). This is done both in a rolling horizon approach or based on triggers such as a lower expected production output with the current configuration.

Scenario Generation

If the system performance is unsatisfactory as shown by significant deviations from the expected production levels, alternative scenarios are generated for evaluation in the digital realm. Alternatives involve both the shift and re-organization of assembly steps along the stations and the operators order and sequence along the line. The scenarios are generated via a smart exploration of the solution space. For instance, if a certain setup has been observed to be productive in the past (either in the real or in the digital world), the alternative scenarios are generated as close as possible to such configuration. Also, given the availability of a graph-model as DT skeleton, promising solutions can be identified as near to the current situation. For instance, the shift of one task from one operator to the next one can be generated and a reasonable number of solutions can be quickly explored.

Decision Support

The expected performance of each identified scenario is evaluated digitally via simulation experiments. The optimal scenarios are implemented automatically by readily giving operators the indications for a re-setup via the dedicated human-machine interfaces. If changes in the assembly kit are required, respective line operators are informed and operate the change at the next available occasion (e.g., change of shift).

CASE STUDY

Figure 2 shows the available setup at the Industry 4.0 laboratory at the Budapest University of Technology and Economics. The decision-support system proposed in this research project will be installed on the same line, and will provide operators with real-time operational instructions on how to (re)-balance and (re)-allocate the assembly operations.

Test Assembly Line

The purpose of the test assembly line is to be able to test manual assembly processes by continuously monitoring the process and collecting data for its optimization. The line includes three workstations for assembling, beside a place for logistics operations and a quality assurance workplace. The line itself is accessible from both sides. The back side of the line is designed for logistics staff. Here, the logistics operator can create a small buffer storage from the parts arriving from the warehouse by AGV, and then arrange the parts organized in kit trays, one for each of the three

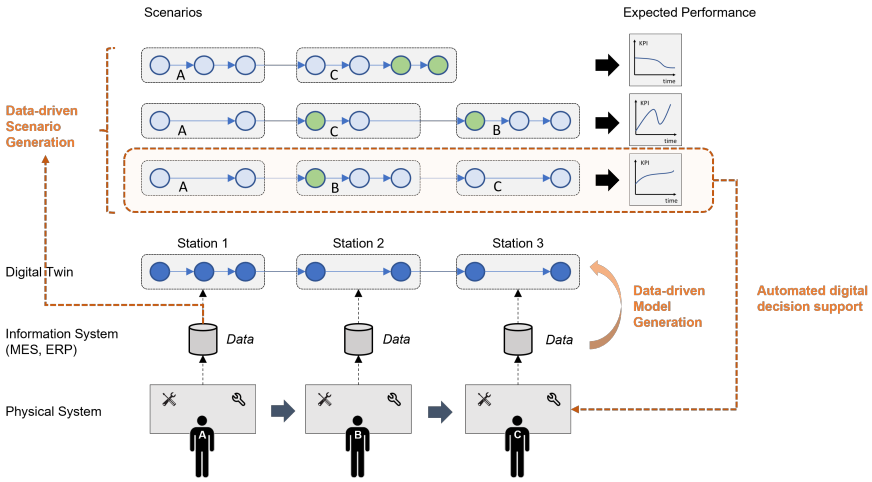


Figure 1: Schematic view of the proposed methodology.

Figure 2: The supervised manual assembly stations (3 operators) in the Industry 4.0 laboratory.

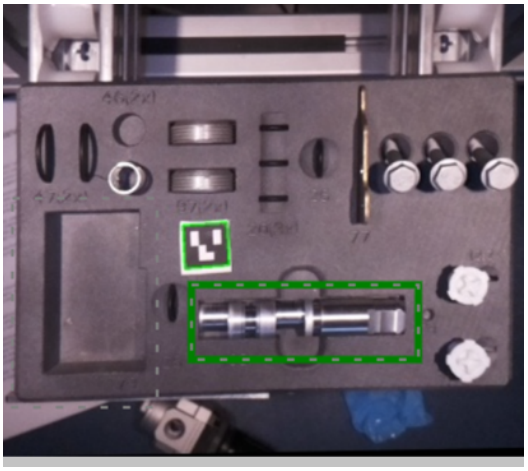


Figure 3: The assembly kit and the identification of components (green box) via the camera according to the assembly sequence.

assembly station. The kit trays are delivered to the operators via a series of inclined rollers. After completing the assembly, the operator returns the empty kit trays on another row of rollers inclined in the opposite direction. During supervised production, the logistics operator receives on-screen instructions on which parts to place on trays 1, 2, and 3, which may change from product to product as a result of optimization.

The front side of the row hosts the workplaces of the the three operators. The kit trays coming from the logistics side stop in a fixed position at the end of the roller row, from where the operator can take out the parts according to the assembly process one by one. A Raspberry PI-based camera units are placed above the kit trays, which monitors the kit trays in real time. Each station is equipped with a monitor which displays the necessary information for the operator and, if necessary, for the shift manager. Additionally, an Andon lamp is placed above each workstation and provides clear visual signals about the current state of the workplace. From the assembly line, the finished workpieces are moved to the quality control workplace, from where the finished products can be placed directly on the AGV.

The assembly process is supported by the Production Line Monitoring (PLM) system. The operators have access to a real-time image of the kit tray on their monitor. When the assembly process starts, the application marks the next component to be installed on the kit tray with a green frame (Figure 3). After the operator has removed the appropriate part from the tray, the image recognition system records the removal of the part as an event with a time stamp in the database. Then, the PLM system marks the next component to be used for the operator. As soon as the operator has completed all the assigned assembly steps, the finished part is placed in the buffer zone between the two workplaces, where the number of waiting parts (depending on the part) is counted with different sensors. The operator can then remove the empty tray from the roller line, thereby sliding the next filled tray into place, and then returns the empty tray to the logistics operator on

Table 2: Extract of data collected from the camera and parts recognition system indicating which component was picked from the assembly kit (assembled parts number 177 and 178).

Time-stamp	Part-Id	Assembly-Name
...	...	...
2/9/24 3:34:17 PM	177	Housing
2/9/24 3:34:18 PM	177	Shaft
2/9/24 3:34:21 PM	177	Screw
2/9/24 3:34:22 PM	177	Spring
2/9/24 3:34:23 PM	177	Plug
2/9/24 3:34:59 PM	178	Housing
2/9/24 3:35:00 PM	178	Shaft
...	...	...

the outgoing roller line.

Collected Data

The data in Table 2 has been extracted in a preliminary experiment, which demonstrates the capability to collect data in real time during the manual assembly operations. The datasets are compatible with model generation techniques and can be used to generate an as-is model of the line. These data demonstrate an ideal usage scenario and do not contain incorrect processes or faulty detection. Each row represents the removal of an item from the tray and therefore, in ideal conditions it corresponds to an assembly step. The columns used in this project are the following: (1) *Time-stamp*: it indicates the date and time of an assembly event; (2) *Part-Id* denotes a specific assembly part. when a product is assembled, a new identifier is created and then used within each step. (3) *Assembly-Name*, which is the specific assembly process which is being performed.

Preliminary Results

A preliminary test has been executed by using the available data set to generate a process model via a process mining-based algorithm (Lugaresi and Matta, 2021), which demonstrated the possibility to generate reliable models describing the real-time system conditions from the available camera setup (Figure 3). Figure 4 shows the obtained process model from the preliminary dataset. The following considerations can be formulated. The video-based acquisition system is sufficient to compose an event log compatible with process mining techniques. The assembly steps are visible in the generated model, as well as the current balancing solution. System performance measures (e.g., activity duration, queuing times) are easily grasped by the generated models and can be readily used to generate alternatives. Process- and resource-based constraints are not visible in the generated representations, and need to be integrated with additional datasets. The generated model is effectively used to evaluate the system performance with the current task assignments. Given the successful model generation on one station, the

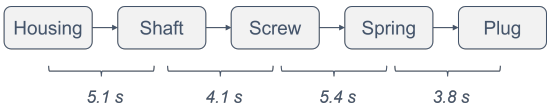


Figure 4: Process model discovered from the preliminary data on station 1 (times are mean values).

next steps within the project involve the generation of a complete model of the whole assembly process and the data-driven generation of line balancing solutions.

FINAL REMARKS

This work proposes a DT driven conceptual framework for assembly line re-balancing in the context of Industry 5.0. Operators can verify the current line configuration based on data driven modeling and performance estimation of alternative line balancing solutions. Preliminary results demonstrated the capability to interface a camera-based real-time supervision system to a model generation procedure.

This research project is subject to several limitations. The literature review can be extended to include works from related fields, such as free-flow production. The proposed framework is incomplete as it lacks of interfaces between components. Also, extensive tests with assembly operators and serial production conditions should be conducted to further validate the proposed setting. The assembly line studied in this work is representative of only a subset of possible production scenarios, with limited constraints that might emerge in other environments. It is well known that line balancing must strictly respect precedence, skill-based (e.g., learning curves), and technical constraints. The inclusion of all these limitations in an automated digital evaluation is not trivial, and will require further work. Future work should also investigate the feasibility of the proposed framework in other production environments with manual operations, such as re- and demanufacturing systems, u-shaped assembly lines, as well as human-robot collaborative settings.

ACKNOWLEDGEMENTS

This paper was supported by: the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (project no. TKP-6-6/PALY-2021), the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund (TKP2021-NVA funding scheme), the National Research Development and Innovation Fund (grant nr. ÚNKP-23-5-BME-80.), the Hungarian Scientific Research Fund (NKFIH FK-146067), the Ministry of Culture and Innovation and the National Research, Development and Innovation Office within the Cooperative Technologies National Laboratory of Hungary (grant nr. 2022-2.1.1-NL-2022-00012). Special thanks to the development team of the Technology Center, and personal thanks to Dániel Bálint and Péter Tkalcec.

References

Ayse Aslan, Hanane El-Raoui, Jack Hanson, Gokula Vasantha, John Quigley, and Jonathan Corney. Data-driven discovery of manufacturing processes and performance from worker

localisation. In *International Conference on Flexible Automation and Intelligent Manufacturing*, pages 592–602. Springer, 2023.

Daria Battini, Serena Finco, and Fabio Sgarbossa. Human-oriented assembly line balancing and sequencing model in the industry 4.0 era. *Scheduling in Industry 4.0 and Cloud Manufacturing*, pages 141–165, 2020.

Christian Becker and Armin Scholl. A survey on problems and methods in generalized assembly line balancing. *European journal of operational research*, 168(3):694–715, 2006.

Georgios L Bounitsis, Lazaros G Papageorgiou, and Vassilis M Charitopoulos. Data-driven scenario generation for two-stage stochastic programming. *Chemical Engineering Research and Design*, 187:206–224, 2022.

Nils Boysen, Philipp Schulze, and Armin Scholl. Assembly line balancing: What happened in the last fifteen years? *European Journal of Operational Research*, 301(3):797–814, 2022.

Tolga Cimen, Adil Baykasoğlu, and S Akyol. Assembly line rebalancing and worker assignment considering ergonomic risks in an automotive parts manufacturing plant. *International Journal of Industrial Engineering Computations*, 13 (3):363–384, 2022.

Marcello Colledani, Anteneh Yemane, Giovanni Lugaresi, Giovanni Borzi, and Daniele Callegaro. A software platform for supporting the design and reconfiguration of versatile assembly systems. *Procedia CIRP*, 72:808–813, 2018.

Andreas Hottenrott and Martin Grunow. Flexible layouts for the mixed-model assembly of heterogeneous vehicles. *OR Spectrum*, 41(4):943–979, 2019.

Dmitry Ivanov. The industry 5.0 framework: Viability-based integration of the resilience, sustainability, and human-centricity perspectives. *International Journal of Production Research*, 61(5):1683–1695, 2023.

Anna M Lassila, Sameh M Saad, Terrence Perera, Tomasz Koch, and Jaroslaw Chrobot. Modelling and simulation of human-centred assembly systems-a real case study. In *International Conference on Information Technology for Balanced Automation Systems*, pages 405–412. Springer, 2004.

Giovanni Lugaresi and Andrea Matta. Automated manufacturing system discovery and digital twin generation. *Journal of Manufacturing Systems*, 59:51–66, 2021.

Giovanni Lugaresi and Andrea Matta. Automated digital twin generation of manufacturing systems with complex material flows: graph model completion. *Computers in Industry*, 151: 103977, 2023.

KISS Monika. The future of work: Trends, challenges and potential initiatives. *EPRS: European Parliamentary Research Service*, 2021.

László Monostori, Botond Kádár, Thomas Bauernhansl, Shinsuke Kondoh, Soundar Kumara, Gunther Reinhart, Olaf Sauer, Gunther Schuh, Wilfried Sihn, and Kenichi Ueda. Cyber-physical systems in manufacturing. *Cirp Annals*, 65(2): 621–641, 2016.

Venkata Krishna Rao Pabolu and Divya Shrivastava. A dynamic job rotation scheduling conceptual framework by a human representing digital twin. *Procedia CIRP*, 104:1367–1372, 2021.

Venkata Krishna Rao Pabolu, Divya Shrivastava, and Makarand S Kulkarni. A digital-twin based worker's work allocation framework for a collaborative assembly system. *IFAC-PapersOnLine*, 55(10):1887–1892, 2022.

Mahsa Pourbafrani and Wil MP van der Aalst. Data-driven simulation in process mining: introducing a reference model. *Communications of the ECMS*, 37(1):411–420, 2023.

Lorenzo Ragazzini, N Saporiti, Elisa Negri, Tommaso Rossi, Marco Macchi, Giovanni Pirovano, et al. A digital twin-based approach to the real-time assembly line balancing problem. In *Proceedings of the 2nd International Conference on Innovative Intelligent Industrial Production and Logistics*, pages 93–99, 2021.

Romão Santos, Catarina Marques, César Toscano, Hugo M Ferreira, and Joel Ribeiro. Deep reinforcement learning-based approach to dynamically balance multi-manned assembly lines. In *International Conference on Flexible Automation and Intelligent Manufacturing*, pages 633–640. Springer, 2023.

E Silva, M Donauer, Américo Azevedo, P Peças, and E Henriques. A case study evaluating the impact of human behavior on a manufacturing process in-line with automatic processes by means of a simulation model. In *2013 IEEE International Conference on Industrial Engineering and Engineering Management*, pages 145–149. IEEE, 2013.

Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, 94(9):3563–3576, 2018.

Thomas H-J Uhlemann, Christian Lehmann, and Rolf Steinhilper. The digital twin: Realizing the cyber-physical production system for industry 4.0. *Procedia Cirp*, 61:335–340, 2017.

Jan-Frederik Uhlenkamp, Jannicke Baalsrud Hauge, Eike Broda, Michael Lütjen, Michael Freitag, and Klaus-Dieter Thoben. Digital twins: A maturity model for their classification and evaluation. *IEEE Access*, 10:69605–69635, 2022.

Wil Van Der Aalst and Wil van der Aalst. *Data science in action*. Springer, 2016.

Wenjun Xu, Zhihao Li, Jiayi Liu, Jia Cui, and Yang Hu. Virtual reconfiguration method of robotic mixed-model assembly line using bees algorithm based on digital twin. *IEEE Transactions on Automation Science and Engineering*, 2023.

Hyungjoon Yang, Je-Hun Lee, and Hyun-Jung Kim. Assembly line worker assignment and balancing problem with positional constraints. In *IFIP International Conference on Advances in Production Management Systems*, pages 3–11. Springer, 2021.

Ding Zhang, Jiewu Leng, Min Xie, Hong Yan, and Qiang Liu. Digital twin enabled optimal reconfiguration of the semi-automatic electronic assembly line with frequent changeovers. *Robotics and Computer-Integrated Manufacturing*, 77:102343, 2022.

Rong Zhang, Jianhao Lv, Jinsong Bao, and Yu Zheng. A digital twin-driven flexible scheduling method in a human–machine collaborative workshop based on hierarchical reinforcement learning. *Flexible Services and Manufacturing Journal*, pages 1–23, 2023.

GIOVANNI LUGARESI is Assistant Professor at the Department of Mechanical Engineering of KU Leuven. His research interests include production planning and control, internal logistics, process mining, and robust optimization. His email address is giovanni.lugaresi@kuleuven.be.

KOVACS LASZLO is the head of Industry 4.0 Technology Center at the Budapest University of Technology and Economics. He worked as IT professional and held CIO positions in the telecommunication industry. He joined BME in 2018, where he is teaching IoT, industrial digitization, and participates in several industrial digitization research projects. His email address is kovacs.laszlo@vik.bme.hu.

KORNEL TAMAS is Associate Professor at Budapest University of Technology and Economics where he has received his M.Sc. and Ph.D. degrees. His professional field is the modelling of granular materials with the use of discrete element methods. His e-mail address is tamas.kornel@gt3.bme.hu and his web-page is http://tinyurl.com/korneltamas.

PERIODIC REVIEW INVENTORY MANAGEMENT WITH BUDGET CONSTRAINTS: DISCRETE-EVENT SIMULATION AND SENSITIVITY ANALYSIS

Natalya Lysova
Federico Solari
Damiana Caccamo
Claudio Suppini
Roberto Montanari
Department of Engineering and Architecture
University of Parma
43124, Parma, Italy
federico.solari@unipr.it

KEYWORDS

Inventory management, periodic review, budget constraint, discrete event simulation, procurement lead time

ABSTRACT

In this study, a periodically reviewed inventory system with an order-up-to-level policy and budget constraints is considered. A discrete-event simulation model is developed to identify the optimal operating conditions, in terms of reorder period (DT) and order up to level (OUTL). Two different scenarios, characterized by different unit costs, are considered. A sensitivity analysis is finally conducted to evaluate the impact of procurement lead time and budget constraints on optimal conditions. A linear correlation between optimal DT and optimal OUTL is found and the equation of the line is derived as a function of the average daily demand and procurement lead time.

INTRODUCTION

Good inventory management is essential, on the one hand, to reduce the management cost and, on the other hand, to ensure high customer satisfaction. The most critical and challenging aspect of inventory management, indeed, is balancing supply and demand so that products are not overstocked or in shortage. Keeping inventory at a low level, indeed, can reduce the costs of keeping stock on hand, but at the same time increases the risk of shortages. On the contrary, maintaining inventory at a high level reduces the risk of shortages, but increases the cost of storage. “How much to order?" and "When to order?" are the two biggest problems that managers must solve when they manage an inventory.

Several approaches can be found in the literature aimed at optimizing inventory management. Some authors focus on economic optimization, by both minimizing the total management cost (Jeenanunta et al., 2021; Kouki et al., 2014; Qiu et al., 2022) or maximizing the profit (Çomez & Kiessling, 2012; Zhang, 2012; Zhang et al., 2008). Generally, the cost items considered are the stock-keeping cost, the order-issuing cost, and the stock-out cost. The latter, when backlogs are admitted, coincides with a shortage cost; in other cases, out-of-stock situations are considered to cause lost sales.

Other authors focus on the customer satisfaction point of view by introducing service level constraints to maximize the level of demand satisfaction (Chen & Li, 2015; Minner & Transchel, 2010; Wang & Chen, 2022).

In real scenarios, making these decisions is a great challenge especially because constraints caused by limited resources have to be faced. One of the most important constraints imposes budget restrictions on the amount that can be spent on stocks. This constraint is studied in different fields using different methods. Bera et al. (2009) have presented a continuous-review inventory model with a budget constraint in which the purchase cost of the system is paid when an order comes.

Also, available storage space is often a constraint to be met. Hariga (2010) has presented a continuous review inventory system with constraints of stochastic space for a single item and random demand in which the quantity

of the order and reorder point are decision variables. Finally, also lead time can be considered as a time constraint that affects system performance and should be taken into account to optimize management policy (Ben-Daya & Raouf, 1994). Depending on the case, it is treated as a predetermined parameter constant or stochastic. To the best of the authors' knowledge in the literature there are no studies that address the optimization of periodically reviewed inventory management systems with budget constraints using a simulation approach. Furthermore, no study evaluates the impact that budget constraint amount, procurement lead time as well as unit costs have on the optimal inventory management policy.

In this study, a discrete-event simulation model of an order-up-to-level periodic review policy with a budget constraint was developed. The simulation model was then used to identify the optimal combination of the operating leverages for minimizing the total management cost. Finally, a sensitivity analysis was conducted to assess the impact that both budget constraints and lead time have on management policy.

The article is organized as follows: the model overview as well as nomenclature and assumptions are reported in section 2; numerical simulation results are presented and discussed in section 3. Finally, in section 4, conclusions are reported and future research activities are outlined.

MODEL OVERVIEW AND ASSUMPTIONS

Table 1 – Nomenclature

Symbol	Description	Unit
i	i-th day (i = 1,…,n)	-
ΔT	Reorder period	days
$OUTL$	Order-up-to-level	kg
LT	Procurement lead time	days
OH_i	On-hand inventory on the day i	units
O_i	Quantity purchased on the day i	kg
$\overline{D}$	Average daily demand	kg/day
σ	Standard deviation of the daily demand	kg/day
OOS_i	Out-of-stock on the day i	kg/day
$C_{oi,i}$	Order-issuing cost on the day i	€/order
c_{so}	Unitary stock-out cost	€/day
c_{inv}	Unitary inventory holding cost	€/kg
c_p	Unitary purchase cost	€/units

The model assumes a normally distributed demand with a known mean and standard deviation. Unitary costs are assumed to be constant and deterministic. Two different scenarios characterized by different unitary costs were considered, as reported in Table 2. The impact of procurement LT was assessed by considering three LT values for each scenario (1, 2, and 3 days). In the case of stock-out situations, shortages are allowed and fully back-ordered.

Table 2 - Operating conditions of the two simulated scenarios

	Scenario 1	Scenario 2
$\overline{D}$	2000	2000
σ	100	100
c_p	2	2
c_{oi}	1500	1500
c_{inv}	0.025	0.050
c_{so}	0.25	0.50

According to the order-up-to-level periodic review policy, the stock on hand is periodically reviewed and, the orders are issued at constant time intervals (ΔT) to restore the target level of stock, i.e., order-up-to-level (OUTL). In this paper, budget constraints are included in the problem formulation, resulting in an upper limit in the quantity that could be ordered each time. Unlike traditional periodic review policy, therefore, it might happen that the maximum orderable quantity is not sufficient to restore the desired stock level. This limitation increases the risk of stock-out conditions, generating the need of reviewing the inventory management policy to adjust the levels of operating leverages ΔT and OUTL based on the system constraints. Operating leverages must therefore necessarily be determined by keeping in mind the budget constraint to maximize system performance.

The simulation model was then used to simulate a sufficient number of periods (days), 50'000 in this study, to consider different combinations of the operating leverages and identify the one that minimizes the total management cost while respecting the budget limit. For each scenario, the reordering period was varied between LT and 10 and the target OUTL between 1000 and 30000, assuming the latter to be a multiple of 500. In addition, to assess how much the budget constraint impacts the system's optimal operating conditions, for each configuration, the budget constraint was varied between 10'000 € and 1'000'000 €. For each period the following activities are performed:

(i) If an order is scheduled for the day *i* a number of items such as to restore OUTL or, if the budget limit prevents us from ordering this quantity, the maximum quantity of items that can be purchased with the available budget, is ordered:

$$O_i = \min\left(OUTL - OH_i; {}^{Budget}/_{c_p}\right) \quad (1)$$

(ii) If an order was issued LT days before, the stock is updated according to the quantity ordered:

$$OH_i = OH_{i-1} + O_{i-LT} \quad (2)$$

(iii) If OH_i meets demand at day i (d_i) is fulfilled and OH_i is consequently updated:

$$OH_i = OH_i - d_i \quad (3)$$

Otherwise, If OH_i doesn't meet d_i, out of stock on day i is computed:

$$OOS_i = d_i - OH_i \quad (4)$$

(iv) The inventory holding cost is determined based on the quantity in stock:

$$C_{inv,i} = c_{inv} \cdot OH_i \quad (5)$$

(v) Stock-out cost is a cost that occurs when the inventory level is not sufficient to satisfy customers' demand:

$$C_{so,i} = c_{so} \cdot OOS_i \quad (6)$$

(vi) The purchase cost is strictly dependent on the quantity purchased:

$$C_{p,i} = c_p \cdot O_i \quad (7)$$

(vii) The order issuing cost $C_{oi,i}$ is fixed and independent of the quantity ordered.

(viii) The total cost of the system is given by the sum of the described cost items:

$$C_{tot,i} = C_{inv,i} + C_{so,i} + C_{p,i} + C_{oi,i} \quad (8)$$

The overall structure of the model is presented in Figure 1.

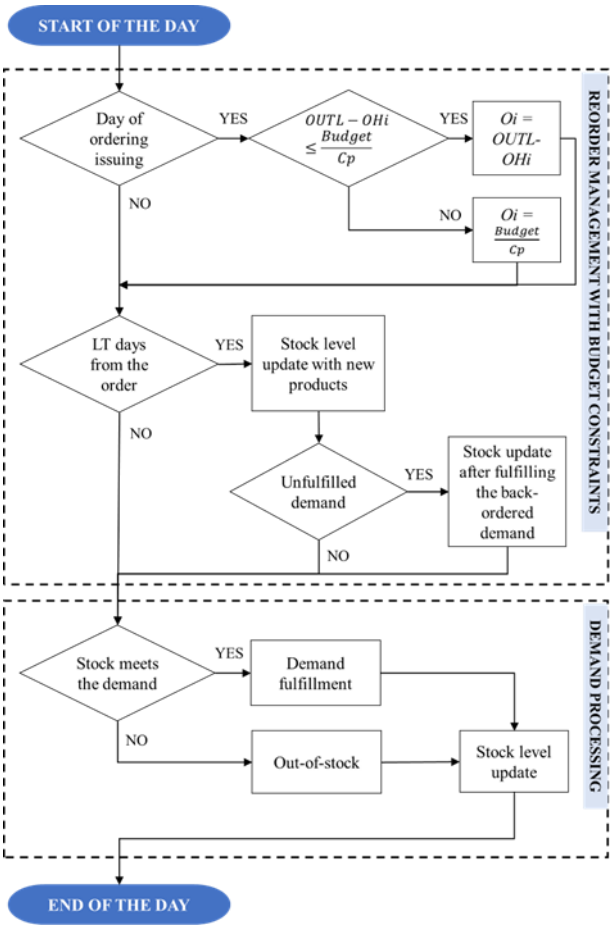


Figure 1 – Flowchart of the proposed approach

RESULTS AND DISCUSSION

The optimal operating conditions, in terms of DT and OUTL, with the associated total cost as the budget constraint changes, for scenario 1 and scenario 2, considering LT=1, are shown in Tables 3 and 4, respectively.

For both scenarios, the optimal operating condition remains unaffected by the budget constraint until the budget value is sufficient to guarantee the feasibility of the unconstrained optimum point, (i.e., 8 days as DT and 16'000 units as OUTL for scenario 1 and 6 days as DT and 12'000 units as OUTL for scenario 2). For lower budget values, the optimal DT and OUTL decrease, resulting in more frequent orders of smaller quantities, and the related total cost increases.

By observing the results obtained for each scenario, the optimal points were found to present a clear linear trend: in particular, it can be noted that as the budget constraint increases, the operating leverages increase

Table 3 - Optimal operating leverages as the budget constraint varies for Scenario 1 and LT1

Budget [€]	DT [days]	OUTL [units]	C_{tot} [€]
1'000'000.00	8	16000	4365.24
100'000.00	8	16000	4365.24
80'000.00	8	16000	4365.24
70'000.00	8	16000	4365.24
65'000.00	8	16000	4365.24
60'000.00	8	16000	4365.24
55'000.00	8	16000	4365.24
50'000.00	8	16000	4365.24
45'000.00	8	16000	4365.24
40'000.00	8	16000	4365.24
35'000.00	8	16000	4365.24
30'000.00	7	14000	4366.79
25'000.00	6	12000	4378.19
20'000.00	4	8000	4454.30
15'000.00	3	6000	4555.01
10'000.00	2	4000	4781.59

and the value of the minimum cost decreases (Figure 2). Moreover, when LT increases, since when the order is issued, the quantity in stock is higher as it must meet the demand of *(LT-1)* days (same-day demand has already been met), the optimal value of OUTL is higher.

On the other hand, the optimal DT does not result to be affected by lead time. The theoretical on-hand inventory (dashed line in Figure 3), which is updated at the moment the order is issued, reflects the optimal OUTL value, as it accounts for the quantity ordered and for the demand expected during the procurement LT period. On the other hand, the real stock-on-hand pattern (solid line in Figure 3) is only shifted forward by LT days. The average quantity ordered, indeed, appears to remain almost the same independently from LT.

Looking at the results of scenario 2, it appears that both the optimal DT and the optimal OUTL decrease.

Table 4 - Optimal operating leverages as the budget constraint varies for Scenario 2 and LT1

Budget [€]	DT [days]	OUTL [units]	C_{tot} [€]
1'000'000.00	6	12000	4507.65
100'000.00	6	12000	4507.65
80'000.00	6	12000	4507.65
70'000.00	6	12000	4507.65
65'000.00	6	12000	4507.65
60'000.00	6	12000	4507.65
55'000.00	6	12000	4507.65
50'000.00	6	12000	4507.65
45'000.00	6	12000	4507.65
40'000.00	6	12000	4507.65
35'000.00	6	12000	4507.65
30'000.00	6	12000	4507.65
25'000.00	6	12000	4507.65
20'000.00	4	8000	4534.72
15'000.00	3	6000	4611.32
10'000.00	2	4000	4814.29

Indeed, since both c_{inv} and c_{so} are higher, it is more convenient to order, more frequently, a lower average amount of items. This reduces both the stock-out probability and the average stock level. The correlation between the optimal DT and the optimal OUTL, as the budget constraint changes, is confirmed to be linear.

This correlation was fitted with a linear model and the equation of the line was then derived (equation 9).

$$OUTL_{opt} = \overline{D} \cdot DT + \overline{D} \cdot (LT - 1) \qquad (5)$$

It can be observed that the procurement LT affects the intercept, while it does not affect the slope of the line, which is only affected by the average daily demand (Figure 4 and Figure 5).

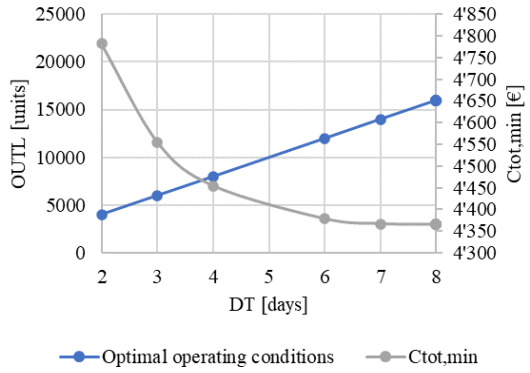


Figure 2 - Trend in optimal operating conditions and total cost values for different budget constraints for Scenario 1 and LT1

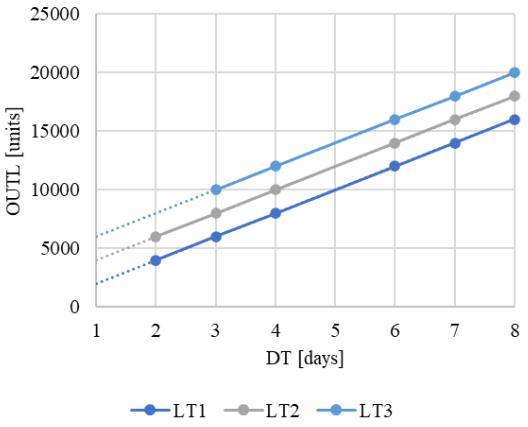


Figure 3 – Scenario 1; trend of the optimal combinations of the operating leverages DT and OUTL

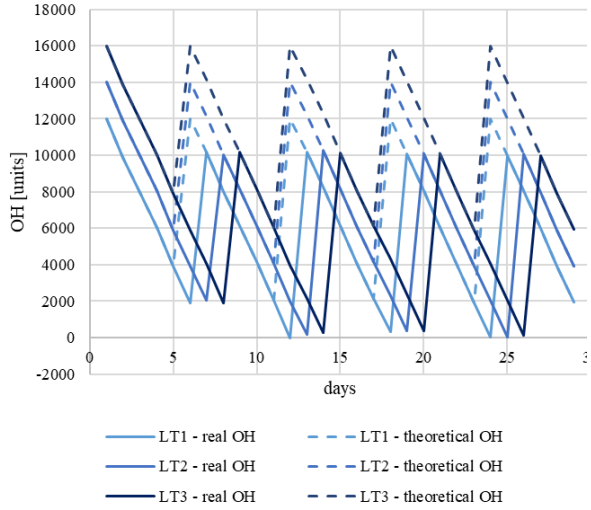


Figure 2 – OH levels in the case of Scenario 2 and three different values of procurement LT values (1, 2, and 3)

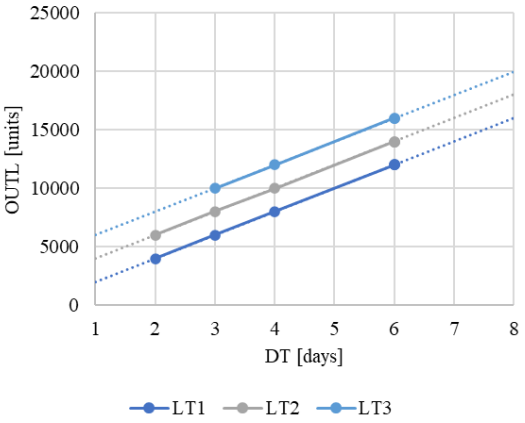


Figure 4 – Scenario 2; trend of the optimal combinations of the operating leverages DT and OUTL

CONCLUSIONS

A discrete event simulation model of an order-up-to-level periodic review policy with a budget constraint was developed.

The simulation model was used to identify the optimal operating condition, in terms of the combination of DT and OUTL, to minimize the total management cost system while meeting the budget constraint.

A sensitivity analysis was then performed to assess how the optimal condition varies as LT, budget constraint, and unit costs (c_{inv} and c_{so}) vary.

Results highlighted that, as the lead time changes, with all other parameters remaining constant, the value of optimal OUTL changes proportionally to the lead time value, while the value of optimal DT does not change.

A linear trend describing the correlation between DT and OUTL was derived, which appears to be strongly influenced by average demand.

In future studies, it will be interesting to evaluate the impact of different costs and system parameters on the correlation identified and described in this study. A simulative campaign based on DOE will also be interesting to assess the significance of each parameter on the output response of the model (i.e. total management cost).

REFERENCES

Ben-Daya, M., & Raouf, A. (1994). Inventory Models Involving Lead Time as a Decision Variable. *The Journal of the Operational Research Society*, *45*(5), 579. https://doi.org/10.2307/2584393

Bera, U. K., Rong, M., Mahapatra, N. K., & Maiti, M. (2009). A multi-item mixture inventory model involving random lead time and demand with budget constraint and surprise function. *Applied Mathematical Modelling*, *33*(12), 4337–4344. https://doi.org/10.1016/j.apm.2009.03.025

Chen, H., & Li, P. (2015). Optimization of (R, Q) policies for serial inventory systems using the guaranteed service approach. *Computers & Industrial Engineering*, *80*, 261–273. https://doi.org/10.1016/j.cie.2014.12.003

Çomez, N., & Kiessling, T. (2012). Joint inventory and constant price decisions for a continuous review system. *International Journal of Physical Distribution & Logistics Management*, *42*(2), 174–202. https://doi.org/10.1108/09600031211219672

Hariga, M. A. (2010). A single-item continuous review inventory problem with space restriction. *International Journal of Production Economics*, *128*(1), 153–158. https://doi.org/10.1016/j.ijpe.2010.06.008

Jeenanunta, C., Kongtarat, V., & Buddhakulsomsiri, J. (2021). A simulation-optimisation approach to determine optimal order-up-to level for inventory system with long lead time. *International Journal of Logistics Systems and Management*, *38*(2), 253. https://doi.org/10.1504/IJLSM.2021.113250

Kouki, C., Jemai, Z., Sahin, E., & Dallery, Y. (2014). Analysis of a periodic review inventory control system with perishables having random lifetime. *International Journal of Production Research*, *52*(1), 283–298. https://doi.org/10.1080/00207543.2013.839895

Minner, S., & Transchel, S. (2010). Periodic review inventory-control for perishable products under service-level constraints. *OR Spectrum*, *32*(4), 979–996. https://doi.org/10.1007/s00291-010-0196-1

Qiu, R., Sun, Y., & Sun, M. (2022). A robust optimization approach for multi-product inventory management in a dual-channel warehouse under demand uncertainties. *Omega*, *109*, 102591. https://doi.org/10.1016/j.omega.2021.102591

Wang, L., & Chen, H. (2022). Optimization of a stochastic joint replenishment inventory system with service level constraints. *Computers & Operations Research*, *148*, 106001. https://doi.org/10.1016/j.cor.2022.106001

Zhang, J.-L. (2012). Integrated Decision On Pricing, Promotion And Inventory Management. *Asia-Pacific Journal of Operational Research*, *29*(06), 1250038. https://doi.org/10.1142/S0217595912500388

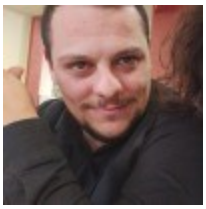
Zhang, J.-L., Chen, J., & Lee, C.-Y. (2008). Joint optimization on pricing, promotion and inventory control with stochastic demand. *International Journal of Production Economics*, *116*(2), 190–198. https://doi.org/10.1016/j.ijpe.2008.09.008

AUTHOR BIOGRAPHIES

NATALYA LYSOVA is a Ph.D. Candidate at the University of Parma, where she has achieved a Masters' Degree in Engineering for the Food Industry in 2021. Her research project is titled "Virtualization approaches for industrial plants control and design". Her research topics include CFD and discrete-event simulation of industrial plants and inventory systems. Her e-mail address is: natalya.lysova@unipr.it

FEDERICO SOLARI is a researcher and lecturer at the Department of Engineering and Architecture of the University of Parma. He authored 30 publications indexed on Scopus. His main research topics are: industrial plant logistics, industrial plant analysis and design, supply chain management, industrial plant analysis and design, advanced industrial plant design also using simulative techniques. His e-mail address is: federico.solari@unipr.it and his Web- page can be found at ttps://personale.unipr.it/it/ugovdocenti/person/102724.

DAMIANA CACCAMO was born in Siracusa, Italy and studied management engineering at the University of Parma, where she took her master's degree in 2022. During her thesis work, she focused on the study of periodic review systems with budget constraints. She is now doing an extra-curricular internship in a multinational Italian company in Parma in the logistics area. Her e-mail address is: damiana.caccamo@studenti.unipr.it



CLAUDIO SUPPINI was born in Vergato, Italy and went to the University of Parma, where he studied management engineering and obtained his master's degree in 2022. He starts as Ph.D. student in industrial engineering at the same University, where he is conducting research into the field of discrete-event simulation for the optimization of industrial plants. His e-mail address is: claudio.suppini@unipr.it

ROBERTO MONTANARI is Full Professor SSD Industrial Mechanical Systems Ing-Ind/17 at the Department of Engineering and Architecture - University of Parma is the Holder of the course of Industrial Plants - Bachelor of Engineering Management, the course of Simulation of Production Systems - Bachelor of Engineering in Food Industry Plants and Machinery. His e-mail address is: roberto.montanari@unipr.it and his Web- page can be found at https://personale.unipr.it/it/ugovdocenti/person/19786.

COMPUTATIONAL FLUID DYNAMICS SIMULATION OF SLOSHING INSIDE BEVERAGE CANS ON A ROTARY FILLING MACHINE

Federico Solari
Natalya Lysova
Federico Scano
Roberto Montanari

University of Parma
Department of Engineering and Architecture
Parco Area delle Scienze 181/A, 43124, Parma, Italy

federico.solari@unipr.it

Enrico Bedogni

Krones AG

Böhmerwaldstraße 5, 93073 Neutraubling, Germany

Gabriele Copelli

Gabriele Copelli Engineering Office

Via Benedetta 75, 43122, Parma, Italy

KEYWORDS

Sloshing, rotary filler, CFD simulation, volume of fluid, single reference frame, openFoam, Fluent

ABSTRACT

Sloshing is a critical issue in many industrial contexts. In the food industry, it becomes particularly relevant during the filling of beverage cans and bottles with automatic rotary machines, when the uncapped filled containers move to the transfer star wheel, suddenly changing direction of motion and potentially causing the spilling of the product. Deep knowledge of the system behavior and the fluid dynamics in the domain is essential to guarantee the safety and quality of the final products and processing environment. In this study, Computational Fluid Dynamics (CFD) was used to simulate sloshing in beverage cans using two CFD software: commercial ANSYS Fluent and open-source OpenFOAM. Some modeling strategies are explored with the aim of making the simulation more efficient without impacting the results, and an approach for tracking the maximum fluid level in the can is proposed. The modeling methodology was validated by means of an analytical model and by comparing the results calculated by the two software. Finally, operational insights were derived based on the results of a sensitivity analysis carried out by varying the star wheel diameter and the system productivity.

INTRODUCTION

Sloshing, the transient movement of liquids within a confined container, poses significant challenges in various industries. In transportation and maritime contexts, it can lead to loss of control over the vehicle or compromise the stability of the ship due to the movement of large volumes of fluid.

The phenomenon of sloshing, however, also affects the processing activities of the food industry, mainly in the stages following the filling of containers, when they are transferred towards the sealing station. In this case, the undesirable effect is related to the spilling of fluids, which happens when the product overflows from the container due to the major stresses and acceleration experienced during the motion. Spilling leads to loss of product, contamination of the processing environment, and possible deposition of fluids on the outer walls of the container that could compromise the integrity of the product and the sealing process, undermining both the safety and the quality of the final product.

This phenomenon is particularly relevant during the filling of beverage cans, due to the very limited headspace and the high rotation speeds of the transfer star wheel that increase the probability of sloshing. In this case, the occurrence of sloshing and spilling mainly impacts the seaming of the cans.

A deep understanding of the dynamics of the fluid inside the cans as they travel between the filling and sealing stations, as well as a detailed knowledge about the behavior of the free surface, are crucial for correctly defining the productivity, thus the rotational velocity of the machines, reducing testing times, wastes, and quality issues. In this context, numerical simulation can be of significant support, providing the decision-makers with useful insights and data for detailed analyses, while considerably reducing experimental testing.

Over time, analytical models of sloshing have been developed and validated (Ibrahim 2005) regarding mostly simple geometries and standard boundary conditions. To increase the modeling precision and correspondence to the actual conditions, simulation can be a crucial tool (Elahi et al. 2015). For example, simulation can be used to predict the severity of sloshing in relevant applications (Zheng et al. 2020). As always when it comes to simulation results, validation is required, and it can be usually based on the comparison with theoretical or experimental results (Elahi et al. 2015; Guo et al. 2010). To capture in detail the fluid dynamics involved with the sloshing phenomenon, Computational Fluid Dynamics (CFD) plays a key role, as it allows for modeling complex fluid-structure interactions under various conditions, considering realistic container geometries, custom transfer trajectories, as well as complex and even multiphase fluids usually processed in the food industry. CFD has been applied to simulate a huge variety of food industry processes in the last years

(Szpicer et al. 2023; Ian Wilson and John Chew 2023), also thanks to the increase in the computation resources and the presence of numerous simulation software. To perform CFD simulation, indeed, several commercial and open-source software are available. While commercial CFD software offer benefits such as technical support, updates and enhanced usability, open-source tools provide flexibility and free access, albeit requiring more expertise.

Despite the relevance of sloshing in the beverage industry, to the best of the authors' knowledge, no studies in the literature have dealt with this particular issue to date. Indeed, most articles in the literature have focused on the sloshing of fluids in tanks (Liu and Lin 2008) during transportation, sloshing occurrences and effects in the aerospace (Saltari et al. 2021; Tang et al. 2018) and naval applications, and the different methods that could be implemented to mitigate this phenomenon, e.g., introduction of different geometries and characteristics of baffles (Yu et al. 2020; Zhang 2019). In the context of sloshing in the field of automatic machines, Guagliumi et al. (2021) have presented a technique for analyzing sloshing in cylindrical containers performing rectilinear movements. Guagliumi et al. (2022) have proposed a discrete linear model of sloshing, implementable in real-time for the feedforward anti-sloshing control of container motion.

This article presents a simulation approach, implemented with both commercial ANSYS Fluent and open-source OpenFOAM CFD software, to predict sloshing during the transfer of beverage containers after filling. The calculated inclination angle of the free surface is validated with theoretical data. In addition, using two different tools allows to inter-validate the simulation models by comparing the respective results. Some modeling strategies are explored, aiming at making the simulations more rapid and efficient. Then, the results of the simulations are discussed, proposing an approach for tracking the level of the fluid during the transfer of the container and detecting the occurrence of sloshing. Finally, a sensitivity analysis is performed to assess the effect of different transfer star wheel diameters and productivity levels on the maximum level reached by the fluid.

Table 1: Characteristics of the two carousels

	Diameter [mm]	Number of heads	Rotational speed [rev min^{-1}]
Filling carousel	1080	24	13.89
Transfer star wheel	720	16	20.83

METHODS

Analyzed scenario

The fluid behavior inside a can during the transition from the filling carousel to the transfer star wheel, as depicted in Figure 1, was analyzed. For this purpose, the period from the end of the filling process to the exit from the transfer star wheel was simulated. In the simulated machine, the filling is completed about 40 degrees before the carousel change; it was assumed that, at that time, the free surface conformation within the can was already stabilized.

Since the filling carousel has a larger diameter than the transfer star wheel, and since the tangential velocity must be the same, the latter has a higher rotational speed.

A real industrial context was considered, characterized by a productivity of 20'000 cans per hour (c/h). 0.5-litre can geometry, having an inner diameter of 0.033 m and height of 0.15 m, was modeled.

Table 1 reports the geometric characteristics and the operating conditions assumed for the two carousels. Based on the data, the law of motion of the can, in terms of the evolution of angular velocity over time, and as a consequence, of centrifugal acceleration over time, can be derived.

In Figure 2, the angular velocity and the centrifugal acceleration characterizing the motion of the can are presented: at 0.5 s the can is assumed to transfer from the filler to the star wheel.

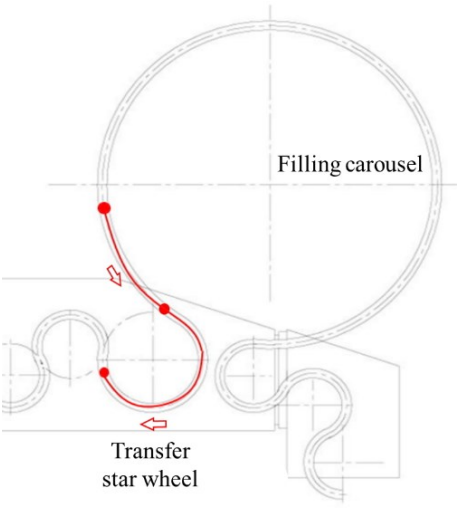


Figure 1: Scheme of the simulated line

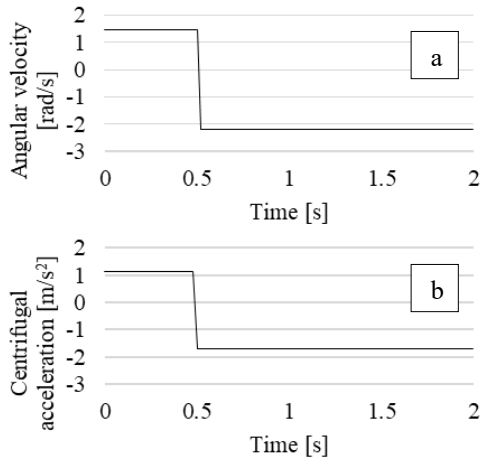


Figure 2: a) angular velocity and b) centrifugal acceleration characterizing the motion of the can in the simulated system

Table 2: Characteristics of the fluids

	Air	Water
Density [kg m^{-3}]	1.225	998.2
Dynamic viscosity [Pa s]	1.8e-5	1.0e-3
Surface tension [N m^{-1}]	0.072	

Numerical simulation

In this study, we explore the capabilities of CFD simulation for studying sloshing phenomena in industrial bottling processes.
Aiming to assess the reliability of the proposed approach, the results obtained with two different software, namely ANSYS Fluent and OpenFOAM, were compared.

The single reference frame (SRF) method was adopted to reproduce the behavior of the fluid inside the can as it moved within an absolute reference frame with a given law of motion. This method consists of an alternative approach, less computationally demanding compared to the traditional moving mesh method, that can be adopted when simulations involve moving regions, allowing to study them in their respective reference frames.

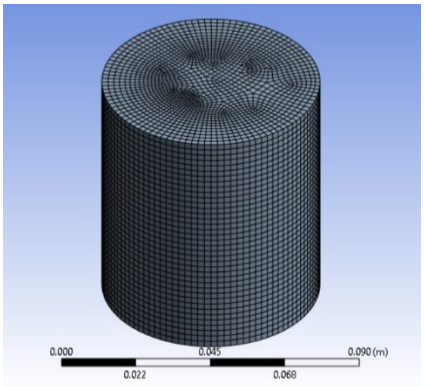


Figure 3: Hexahedral mesh of the computational domain

A multiphase simulation was set up using the Volume-of-Fluid (VOF) method, which is appropriate for the simulation of immiscible fluids being separated by a well-defined interface surface (free surface) (Hirt and Nichols, 1981). Air and water were considered as fluids, whose properties, although well known, are summarized in Table 2.
Under the conditions considered, a significant but not abrupt free surface perturbation was expected, not affecting the entire volume of fluid contained in the can, but only the upper part. For this reason, it was decided to include only the upper 3 cm of the can in the calculation, as the lower part of the cylinder contributed minimally to the sloshing dynamics. This assumption was validated by comparing the results obtained by simulating the whole can geometry with those obtained with the reduced domain.
To ensure a high-resolution capture of the fluid behavior, a hexahedral mesh, with equilateral elements having a dimension of 3 mm, was defined (Figure 3). To better compare the results obtained, the same mesh was used with both solvers.

The law of motion was finally assigned to the mesh by imposing the acceleration field represented in Figure 2. Moreover, a gravitational field, with a downward acceleration of 9.81 m/s^2, was applied.

Inclination of the free surface

To identify the initial condition, i.e., the shape of the liquid-free surface inside the can at the end of the filling process, especially in terms of inclination angle (θ) with respect to the horizontal plane, an ad-hoc simulation was conducted. Starting from a resting condition (θ=0°), the revolution of the can on the filling carousel was simulated until the stationarity condition was reached. The inclination of the free surface was then measured and compared with the angle calculated by solving the balance of forces acting on the fluid, as described also in Elahi et al. (2015):

$$\theta = arc\ tan\left(\frac{a_r}{g}\right) \quad (1)$$

Where a_r is the centrifugal acceleration in the radial direction and g is the gravitational acceleration, acting in the y direction.
This comprehensive procedure served as a robust foundation for determining the initial free surface angle in the first carousel and ensured accurate initialization for subsequent simulations. Furthermore, it allowed to validate the simulation approach with well-established theoretical notions.
After the initial inclination angle was obtained and validated as described, the simulation of the transition phase from the filling carousel to the transfer star wheel was performed. The simulation was run using both software packages, setting the angle determined as the initial condition. The simulation outcomes were then compared, to check whether the results obtained with

both software, in terms of liquid sloshing inside the can, starting from an identical initial condition, were the same. This further step is intended as an inter-validation step of obtained results.

Maximum fluid level Ymax

To qualitatively assess sloshing, it is sufficient to observe the behavior of the free surface during the movement of the can. To make a more accurate and precise assessment, the maximum height reached by the free surface during the can motion, *Ymax*, was also calculated and traced over time. This parameter can be very useful in assessing the impact sloshing may have on the process quality and efficiency. When *Ymax* is lower than the total height of the can, even if sloshing is happening, the liquid doesn't leak out. In this scenario, therefore, sloshing does not represent an issue. On the other hand, when *Ymax* exceeds the maximum height of the can, the liquid may overflow from the can and negatively affect the filling process.

The comparison of the results obtained with the two software was then performed in two ways:

- a comparison of the trend over time of the maximum height reached by the free surface.
- a frame-by-frame comparison between the shape of the free surface computed by the two software in specific time intervals.

Finally, the proposed method was used to evaluate the effect that some operational and design parameters have on liquid behavior. In particular, the impact of transfer star wheel diameter and line productivity was evaluated.

RESULTS

The first phase of the study aimed to identify the computational domain. To have a trade-off between computational time and results accuracy, it was decided to compare the results achieved considering only the upper part of the can (30 mm), with the results obtained considering the whole can. The results are shown in Figure 5a.

Based on the graph, it can be observed that as the can moves from the filling carousel to the transfer star wheel (t=0.5 s) sloshing occurs. A wave is generated that reaches its maximum peak 0.14 s after the transfer and then continues fluctuating throughout the period considered. The results of the two simulations are perfectly overlapping: the percentage deviation between the two curves is always less than 3.9%. It can therefore be concluded that the liquid in the bottom of the can does not significantly affect the dynamics of sloshing and can consequently be excluded from the computational domain.

The second phase of the study focused on identifying the liquid condition at the exit of the filling carousel. To validate this condition, the angle of inclination of the free surface under the effect of the centrifugal force generated by the rotation of the carousel was calculated in two different ways. The balance of forces acting on the free surface (eq. 1) leads to an angle value of 6.64 degrees.

Below are the results of a simulation that, starting from a resting condition (θ=0°), simulates the revolution of the can on the filling carousel until the inclination angle of the free surface is stabilized (Figure 4).

Measuring the inclination angle that the free surface reaches under stationary conditions yields a theta value of 6.24 degrees which results in a deviation of 6% from the theoretical value, calculated with eq.1. This deviation can be considered acceptable because it is included within the range of accuracy that the mesh size allows for.

The following phase of the study focused on the sloshing occurring as the beverage can moved from the filling carousel to the transfer star wheel. A comparison between the results obtained with ANSYS Fluent and those obtained with OpenFOAM, is shown in Figure 5b for the value of *Ymax*, while a representation of the free surface, calculated with the two software, at several time steps is presented in Figure 6.

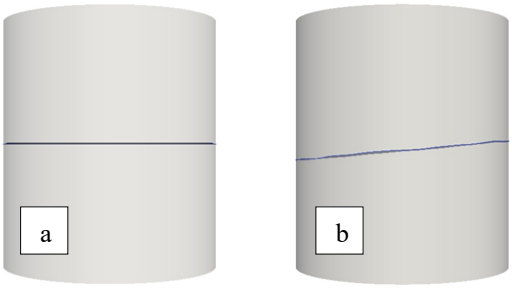


Figure 4: Initial (a) and final (b) state of the free surface

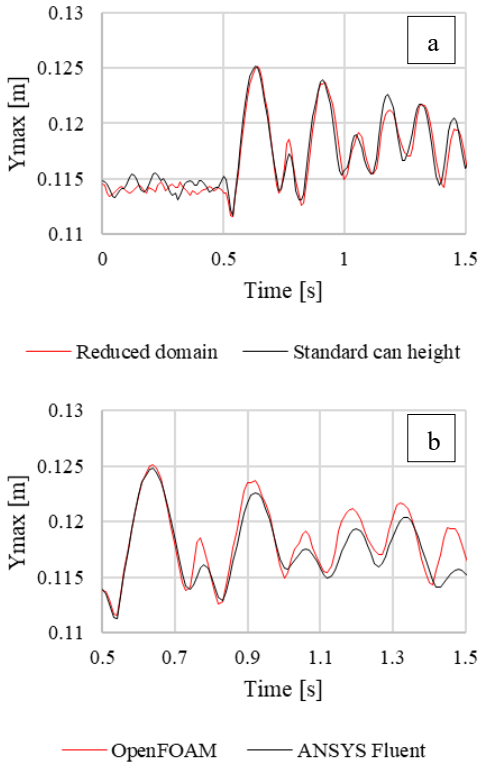


Figure 5: a) comparison of standard vs reduced domain; b) comparison of results calculated with OpenFOAM vs ANSYS Fluent

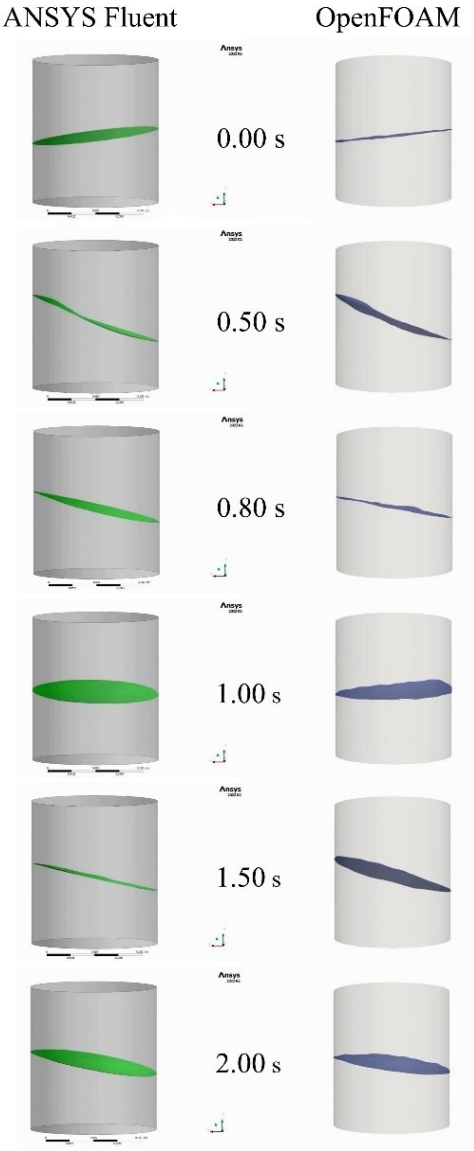


Figure 6: Free surface of the fluid inside the can at different time steps, calculated with ANSYS Fluent (left) and OpenFOAM (right)

It can be concluded that the results of the two CFD software agree both in the prediction of the maximum fluid level reached inside the can and in the oscillation frequency of the wave itself. Even the shapes of the free surface calculated by the two software appear to be in good agreement.

Finally, the proposed simulation approach was used to evaluate the impact that some processing and geometric parameters have on liquid behavior. The evolution of *Ymax* reached by the liquid inside the can for different diameters of the transfer star wheel is presented in Figure 7a, while the levels observed with different productivity values are shown in Figure 6
7b.

It can be seen that the maximum height reached by the liquid decreases as the diameter of the transfer star increases.

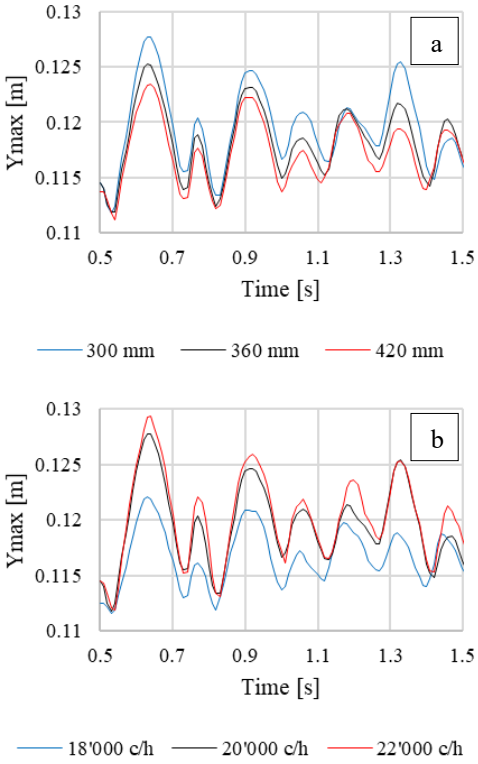


Figure 7: *Ymax* with a) different star wheel diameters and b) different productivity levels

This correlation appears to be linear. On the other hand, regarding the correlation between *Ymax* and productivity, it can be seen that between 18’000 and 20’000 cans per hour, there is a much greater difference than between 20’000 and 22’000 cans per hour. This indicates that this correlation is probably not linear in nature. This aspect needs to be investigated in more detail in future research studies.

CONCLUSIONS

In the present study, a method for predicting liquid sloshing occurring during container motion in transfer carousels during the filling process is presented.
In-depth knowledge of this phenomenon is a big advantage in both the design and management phases of automatic filling. A rigorous step-by-step approach was followed to identify the optimal simulation settings that would allow to obtain accurate results in times consistent with industrial needs. The reliability of the simulation approach was first validated by comparing the results with those obtained from an analytical model in a simple case study (i.e., can in uniform rotary motion around a fixed axis).
A subsequent inter-validation phase, conducted by comparing the results obtained with two different software (i.e., ANSYS Fluent and OpenFOAM) in a more complex case study, reproducing a real-world

context in which the can passes from the filling carousel to the transfer star, confirmed the consistency of the results. The findings suggest that both tools can significantly support the optimization phase of the industrial processes by accurately modeling complex fluid dynamics, thereby reducing testing time and enhancing production efficiency.

The developed approach represents a novelty within the scientific literature because to date there is no study outlining fluid dynamic simulation models to simulate sloshing during filling processes. The presented approach allows to evaluate the behavior of the liquid inside the can between the end of the filling process and the seaming by allowing to assess the impact that some design and operational parameters have on the behavior of the liquid. In future activities, experimental tests will be performed to further validate the proposed approach.

REFERENCES

Elahi, R., Passandideh-Fard, M., & Javanshir, A. (2015). Simulation of liquid sloshing in 2D containers using the volume of fluid method. Ocean Engineering, 96, 226–244. https://doi.org/10.1016/j.oceaneng.2014.12.022

Guo, L. C., Zhang, S., Morita, K., & Fukuda, K. (2010). Fundamental validation of the finite volume particle method for 3D sloshing dynamics. International Journal for Numerical Methods in Fluids, 68(1), 1–17. Portico. https://doi.org/10.1002/fld.2490

Hirt, C. W., & Nichols, B. D. (1981). Volume of fluid (VOF) method for the dynamics of free boundaries. Journal of Computational Physics, 39(1), 201–225. https://doi.org/10.1016/0021-9991(81)90145-5

Guagliumi, L., Berti, A., Monti, E., & Carricato, M. (2022). Antisloshing Trajectories for High-Acceleration Motions in Automatic Machines. Journal of Dynamic Systems, Measurement, and Control, 144(7). https://doi.org/10.1115/1.4054224

Guagliumi, L., Berti, A., Monti, E., & Carricato, M. (2021). A Simple Model-Based Method for Sloshing Estimation in Liquid Transfer in Automatic Machines. IEEE Access, 9, 129347–129357. https://doi.org/10.1109/access.2021.3113956

Ian Wilson, D., & John Chew, Y. M. (2023). Fluid mechanics in food engineering. Current Opinion in Food Science, 51, 101038. https://doi.org/10.1016/j.cofs.2023.101038

Ibrahim, R. A. (2005). Liquid Sloshing Dynamics. https://doi.org/10.1017/cbo9780511536656

Liu, D., & Lin, P. (2008). A numerical study of three-dimensional liquid sloshing in tanks. Journal of Computational Physics, 227(8), 3921–3939. https://doi.org/10.1016/j.jcp.2007.12.006

Saltari, F., Traini, A., Gambioli, F., & Mastroddi, F. (2021). A linearized reduced-order model approach for sloshing to be used for aerospace design. Aerospace Science and Technology, 108, 106369. https://doi.org/10.1016/j.ast.2020.106369

Szpicer, A., Bińkowska, W., Wojtasik-Kalinowska, I., Salih, S. M., & Półtorak, A. (2023). Application of computational fluid dynamics simulations in food industry. European Food Research and Technology, 249(6), 1411–1430. https://doi.org/10.1007/s00217-023-04231-y

Tang, Y., Yue, B., & Yan, Y. (2018). Improved method for implementing contact angle condition in simulation of liquid sloshing under microgravity. International Journal for Numerical Methods in Fluids, 89(4–5), 123–142. Portico. https://doi.org/10.1002/fld.4685

Yu, L., Xue, M.-A., & Zhu, A. (2020). Numerical Investigation of Sloshing in Rectangular Tank with Permeable Baffle. Journal of Marine Science and Engineering, 8(9), 671. https://doi.org/10.3390/jmse8090671

Zhang, E. (2019). Numerical research on sloshing of free oil liquid surface based on different baffle shapes in rectangular fuel tank. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 234(2–3), 363–377. https://doi.org/10.1177/0954407019855569

Zheng, M.-Z., Gou, Y., Teng, B., & Jo, H. (2020). A practical prescreening method for sloshing severity evaluation. Petroleum Science, 17(4), 1119–1134. https://doi.org/10.1007/s12182-020-00453-x

AUTHOR BIOGRAPHIES

FEDERICO SOLARI is a researcher and lecturer at the Department of Engineering and Architecture of the University of Parma since September 2021. He graduated (with distinction) in Mechanical Engineering for the Food Industry in 2008 and got his Ph.D. in Industrial Engineering in 2014, both at the University of Parma. He authored 39 publications indexed on Scopus. His main research topics are industrial plant logistics, industrial plant analysis and design, supply chain management, advanced industrial plant design also using simulative techniques. In his research activities, he explores the use of simulation and advanced numerical techniques for the design, control and maintenance of industrial plants and processes. His e-mail address is: federico.solari@unipr.it and his Web- page can be found at ttps://personale.unipr.it/it/ugovdocenti/person/102724.

ORCID: 0000-0001-8365-965X

NATALYA LYSOVA is a Ph.D. Candidate at the University of Parma, where she graduated in Engineering for the Food Industry in 2021. Her doctoral research project is titled “Virtualization approaches for industrial plants control and design”. In her research activities, she aims to leverage

numerical techniques for the analysis, design and optimization of industrial systems, plants, and devices. Her e-mail address is: natalya.lysova@unipr.it.
ORCID: 0000-0001-6066-6398

FEDERICO SCANO is a Master's student at the University of Parma. He is now concluding his studies in Engineering for Food Industry and this work will be part of his dissertation.

ROBERTO MONTANARI is a Full Professor at the Department of Engineering and Architecture - University of Parma. He is the Holder of the course in Industrial Plants – Bachelor of Engineering Management, and the course in Simulation of Production Systems – Master Degree in Engineering for the Food Industry. His e-mail address is: roberto.montanari@unipr.it and his Web- page can be found at https://personale.unipr.it/it/ugovdocenti/person/19786.

ENRICO BEDOGNI is an R&D Mechanical Engineer. He got his Ph.D. in Industrial Engineering in 2013 at the University of Parma. After a few years as a researcher at the University of Parma, he spent several years in the R&D field, in Automatic Machines and Automotive sectors. He is now leading the Krones Innovation Hub. His e-mail address is: enrico.bedogni@krones.com

GABRIELE COPELLI is a Mechanical Engineer; from 2005 to 2013 research fellow at the Department of Industrial Engineering of the University of Parma, in the machinery area. Freelance since 2013. He is expert in mechanical design and development of machinery and processes for the food industry by means of CAE (Computer Aided Engineering) and Computational Fluid Dynamics (CFD). www.gabrielecopelli.com

Model Generalisation for Predicting the Amount of Photosynthetically Available Radiation in the Water Column from Freefall Profiler Observations

Christoph Tholen[1], Lars Nolle[1,2], Jochen Wollschläger[3] and Frederic Stahl[1]

[1]German Research Center for Artificial Intelligence
Marie-Curie-Straße 1
26129 Oldenburg, Germany
Email: {christoph.tholen|lars.nolle|frederic_theodor.stahl}@dfki.de

[2]Jade University of Applied Sciences
Friedrich-Paffrath-Straße 101
26389 Wilhelmshaven, Germany
Email: lars.nolle@jade-hs.de

[3] Carl von Ossietzky Universität Oldenburg
School of Mathematics and Science
Institute for Chemistry and Biology of the Marine Environment (ICBM)
Ammerländer Heerstraße 114-118
26129 Oldenburg
Email: jochen.wollschlaeger@uni-oldenburg.de

KEYWORDS
Machine Learning, Underwater Light Field, Photosynthetic Active Radiation, Freefall Profiler, KNIME

ABSTRACT

In modern oceanography Photosynthetically Available Radiation (PAR) is used for modelling vegetation growth as it is a requirement for the process of photosynthesis. PAR as integrated value of the light spectrum between 400-700 nm can be measured directly using respective sensor systems. However, PAR can also be determined indirectly using measurements from only a small number of discrete wavelengths. In this paper, such a modelling approach is presented for predicting PAR in the water column. The approach uses spectral information within the water column and from above the sea surface. Three different modelling approaches based on artificial intelligence (AI) were used. It was shown that the artificial neural network (ANN) approach outperformed the regression tree (RT) and the linear regression (LR) approaches. It was also shown that the models generalise well, with an accuracy loss of 10 % based on the median, on data recorded in other geolocations without additional modification or re-training.

INTRODUCTION

In modern oceanography, one of the important parameters is Photosynthetically Available Radiation (PAR), which is the integrated radiation between 400-700 nm. It can be used for modelling vegetation growth due to being a requirement for the photosynthesis process (Holinde and Zielinski, 2016; Wang et al., 2013).

Therefore, measuring PAR is important. As proven in previous work, the PAR values can be re-constructed using only discrete wavelengths from the underwater light field and, if necessary, additional environmental parameters (Stahl et al., 2022; Kumm et al., 2022). Predicting PAR has been explored in the context of autonomous Argo Float devices (Sloyan et al., 2018) in (Stahl et al., 2022) using multiple linear regression and regression trees. Kumm et. al. (2022) showed that these results can be improved by using artificial neural networks-based models and further improved by incorporating additional environmental parameters, i.e. pressure. Due to the heavy dependency of the underwater light field on the incoming surface irradiance (E_s) (Wollschläger et al., 2020d), an alternative to incorporate pressure measurements to improve accuracy would be using these surface light field measurements. However, since Argo floats operate autonomous underwater for long time, simultaneous measurements of the surface light field is not an option.

A similar way of measuring PAR is being conducted by Freefall Profilers (Figure 1). However, different to Argo Floats, these measurements also comprise E_s. Therefore, this study tries to map the approaches from Kumm et al (2022) and Stahl et al (2022) to the freefall profiler platform. In addition, it will be investigated if incorporating E_s into the model building increases the accuracy. It will also be investigated if models trained on one set of experiments can be generalised to data from other measurements, i.e. other geolocations. If possible, it would allow marine scientists to reuse the developed models without re-training.

Figure 1 – Freefall profiler.

RADIOMETRIC PROFILING

For the data acquisition, the underwater light field was investigated using a free-falling profiling system (HyperPro II; Sea-Bird Scientific, USA, former Satlantic), which is designed to slowly sink vertically through the water column (Figure 1). The HyperPro II was equipped with two hyperspectral HyperOCR radiometers (Sea-Bird Scientific, USA, λ=350-800 nm) measuring different parts of the underwater light field: A planar cosine radiometer was mounted looking upward in order to determine the downwelling irradiance $E_d(\lambda)$, thus the overall light field propagating from the from the sea surface into the depth. Another, radiance-type radiometer with a field-of-view of 8.5° was mounted looking downward to measure the upwelling radiance $L_u(\lambda)$, thus the light field scattered back from the depth in a narrow cone in sinking direction. A third planar cosine radiometer was placed as reference in an unshaded, upright position on an elevated position on the ship in order to determine the downwelling irradiance $E_s(\lambda)$ above the seawater, thus the light field impinging on the sea surface. Its measurements allow the correction of the in-water measurements for general changes in the light field (e.g. temporary cloud coverage) during the deployment of the HyperPro II. The HyperPro II also contains sensors for additional parameters, like temperature, conductivity, depth, chlorophyll-a fluorescence, backscatter, and tilt of the instrument. All sensors on the instrument were pre-calibrated by the manufacturer, and the radiometers were checked with a reference lamp (FieldCal, TriOS GmbH, Germany) before and after the cruise, confirming that the initial calibration was still valid.

The handling of the HyperPro II followed the same protocol as in Holinde and Zielinski (2016), Mascarenhas et al. (2017), and Wollschläger et al. (2020d): Prior to the deployment of the HyperPro II at a station, the depth sensor was tared on deck of RV *Heincke* (Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 2017) with the instrument in an upright position in order to adjust it to the current air pressure and ensuring correct in-water readings of the depth. Afterwards, the HyperPro II was deployed from the ship's stern, letting it drift to a distance of approx. 30 m to avoid shadow anomalies on the underwater measurements caused by the ship and its superstructures. Per station, one to three profiles were taken, depending on available time. All profiles were done as deep as possible (limited by the length of the instrument cable), but at least until the lower limit of the euphotic zone (depth in which 1% of surface PAR is available). Data were recorded using the SatView software (version 2.9.5_7). During data processing readings corresponding to an instrument tilt of >5° were discarded, as a vertical orientation of the instrument is necessary for correct measurements.

MODELLING

For modelling purposes, data from the HE533 Expedition (Voß et al., 2020e) was used after pre-processing, i.e. normalisation and removal of data records with missing values. Random sampling without replacement was applied, to split the HE533 data into a training set (70 %) and a test set (30 %). The training set was used to learn three different AI based models, i.e. a Linear Regression model (LR), an Artificial Neural Network model (ANN), and a Regression Tree model (RT).

The test set was then used to validate the models generated in terms of accuracy. The outcome of this validation serves as a baseline to investigate the generalisability of the different models to measurements in other geolocations.

The models generated on HE533 where then applied on the other datasets available and evaluated in terms of accuracy. This accuracy was then compared with the baseline accuracy calculated from the HE533 test data. The modelling approach described is visualised in Figure 2.

EXPERIMENTAL SETUP

Publicly available datasets from different ship cruises are used. All datasets can be found on the data portal Pangaea (www.pangaea.de). The data from the cruise HE533 (Voß et al., 2020e) was used to train the different models, while the data from the other cruises was used for validation (Friedrichs et al., 2020; Mascarenhas et al., 2020; Voß et al., 2020f, 2020a, 2020b, 2020c, 2020d; Wollschläger et al., 2020a, 2020b, 2020c). The HE533 dataset contains originally 9858 tuples of which

37.77 % had to be discarded because of missing values. The combined dataset for validation contains 64060 tuples of which 23.05 % for experiment 1 and 23.14 % for experiments 2 and 3 had to be discarded also because of missing values.

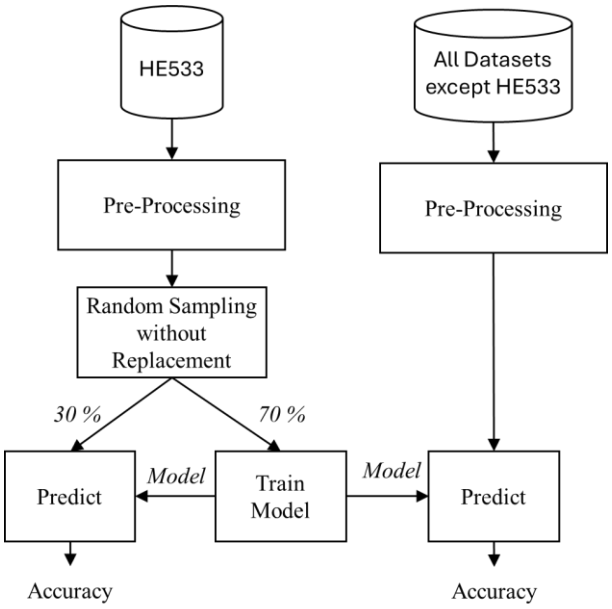


Figure 2: Modelling approach used

All models were generated using the KNIME workbench (Berthold et al., 2009). An ANN was used with one hidden layer containing 100 hidden units and trained for 1,000 epochs, using adaptive RProp (Riedmiller and Braun, 1993). For the RT the procedure described by Breiman et al. (1984) is applied with a couple of simplification, for instance no pruning, not necessarily binary trees. LR model uses standard multiple linear regression (Freedman, 2009).

Three sets of experiments were carried out. In all the experiments, the models were trained using the HE533 dataset. In the first set of experiments, the models were trained on three wavelengths measured in the water column (E_d), 400 nm, 412 nm, and 490 nm, based on (Stahl et al., 2022). In the second set of experiments, the full spectrum of the surface light (E_s) between 400 nm and 700 nm, in 1 nm steps, was added to the inputs. In the third set of experiments, the full E_s spectrum was replaced by the same wavelengths that were used from the underwater light field. The results of the experiments are presented in the next section.

EXPERIMENTAL RESULTS AND DISCUSSION

For comparing the models, the R^2 values were calculated on the test data (see Figure 2). The R^2 value was chosen as metric to ensure comparability with previously published results (Kumm et al., 2022; Stahl et al., 2022). The results on the three experiments can be found in Table 1, where the R^2 values on HE533 correspond to the left hand side of Figure 2, whereas the R^2 for all datasets except HE533 correspond to the right hand side of Figure 2.

As can be seen in Table 1, the R^2 values on all datasets are lower compared with R^2 values on test data from HE533. This was expected since the additional data was not involved in training the models and were recorded in different geolocations with different physical properties.

Table 1: R^2 values for different models using Multiple Linear Regression (LR), Neural Network (ANN) and Regression Tree (RT).

Experiment #	Trained on	Model	R^2 on HE533	R^2 (all Datasets except HE533)
1	**HE533 E_d(400), E_d(412), E_d(490)**	LR	0.984	0.884
		ANN	0.986	0.879
		RT	0.972	0.821
2	**HE533 E_s(full spectrum) and E_d(400), E_d(412), E_d(490)**	LR	0.984	0.035
		ANN	0.989	0.899
		RT	0.977	0.795
3	**HE533 E_s(400), E_s(412), E_s(490) and E_d(400), E_d(412), E_d(490)**	LR	0.982	0.880
		ANN	0.986	0.919
		RT	0.973	0.822

When comparing Experiment 2 with Experiment 1, one can observe that the R^2 values are in the same order of magnitude for the evaluation on HE533, i.e. there was no improvement. However, when comparing results for all datasets, it can be observed that for ANNs the accuracy increases by 2.0 % whereas the performance for the regression tree deceases by 2.6 %. Noticeable, the linear regress decreases in performance by 84.9 %. It is believed that this underperformance is caused by outliers in some of the additional spectral information from the surface light. The linear regression approach will consider all spectral information including outliers. On the other hand, regression trees perform an internal selection of the best spectral information for branching and building the tree structure. Therefore, outliers may not be selected for branching. A neural network can also cope very well with outliers, since they can model non-linear dependencies.

When comparing Experiment 3 with Experiment 1 one can observe that the R^2 values are in the same order of magnitude, even for linear regression. This is in line with the observations about linear regression performance in Experiment 2, since in Experiment 3 a limited spectrum, i.e. number of input variables, was used. The datasets were normalised before training and validation took place.

Comparing results on HE533, with the results on all datasets except HE533 and for all experiments, once can see that the accuracy drops by approximately 10 % using median. The neural network-based model outperformed linear regression and regression tree-based models. This is probably because there are some non-linear factors that a neural network can compensate better. These results are in line with the findings reported by Kumm et al. (2022).

It was shown that spectral information from the surface light can be used to improve the generalisability of the models, especially of the ANN.

CONCLUSIONS AND FUTURE WORK

The paper presented a modelling approach for predicting PAR in the water column, which uses selected spectral information within the water column and additionally surface spectral information. Three different AI-based modelling approaches where used. It was shown that the ANN approach outperformed the RT and LR models. It was also shown that the models generalise well on data recorded in other geolocations without additional modification or re-training.

It should be noted that the parameter settings of the models have not been optimised yet. Therefore, further improvements are potentially possible. The selection of spectral variables was based on the literature. However, it is conceivable that different spectral information may result in more accurate models. Also, other environmental parameters such as e.g. pressure or salinity could potentially improve the models. Therefore, a more systematic variable selection process will be investigated in the future. In addition, methods to improve linear regression models, such as regression splines (Friedman, 1991) or generalised additive models (Wood et al., 2015), will be investigated.

ACKNOWLEDGEMENTS

This work was funded by the Ministry of Science and Culture, Lower Saxony, Germany, through funds from the Niedersächsische Vorab (ZN3480).

REFERENCES

Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 2017. Research Vessel HEINCKE Operated by the Alfred-Wegener-Institute. Journal of large-scale research facilities JLSRF 3, A120–A120. https://doi.org/10.17815/jlsrf-3-164

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., 2009. KNIME - the Konstanz information miner: version 2.0 and beyond. SIGKDD Explor. Newsl. 11, 26–31. https://doi.org/10.1145/1656274.1656280

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Routledge, New York. https://doi.org/10.1201/9781315139470

Freedman, D.A., 2009. Statistical Models: Theory and Practice, 2nd ed. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511815867

Friedman, J.H., 1991. Multivariate Adaptive Regression Splines. The Annals of Statistics 19, 1–67. https://doi.org/10.1214/aos/1176347963

Friedrichs, A., Schwalfenberg, K., Voß, D., Wollschläger, J., Zielinski, O., 2020. Hyperspectral underwater light field measured during the cruise MSM56 with RV MARIA S. MERIAN. https://doi.org/10.1594/PANGAEA.917534

Holinde, L., Zielinski, O., 2016. Bio-optical characterization and light availability parameterization in Uummannaq Fjord and Vaigat–Disko Bay (West Greenland). Ocean Science 12, 117–128. https://doi.org/10.5194/os-12-117-2016

Kumm, M.M., Nolle, L., Stahl, F., Jemai, A., Zielinski, O., 2022. On an Artificial Neural Network Approach for Predicting Photosynthetically Active Radiation in the Water Column, in: Bramer, M., Stahl, F. (Eds.), Artificial Intelligence XXXIX, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 112–123. https://doi.org/10.1007/978-3-031-21441-7_8

Mascarenhas, V.J., Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020. Hyperspectral underwater light field measured during the cruise MSM65 with RV MARIA S. MERIAN. https://doi.org/10.1594/PANGAEA.917564

Mascarenhas, V.J., Voß, D., Wollschläger, J., Zielinksi, O., 2017. Fjord light regime: Bio-optical variability, absorption budget, and hyperspectral light availability in Sognefjord and Trondheimsfjord, Norway. Journal of Geophysical Research: Oceans 122, 3828–3847. https://doi.org/10.1002/2016JC012610

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: the RPROP algorithm, in: IEEE International Conference on Neural Networks. Presented at the IEEE International Conference on Neural Networks, pp. 586–591 vol.1. https://doi.org/10.1109/ICNN.1993.298623

Sloyan, B., Roughan, M., Hill, K., 2018. Global Ocean Observing System.

Stahl, F., Nolle, L., Zielinski, O., Jemai, A., 2022. A Model for Predicting the Amount of Photosynthetically Available Radiation from BGC-ARGO Float Observations in the Water Column, in: ECMS 2022 Proceedings Edited by Ibrahim A. Hameed, Agus Hasan, Saleh Abdel-Afou Alaliyat. Presented at the 36th ECMS International Conference on Modelling and Simulation, ECMS, pp. 174–180. https://doi.org/10.7148/2022-0174

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020a. Hyperspectral underwater light field measured during the cruise SO248 with RV SONNE. https://doi.org/10.1594/PANGAEA.911988

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020b. Hyperspectral underwater light field measured during the cruise SO267/2 with RV SONNE. https://doi.org/10.1594/PANGAEA.912028

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020c. Hyperspectral underwater light field measured during the cruise SO245 with RV SONNE. https://doi.org/10.1594/PANGAEA.911558

Voß, D., Henkel, R., Wollschläger, J., Zielinski, O., 2020d. Hyperspectral underwater light field measured during the cruise SO254 with RV SONNE. https://doi.org/10.1594/PANGAEA.912001

Voß, D., Wollschläger, J., Henkel, R., Zielinski, O., 2020e. Hyperspectral underwater light field measured during the cruise HE533 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.918041

Voß, D., Wollschläger, J., Henkel, R., Zielinski, O., 2020f. Hyperspectral underwater light field measured during the cruise HE492 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.918047

Wang, L., Gong, W., Li, C., Lin, A., Hu, B., Ma, Y., 2013. Measurement and estimation of photosynthetically active radiation from 1961 to 2011 in Central China. Applied Energy 111, 1010–1017. https://doi.org/10.1016/j.apenergy.2013.07.001

Wollschläger, J., Henkel, R., Voß, D., Zielinski, O., 2020a. Hyperspectral underwater light field measured during the cruise HE503 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.912073

Wollschläger, J., Henkel, R., Voß, D., Zielinski, O., 2020b. Hyperspectral underwater light field measured during the cruise HE516 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.912033

Wollschläger, J., Henkel, R., Voß, D., Zielinski, O., 2020c. Hyperspectral underwater light field measured during the cruise HE527 with RV HEINCKE. https://doi.org/10.1594/PANGAEA.912054

Wollschläger, J., Tietjen, B., Voß, D., Zielinski, O., 2020d. An Empirically Derived Trimodal Parameterization of Underwater Light in Complex Coastal Waters – A Case Study in the North Sea. Frontiers in Marine Science 7.

Wood, S.N., Goude, Y., Shaw, S., 2015. Generalized Additive Models for Large Data Sets. Journal of the Royal Statistical Society Series C: Applied Statistics 64, 139–155. https://doi.org/10.1111/rssc.12068

AUTHOR BIOGRAPHY

CHRISTOPH THOLEN is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), in the Marine Perception research department. His current research interests including the application of Artificial Intelligence applied to the maritime context, with a special focus on the identification and quantification of plastic litter using remote sensing. He received his doctoral degree in 2022 from the Carl von Ossietzky University of Oldenburg. From 2016 to 2022, he worked on a joint project between the Jade University of Applied Science and the Institute for Chemistry and Biology of the Marine Environment (ICBM), at the Carl von Ossietzky University of Oldenburg for the development of a low cost and intelligent environmental observatory.

LARS NOLLE graduated from the University of Applied Science and Arts in Hanover, Germany, with a degree in Computer Science and Electronics. He obtained a PgD in Software and Systems Security and an MSc in Software Engineering from the University of Oxford as well as an MSc in Computing and a PhD in Applied Computational Intelligence from The Open University. He worked in the software industry before joining The Open University as a Research Fellow. He later became a Senior Lecturer in Computing at Nottingham Trent University and is now a Professor of Applied Computer Science at Jade University of Applied Sciences. He also is affiliated with the Marine Perception research department at the German Research Center for Artificial Intelligence (DFKI). His main research interests are computational

optimisation methods for real-world scientific and engineering applications.

JOCHEN WOLLSCHLÄGER is a senior scientist in the working group Marine Sensor systems at the Institute for Chemistry and Biology of the Marine Environment at the Carl-von-Ossietzky-Universität Oldenburg. Being a biologist by training, he got his PhD in 2013 from the Jacobs University (now Constructor University) in Bremen and is working in the field of aquatic sensors for almost 14 years. His work focuses on the application of new and established bio-optical *in situ* instruments for the characterization of the inherent and apparent optical properties of the water. From these data, the relationship to the optically active substances (phytoplankton, CDOM, and non-algal particles) in the water are investigated to obtain ecologically relevant information.

FREDERIC STAHL is Principal Researcher at the German Research Center for Artificial Intelligence (DFKI), where he is heading the Marine Perception research department. He has been working in the field of Data Mining for more than 17 years. His particular research interests are in (i) developing scalable algorithms for building adaptive models for real-time streaming data and (ii) developing scalable parallel Data Mining algorithms and workflows for Big Data applications. In previous appointments Frederic worked as Associate Professor at the University of Reading, UK, as Lecturer at Bournemouth University, UK and as Senior Research Associate at the University of Portsmouth, UK. He obtained his PhD in 2010 from the University of Portsmouth, UK and has published over 85 articles in peer-reviewed conferences and journals.

On the performance evaluation of synchronous and asynchronous parallel particle swarm optimisation

Christoph Tholen[1] and Lars Nolle[1,2]

[1]German Research Center for Artificial Intelligence
Research Department Marine Perception
Marie-Curie-Straße 1
26129 Oldenburg, Germany
Email: {christoph.tholen|lars.nolle}@dfki.de

[2]Jade University of Applied Science
Friedrich-Paffrath-Straße 101
26919 Wilhelmshaven, Germany
Email: lars.nolle@jade-hs.de

KEYWORDS

Artificial Intelligence, Optimisation, Search Heuristics, Distributed Computing, PAPSO, PSPSO.

ABSTRACT

In this work, the efficiency (time) and effectivity (fitness) of two parallel variants of Particle Swarm Optimisation (PSO) have been evaluated, the synchronous PSPSO and the asynchronous PAPSO. In this study, an implementation of PAPSO is utilised, which deviates from the master-slave principle. Instead, all particles function as independent workers, competing for the available computing resources. If a particle discovers a new best position, it shares this information with the other particles. Two well-known test functions, the Rosenbrock function and the Rastigin function, were applied for evaluating the efficiency and effectivity of PSPSO and PAPSO. Firstly, versions of the test functions with 10, 30, and 60 dimensions were used. The population size was increased for each dimensionality from 50 to 100 and finally 200 particles. The results of this set of experiments showed that both variants of PSO performed similar regarding to their effectiveness of finding the optimum solutions. The computing time used by PAPSO, on the other hand, is significantly smaller than the computing time needed by PSPSO. On average the PAPSO was 69.1 % faster than the PSPSO on the Rosenbrock function and 90.3 % faster on the Rastigin function. In a second set of simulations, the maximum waiting time was varied from 5 ms to 1,000 ms. It is shown for both algorithms, that the average computing time rises linearly with the maximum waiting time.

INTRODUCTION

Computational optimisation is an important tool for science and engineering. In the past, various optimisation algorithms were proposed, for instance Genetic Algorithm (GA) (Holland, 1975), Simulated Annealing (SA) (Kirkpatrick, 1984), or Particle Swarm Optimisation (PSO) (Kennedy and Eberhart, 1995). Usually, optimisation algorithms start with an initial candidate solution, which is refined iteratively, until a stopping criterion is met. The refinement of the solution is undertaken based on method-dependent strategies. For each refinement the fitness function needs to be evaluated. The time required for evaluating the fitness function depends on the optimisation problem at hand. Therefore, fitness function evaluation can be seen as the bottleneck of optimisation algorithms. Parallel optimisation algorithms can be used to overcome this bottleneck and make use of the computing power of modern computers (Nolle and Werner, 2017). However, if the time needed for fitness evaluation is dependent on the input parameters, synchronous parallelisation might not be able to unfold its full potential (Nolle and Werner, 2017; Tholen et al., 2019).

In this research the performance of a variant of PSO, called parallel asynchronous particle swarm optimisation (PAPSO) is evaluated empirically.

Particle Swarm Optimisation

Particle Swarm Optimisation (PSO) is inspired by the collective behaviour of real-world entities, such as fish schools or flocks of birds, that collaborate to achieve a shared objective (Kennedy and Eberhart, 1995). Each member of the swarm conducts an individual search, yet the search behaviour of each particle is influenced by other swarm members. At the start of a search, every particle in the swarm begins at a random position and is assigned a randomly selected velocity for each dimension of the n-dimensional search space. Subsequently, the particles traverse the search space with an adjustable velocity determined by factors including their current fitness value, the best solution identified by the particle (cognitive knowledge), and the best solution found by the entire swarm (social knowledge) (1):

$$\vec{v}_{i+1} = \vec{v}_i \cdot \omega + r_1 \cdot c_1(\vec{p}_b - \vec{x}_i) + r_2 \cdot c_2(\vec{g}_b - \vec{x}_i). \quad (1)$$

Where $\vec{v}_{i+1}$ denotes the new velocity of a particle and $\vec{v}_i$ represents the current velocity of a particle. The variables ω, c_1, and c_2 denote the control parameters of the algorithm, named inertia weight, cognitive scaling factor, and social scaling factor respectively. The variables r_1 and r_2 represent random numbers generated from the interval $\{0,1\}$. The variable $\vec{x}_i$ represents the current position of a particle, while $\vec{p}_b$ denotes the best-known position of a particle, and $\vec{g}_b$ the best-known position of the entire swarm.

The next position of a particle can be calculated as follows:

$$\vec{p}_{i+1} = \vec{p}_i + \vec{v}_{i+1} \cdot \Delta t. \qquad (2)$$

Where $\vec{p}_{i+1}$ denotes the new position of a particle, $\vec{p}_i$ denotes current position of a particle, $\vec{v}_{i+1}$ represents the new velocity of a particle, and Δt correspond to a time step (Nolle, 2015).

The performance of PSO heavily depends on the chosen value of the control parameters (Shi and Eberhart, 1998). In real world application, usually a parameter search is conducted to find suitable values of the parameters. However, in this research the control parameters were set to standard values (Jiang et al., 2017; Umarani and Selvi, 2010) ω= 0.729, $c_1 = c_2$= 1.49. The velocity vector $\vec{v}_0$ was initialised to zero, to speed up the search (Engelbrecht, 2012). However, the intention of this study is not to find optimal control parameter settings for the test functions utilised, but to compare the performance of the different parallel implementations of PSO.

Parallel Particle Swarm Optimisation

Due to the substantial computational expenses required for optimising real-world problems, various parallel adaptations of the Particle Swarm Optimization (PSO) have been proposed in previous studies (Koh et al., 2006; Schutte et al., 2004). The categorisation of parallel optimisation algorithms includes synchronous and asynchronous variants (Koh et al., 2006).

While the majority of parallel algorithms suggested in the literature utilise a synchronous software architecture (Koh et al., 2006), a notable drawback of synchronous optimisation algorithms is the necessity to balance the workload among all workers to prevent idle states. This balance cannot be guaranteed if the time for fitness evaluation is dependent on the input vector of the fitness function (Koh et al., 2006).

The Parallel Asynchronous Particle Swarm Optimization (PAPSO) introduced by Koh et al. (2006) mitigates the downsides of the synchronous PSO version. This PAPSO algorithm adheres to the master-slave principle, where the master thread maintains a queue of all particles ready for evaluation and handles the decision-making process, including calculating the next position for all particles. The master assigns the initial particle (candidate solution) in the queue to a free thread (slave), which then evaluates the fitness function and reports the fitness value back to the master. The master compares the returned value with the personal best of the particle or the global best and updates the relevant values if necessary. Afterward, the master assigns the next particle in the queue to the slave. The task-queue is designed to ensure that all particles undergo roughly the same number of function evaluations (Koh et al., 2006).

In this study, a distinct implementation of PAPSO, introduced in a previous study (Tholen et al., 2019) is utilised. This implementation deviating from the master-slave principle. Instead, all particles function as independent workers, competing for the available computing resources. If a particle discovers a new best position, it shares this information with the other particles.

The next section describes the fitness functions used for comparing the performances of Parallel Synchronous Particle Swarm Optimisation (PSPSO) and PAPSO.

Fitness Functions Used

In this study, two well-known fitness functions, the Rosenbrock function (Rosenbrock, 1960) and the Rastrigin function (Rastrigin, 1974), are used for benchmarking the PSO variants under investigation.

The Rosenbrock function in a unimodal *n*-dimensional function often used for evaluating optimisation algorithms:

$$f(x) = \sum_{i=1}^{d-1}[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]. \qquad (3)$$

Here, d denotes the number of dimensions, i.e. the number of inputs, of the function. The function has a distinctive long, narrow, and curved valley, which presents a challenging landscape for optimisation algorithms.

The Rastrigin function is a multimodal *n*-dimensional function also often used as a test function for evaluating optimisation algorithms:

$$f(x) = 10d + \sum_{i=1}^{d}[x_i^2 - 10\cos(2\pi x_i)]. \qquad (4)$$

In (4) d also represents the dimensionality of the function. In this highly multimodal function, the locations of the minima are regularly distributed.

Researchers commonly employ these two test functions in comparative studies to gauge the efficacy and adaptability of various optimization techniques in navigating its intricate solution space (Clerc and Kennedy, 2002; Gaviano et al., 2012).

According to Venter (2006), synchronous PSO implementations led to poor parallel speedup in cases were the calculation of the fitness value depends on the candidate solutions x being analysed.

Since the evaluation time of the test functions mentioned above do not depend on x, a time delay based on x is introduced as follows:

$$t_w(x) = \frac{\sum_{i=1}^{d} x_i}{x_{max}} \cdot t_{max} \quad (5)$$

Where t_w is the time delay, x_{max} represents the maximum sum of the input vector x, and t_{max} corresponds to the maximum waiting time. Figure 1 shows the pseudocode for fitness function evaluation used for the experiments.

```
Fitness(x) do
  if Rosenbrock do
    f(x) := Eq. 3
  else do
    f(x) := Eq. 4
  endif
  tw := Eq. 5
  sleep for tw
  return f(x)
end
```

Figure 1: Pseudocode for fitness function used

As it is shown in Figure 1, the fitness values are calculated using the original test functions before waiting for t_w.

EXPERIMENTAL SETUP

Similar experiments were carried out for both algorithms under investigation, i.e. PAPSO and PSPSO. Within these experiments, all possible permutations of the parameters dimensionality d and number of particles n were evaluated using the following values:

$$d = \{10,30,60\}, \quad n = \{50,100,200\}. \quad (6)$$

For each combination, 200 runs of the algorithm were conducted. In each run, the number of iterations was chosen to be 1,000. For all experiments, the maximum waiting time t_{max} was set to 100 ms.

Since for real-world applications, the time consumed by the fitness evaluation is unknown, a second set of experiments was conducted, varying the maximum waiting time t_{max} as follows:

$$t_{max} = \{5, 10, 50, 100, 250, 500, 1000\}\, ms. \quad (7)$$

For this set of experiments, only the Rosenbrock function with d=30 was used, while the number of particles was set to n=100.
The results of both sets of experiments are given and discussed in the next section.

RESULTS AND DISCUSSION

Figure 2 provides an example of the fitness over time for 200 runs as a waterfall chart for the PAPSO, with n = 200 on the Rastrigin function with d = 60. It can be observed that in all runs the PAPSO converged after approximately 500 ms. The maximum time for completing the 1,000 iterations was 3.66 seconds, whereas the minimum time was 1.89 seconds.

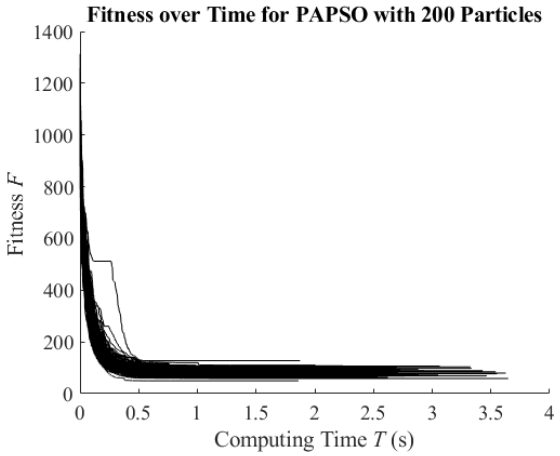


Figure 2: Fitness over time for Rastrigin with d =60 PAPSO with n = 200

Figure 3 depicts another example of the fitness over time as a waterfall chart for the 200 runs for the PSPSO, using the same parameters as above. It is shown that in all runs the PSPSO converged after approximately 10,000 ms. The maximum time for completing the 1,000 iterations was 47.20 seconds, whereas the minimum time was 34.46 seconds.

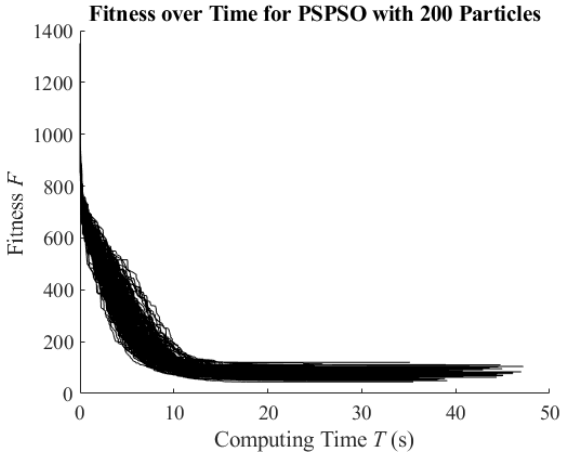


Figure 3: Fitness over time for Rastrigin with d =60 PSPSO with n = 200

The median fitness F achieved in this experiment by PAPSO was 81.59, while the median fitness achieved by

PSPSO was 79.10. Their significance level here is p =0.1159 and hence, both algorithms performed similar regarding the quality of the solutions.

The summarised results of the first set of experiments can be found in Tables 1-4. While Table 1 and 3 are summarising the computing time T of the different experiments for the Rosenbrock and the Rastrigin function respectively, and in Table 2 and 4 the statistics of the fitness F values are given.

It can be seen from Table 2 that for PSPSO performing on the 10-dimensional Rosenbrock function, one run resulted in a local optimum, i.e. the algorithm was not able to find the global optimum. However, this is not uncommon when applying search heuristics.

The following discussion of the first set of experiments is based on the median values. For all functions and chosen values of d it can be seen that $\tilde{F}_{50} \leq \tilde{F}_{100} \leq \tilde{F}_{200}$, which was expected. For the Rosenbrock function, the median of the computing time T for PAPSO seems to be unaffected by the number of particles, while for PSPSO it can be observed that $\tilde{T}_{n=50} \geq \tilde{T}_{n=100} \geq \tilde{T}_{n=200}$. For all experiments, it was shown that $\tilde{T}_{PSPSO} \geq \tilde{T}_{PAPSO}$.

For the computing time T, it can be observed that the significance level p is <0.0001 for all combinations. Hence, the computing time used by PAPSO is significantly smaller than the computing time needed by PSPSO.

For the Rosenbrock function, seven out of nine combinations resulted in a *p–value* > 0.05. Therefore, the results are not significantly different, while for the Rastigin Function only two out of nine combinations showed this behaviour.

Table 1: Summarised Results (Computing Time T) for different values of d and n on Rosenbrock function

		10 Dimensions			30 Dimensions			60 Dimensions		
	n	**50**	**100**	**200**	**50**	**100**	**200**	**50**	**100**	**200**
PAPSO	**Avg.**	7.29	7.79	8.27	5.79	5.49	5.51	8.38	7.90	7.45
	Med.	7.06	7.50	8.17	5.32	4.90	4.61	8.38	8.01	7.50
	Std.	1.69	1.47	0.88	2.41	2.30	2.14	1.54	1.40	1.55
	Min.	4.30	4.28	6.26	1.70	2.73	2.90	4.57	4.21	3.73
	Max.	17.49	16.53	17.85	11.61	12.52	12.02	12.70	12.19	11.05
PSPSO	**Avg.**	16.72	25.89	46.57	14.68	23.25	42.28	14.35	22.86	42.94
	Med.	15.05	24.39	45.90	14.08	22.97	41.96	14.34	22.62	42.21
	Std.	4.29	3.14	4.44	2.69	2.70	3.26	1.38	1.71	3.23
	Min.	11.52	22.23	41.14	10.11	19.22	36.68	10.98	19.22	37.12
	Max.	55.59	37.93	64.46	25.05	33.07	52.21	19.64	29.43	51.54
	p	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 2: Summarised Results (Fitness F) for different values of d and n on Rosenbrock function

		10 Dimensions			30 Dimensions			60 Dimensions		
	n	**50**	**100**	**200**	**50**	**100**	**200**	**50**	**100**	**200**
PAPSO	**Avg.**	1.32	0.63	0.39	36.42	30.22	29.27	167.83	124.88	114.95
	Med.	0.76	0.43	0.28	23.48	22.16	21.20	153.07	125.05	102.13
	Std.	1.44	0.92	0.62	29.01	24.35	23.12	247.54	55.82	47.36
	Min.	0.00	0.00	0.01	0.00	0.11	0.01	4.21	4.65	2.02
	Max.	5.73	6.51	4.28	121.21	107.86	87.47	3525.94	409.89	285.39
PSPSO	**Avg.**	59.73	1.02	0.36	38.81	33.96	30.06	138.87	110.25	103.98
	Med.	0.68	0.47	0.32	24.15	22.81	22.12	146.61	105.81	101.54
	Std.	828.68	5.39	0.52	30.33	26.84	24.16	50.51	41.65	44.22
	Min.	0.00	0.00	0.00	0.13	0.00	0.06	35.77	22.07	17.25
	Max.	11720.40	75.70	4.80	139.86	142.30	109.78	305.52	234.73	227.52
	p	0.3194	0.3137	0.6004	0.4211	0.1452	0.7385	0.1058	0.0032	0.0171

Table 3: Summarised Results (Computing Time T) for different values of d and n on Rastrigin function

		10 Dimensions			30 Dimensions			60 Dimensions		
	n	**50**	**100**	**200**	**50**	**100**	**200**	**50**	**100**	**200**
PAPSO	**Avg.**	1.84	2.36	2.84	1.41	1.80	2.12	1.37	1.69	2.32
	Med.	1.18	1.75	2.59	1.00	1.38	1.83	1.18	1.60	2.21
	Std.	1.52	1.56	1.18	1.17	1.07	0.86	0.58	0.37	0.35
	Min.	0.46	0.71	1.10	0.80	0.90	1.22	1.03	1.41	1.89
	Max.	7.81	9.67	6.47	11.62	6.01	6.06	4.57	3.99	3.66
PSPSO	**Avg.**	12.37	20.86	40.27	11.45	20.51	38.87	10.79	19.47	37.77
	Med.	11.43	20.29	39.78	10.42	19.99	38.32	9.93	18.43	37.51
	Std.	2.61	2.64	3.35	2.65	2.45	3.02	1.93	2.15	2.85
	Min.	9.03	17.29	34.93	8.85	17.58	34.87	9.01	17.20	34.86
	Max.	21.08	32.50	51.62	25.64	28.10	49.96	17.96	29.78	47.20
	p	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 4: Summarised Results (Fitness F) for different values of d and n on Rastrigin function

		10 Dimensions			30 Dimensions			60 Dimensions		
	n	**50**	**100**	**200**	**50**	**100**	**200**	**50**	**100**	**200**
PAPSO	**Avg.**	5.17	3.37	2.22	40.36	33.72	28.07	105.92	95.91	81.46
	Med.	4.97	2.98	1.99	40.30	33.83	27.36	104.47	95.52	81.59
	Std.	2.61	1.69	1.25	9.04	7.61	6.72	18.20	14.73	11.83
	Min.	0.00	0.00	0.00	17.91	16.91	11.94	62.68	61.69	49.75
	Max.	16.91	8.95	6.96	68.65	58.70	46.76	164.21	132.33	127.36
PSPSO	**Avg.**	4.43	2.99	1.77	37.64	31.50	27.02	96.54	88.96	79.46
	Med.	3.98	2.98	1.99	35.82	31.34	27.36	96.51	87.56	79.10
	Std.	2.01	1.62	1.10	8.72	7.85	6.26	15.89	15.88	13.50
	Min.	0.99	0.00	0.00	12.93	12.93	9.95	53.74	44.77	44.77
	Max.	10.94	8.95	4.97	69.65	57.71	42.78	143.27	146.26	120.39
	p	0.0016	0.0222	0.0002	0.0023	0.0043	0.1067	< 0.0001	< 0.0001	0.1159

Table 5: Summarised Results (Computing Time T) for different values of maximum waiting time t_{max}

	t_{max}	**5**	**10**	**50**	**100**	**250**	**500**	**1000**
PAPSO	**Avg.**	1.79	2.76	3.58	5.49	12.69	23.21	43.44
	Med.	1.79	2.69	3.26	4.90	11.01	21.53	37.11
	Std.	0.10	0.73	1.36	2.30	5.78	10.79	21.61
	Min.	1.58	1.63	1.78	2.73	5.36	10.13	16.02
	Max.	2.11	4.50	7.46	12.52	31.33	55.92	118.80
PSPSO	**Avg.**	18.15	18.03	20.38	23.25	31.95	48.29	75.31
	Med.	17.80	17.81	20.33	22.97	31.03	45.47	69.80
	Std.	0.75	0.63	1.61	2.70	5.76	12.98	22.86
	Min.	17.64	17.48	18.08	19.22	23.74	29.63	42.98
	Max.	20.35	21.89	26.42	33.07	52.73	98.13	144.60
	p	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 6: Summarised Results (Fitness F) for different values of maximum waiting time t_{max}

	t_{max}	**5**	**10**	**50**	**100**	**250**	**500**	**1000**
PAPSO	**Avg.**	33.28	32.60	31.35	30.22	34.67	34.95	37.33
	Med.	21.77	21.76	21.98	22.16	22.47	22.34	22.81
	Std.	29.71	28.25	23.99	24.35	26.63	28.36	28.14
	Min.	0.00	0.00	0.11	0.11	0.05	0.04	0.15
	Max.	90.14	91.96	88.09	107.86	88.63	140.59	138.36
PSPSO	**Avg.**	36.42	36.70	34.65	33.96	35.13	32.61	32.59
	Med.	23.06	23.06	22.98	22.81	22.82	22.89	22.97
	Std.	26.15	27.60	27.69	26.84	26.45	27.35	24.83
	Min.	0.17	0.02	0.01	0.00	0.06	0.00	0.00
	Max.	106.46	87.18	152.92	142.30	99.55	195.20	89.60
	p	0.2626	0.1429	0.2035	0.1452	0.8625	0.4015	0.0748

The results achieved by the second set of experiments can be found in Table 5 and Table 6 and are summarised in Figure 4. In this figure, the average computing time is plotted over the maximum waiting time. It can be observed that for both algorithms the computing time increases linearly with increasing maximum waiting time. The R^2 value of the linear regression is 0.999 for PSPSO and 0.9992 for PAPSO. Therefore, the average computing time rises linearly with the maximum waiting time for both algorithms.

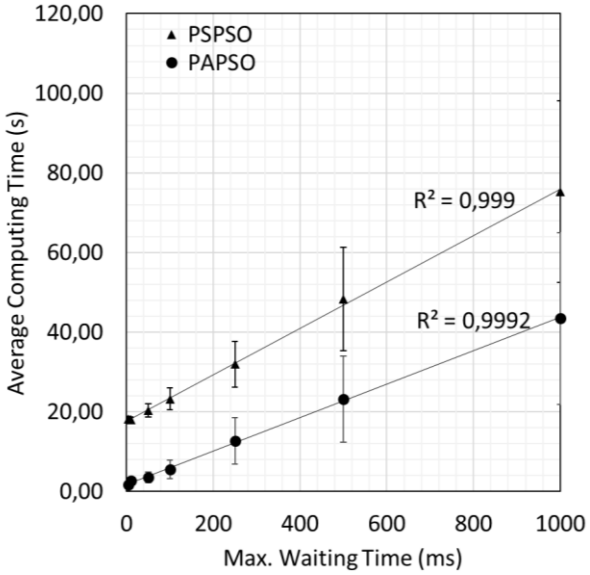


Figure 4: Average computing time per run for different maximum waiting times t_{max}

CONCLUSIONS AND FUTURE WORK

In this work, the efficiency (time) and effectivity (fitness) of two parallel variants of PSO have been evaluated, the synchronous PSPSO and the asynchronous PAPSO. Two well-known test functions, the Rosenbrock function and the Rastigin function, were applied. For each experiment, 200 runs were carried out in order to analyse the results statistically.

In a first set of simulations, versions of the test functions with 10, 30, and 60 dimensions, i.e. inputs, were used. The population size was increased for each dimensionality from 50 to 100 and finally 200 particles. The results of this set of experiments showed that both variants of PSO performed similar regarding to their effectiveness of finding the optimum solutions. The computing time used by PAPSO, on the other hand, is significantly smaller than the computing time needed by PSPSO.

In a second set of simulations, the maximum waiting time was varied from 5 ms to 1,000 ms. Simulations were carried out on the Rosenbrock function only. In this set of experiments, the number of dimensions was set to30 and the population size was set to 200. It was showed for both algorithms, that the average computing time rises linearly with the maximum waiting time.

In conclusion, it can be said that for optimisation problems were the evaluation of a candidate solution depends heavily on the candidate solution itself, both algorithms achieve the same level of effectiveness, i.e. find similar good solutions. However, in terms of efficiency, PAPSO clearly outperforms PSPSO, i.e. uses less run time.

In this research the performance of PAPSO and PSPSO was evaluated on two different test functions utilising a set of standard values for the three control parameters. The next step of this research, the performance of PAPSO and PSPSO will be evaluated on real world scenarios, for instance to optimise the candidate selection system of the PlasticObs+ system (Tholen and Wolf, 2023).

ACKNOWLEDGEMENTS

This work was funded by the Ministry of Science and Culture, Lower Saxony, Germany, through funds from the Niedersächsische Vorab (ZN3480).

REFERENCES

Clerc, M., Kennedy, J., 2002. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evol. Computat. 6, 58–73. https://doi.org/10.1109/4235.985692

Engelbrecht, A., 2012. Particle swarm optimization: Velocity initialization, in: 2012 IEEE Congress on Evolutionary Computation. Presented at the 2012 IEEE Congress on Evolutionary Computation, pp. 1–8. https://doi.org/10.1109/CEC.2012.6256112

Gaviano, M., Lera, D., Mereu, E., 2012. A Parallel Algorithm for Global Optimization Problems in a Distribuited Computing Environment. AM 03, 1380–1387. https://doi.org/10.4236/am.2012.330194

Holland, J., 1975. Adaptation in Natural and Artificial Systems.

Jiang, C., Zhang, C., Zhang, Y., Xu, H., 2017. An improved particle swarm optimization algorithm for parameter optimization of proportional–integral–derivative controller. Traitement du signal 34, 93–110. https://doi.org/10.3166/ts.34.93-110

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization, in: Proceedings of ICNN'95-International Conference on Neural Networks. Presented at the Proceedings of ICNN'95-international conference on neural networks, IEEE, pp. 1942–1948.

Kirkpatrick, S., 1984. Optimization by simulated annealing: Quantitative studies. J Stat Phys 34, 975–986. https://doi.org/10.1007/BF01009452

Koh, B.-I., George, A.D., Haftka, R.T., Fregly, B.J., 2006. Parallel asynchronous particle swarm optimization. Int J Numer Methods Eng 67, 578–595. https://doi.org/10.1002/nme.1646

Nolle, L., 2015. On a search strategy for collaborating autonomous underwater vehicles. Mendel 2015, 159–164.

Nolle, L., Werner, J., 2017. Asynchronous Population-Based Hill Climbing Applied to SPICE Model Generation from EM Simulation Data, in: Bramer, M., Petridis, M. (Eds.), Artificial Intelligence XXXIV, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 423–428. https://doi.org/10.1007/978-3-319-71078-5_37

Rastrigin, L., 1974. Systems of Extreme Control.

Rosenbrock, H.H., 1960. An Automatic Method for Finding the Greatest or Least Value of a Function. The Computer Journal 3, 175–184. https://doi.org/10.1093/comjnl/3.3.175

Schutte, J.F., Reinbolt, J.A., Fregly, B.J., Haftka, R.T., George, A.D., 2004. Parallel global optimization with the particle swarm algorithm. Int J Numer Methods Eng 61, 2296–2315. https://doi.org/10.1002/nme.1149

Shi, Y., Eberhart, R.C., 1998. Parameter selection in particle swarm optimization, in: Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E. (Eds.), Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 591–600. https://doi.org/10.1007/BFb0040810

Tholen, C., Nolle, L., El-Mihoub, T., Dierks, J., Burger, A., Zielinski, O., 2019. Automated Tuning Of A Cellular Automata Using Parallel Asynchronous Particle Swarm Optimisation, in: ECMS 2019 Proceedings Edited by Mauro Iacono, Francesco Palmieri, Marco Gribaudo, Massimo Ficco. Presented at the 33rd International ECMS Conference on Modelling and Simulation, ECMS, pp. 30–36. https://doi.org/10.7148/2019-0030

Tholen, C., Wolf, M., 2023. On the Development of a Candidate Selection System for Automated Plastic Waste Detection Using Airborne Based Remote Sensing, in: Bramer, M., Stahl, F. (Eds.), Artificial Intelligence XL, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 506–512. https://doi.org/10.1007/978-3-031-47994-6_45

Umarani, R., Selvi, V., 2010. Particle swarm optimization-evolution, overview and applications. International Journal of Engineering Science and Technology 2.

Venter, G., Sobieszczanski-Sobieski, J., 2006. Parallel Particle Swarm Optimization Algorithm Accelerated by Asynchronous Evaluations. Journal of Aerospace Computing, Information, and Communication 3, 123–137. https://doi.org/10.2514/1.17873

AUTHOR BIOGRAPHY

CHRISTOPH THOLEN is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), in the Marine Perception research department. His current research interests including the application of Artificial Intelligence applied to the maritime context, with a special focus on the identification and quantification of plastic litter using remote sensing. He received his doctoral degree in 2022 from the Carl von Ossietzky University of Oldenburg. From 2016 to 2022, he worked on a joint project between the Jade University of Applied Science and the Institute for Chemistry and Biology of the Marine Environment (ICBM), at the Carl von Ossietzky University of Oldenburg for the development of a low cost and intelligent environmental observatory.

LARS NOLLE graduated from the University of Applied Science and Arts in Hanover, Germany, with a degree in Computer Science and Electronics. He obtained a PgD in Software and Systems Security and an MSc in Software Engineering from the University of Oxford as well as an MSc in Computing and a PhD in Applied Computational Intelligence from The Open University. He worked in the software industry before joining The Open University as a Research Fellow. He later became a Senior Lecturer in Computing at Nottingham Trent University and is now a Professor of Applied Computer Science at Jade University of Applied Sciences. He also is affiliated with the Marine Perception research department at the German Research Center for Artificial Intelligence (DFKI). His main research interests are computational optimisation methods for real-world scientific and engineering applications.

IMPACT OF THE VOLUME OF DEVELOPER HOUSING UNITS ON REAL ESTATE PRICES IN POLAND: CORRELATION AND COHERENCE ANALYSIS

Daria Wotzka[1], Łukasz Mach[2], Paweł Frącz[2], Bartosz Chorkowy[2]
Marzena Stec[3] and Ireneusz Dąbrowski[4]

[1]Faculty of Electrical Engineering Automatic Control and Informatics,
Opole University of Technology, e-mail: d.wotzka@po.edu.pl
[2]Faculty of Economics, University of Opole,
e-mail: {lmach@uni.opole.pl, p.fracz@gmail.com, bchorkowy@uni.opole.pl}
[3]Narodowy Bank Polski, Regional Branch in Opole, e-mail: 7m.stec@gmail.com
[4]Collegium of Management and Finance, Warsaw School of Economics,
e-mail: ireneusz.dabrowski@sgh.waw.pl

KEYWORDS

Real estate market, seasonality, price volatility, market equilibrium, wavelet coherence, correlation

ABSTRACT

This article conducts a comprehensive correlation analysis to explore the relationship between the development volume by developers and average market prices in Poland's housing sector from 2010 to 2023, utilizing quarterly data on the number of apartments released and their average prices. It employs both linear and nonlinear correlation analysis alongside wavelet coherence analysis. Preliminary correlational analysis offered insights into the basic interdependency patterns, highlighting how developer-supplied apartment numbers impact average prices. Wavelet coherence analysis, a more sophisticated approach, decomposed the data across various frequencies to uncover complex, nonlinear relationship patterns potentially missed by conventional correlation methods. The findings of the study highlight significant connections between the variables that vary over time and space, emphasizing the complexity of the housing market and underscoring the necessity for thorough analysis. These results carry important implications for developers, investors, and housing policy formulation, contributing significantly to understanding Poland's residential real estate market dynamics and supporting further research and strategic real estate industry planning.

INTRODUCTION

The residential real estate market is one of the key markets significantly impacting a country's economic situation. The condition of the real estate market directly influences the macroeconomic and financial stability of the economic system as well as the financial situation of the entities operating within it (Wang, 2021; Mach, 2019; Mach and Rącka, 2018). The condition of the real estate market is undoubtedly affected by the phenomenon of cyclicality, which occurs in the economy (Ben Zeev et al., 2017; Caunedo, 2020; Mandler and Scharnagl, 2022) as well as in the real estate market itself (Pyhrr et al., 1999; Jones & Trevillion, 2022; Devaney and Xiao, 2017). Depending on the phase of the business cycle, we can assess the real estate market as being in a boom or a bust. If there is a boom in the real estate market, it leads to increased activity of its users, whereas a bust decreases user activity (Łaszek and Olszewski, 2018; Łaszek et al., 2017; Chang, 2019; Devaney and Xiao, 2017; Gabrovski and Ortego-Marti, 2019). Prices of mixed-use properties undoubtedly also affect this market (Ghysels et al., 2013; Tsai, 2019; Wang, 2021). According to economic theory, the price level is the result of the actions of the demand and supply sides operating in the market. Market players' actions, over a long period, lead to the achievement of an equilibrium point in the real estate market, which, at least from a theoretical point of view, establishes the market equilibrium price (Fan et al., 2019; Ionașcu et al., 2019; Łaszek et al., 2016; Tsai, 2019). In economic theory, the issue seems simple and easy to model; however, due to its key role in the economy, the residential real estate market is often monitored by the government. State interventionism in the real estate market aims to minimize the risk of crisis on the one hand, but on the other hand can contribute to the disruption of economic laws. (Fernández Muñoz and Collado Cueto, 2017; Tomal, 2019). In Poland, such an example was the government's subsidies for taking out mortgages (the so-called 2% safe mortgage), which, on the one hand, avoid stagnation in the real estate market, while artificially stimulating the demand side of the market.
It is therefore crucial to study the impact of the volume of sales of developer apartments on the sales prices appearing on the market. Investigating and identifying the relationship between the volume of sales of apartments on the primary market and their price will yield knowledge useful to developers for, among other

things, estimating the profit margin they are able to earn on their planned development projects for sale or rent.

DATA USED IN THE RESEARCH

The data subjected to analysis in this article were obtained from the Local Data Bank website, made available by the Statistical Information Center of the Central Statistical Office in Poland. The variables under consideration included: the quantity of apartments released for occupancy by developers, the volume of apartments sold in the primary market, and the mean price of apartments within the primary market, all in form of quarterly time-series from 2010 to 2023. To enhance comparability and facilitate the exploration of interrelations, the data underwent normalization utilizing the Z-score methodology. The Z-score methodology is a statistical technique that quantifies the distance of a data point from the mean of a data-set, in terms of standard deviations. It essentially measures how much a data point differs from the average, considering the variability of the data-set. This method transforms the data into a standardized form, making it possible to compare observations from different scales or distributions. The calculation involves subtracting the mean from the data point and then dividing this difference by the standard deviation of the data-set. By doing so, it facilitates the identification of outliers and the comparison of data across different contexts, providing insights into the relative position of each data point within its distribution. The dynamic behavior of the Z-scored data-sets was graphically depicted across Figures 1 to 3 as blue line.

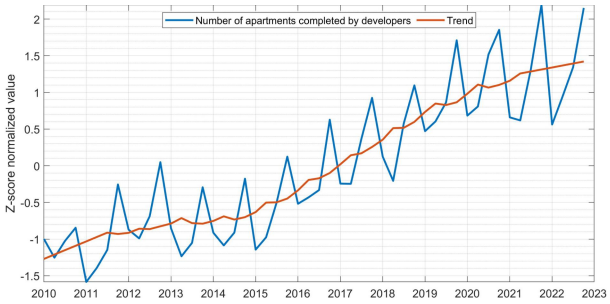


Figure 1: The quantity of apartments released for occupancy by developers and its trend

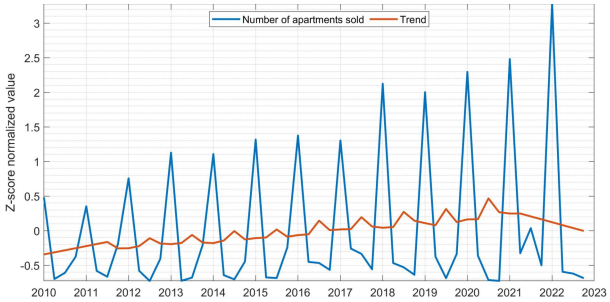


Figure 2: The quantity of apartments sold within the primary market and its trend

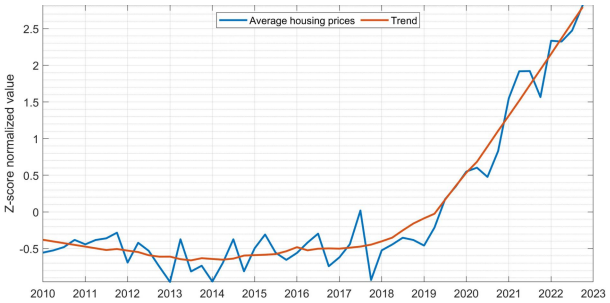


Figure 3: The prise of apartments sold within the primary market and its trend

Following the normalization process, the trend was delineated and the cyclical component extracted employing Sawitzky-Golay digital filters. The trend is indicated as red line in Figures 1-3. In Figure 4, the seasonal components of these variables are juxtaposed for their direct comparison, revealing a similar character in their dynamics. Savitzky-Golay filtration is a technique for smoothing numerical data, used to reduce noise while preserving the shape of the signal. The method involves fitting low-degree polynomials to subsets of data points in a moving window. This allows for the precise determination of local trends and signal characteristics without significantly distorting the data.
Based on the conducted signal processing procedure, the following variables were determined: the number of apartments made available for occupancy (completed) (V3), the number of apartments sold (V2), the average price of apartments (V1), and derivatives: the trend of the number of apartments made available, trend for the average price, and trend for the number of apartments sold, as well as, correspondingly: the seasonal component of the number of apartments made available for occupancy, the seasonal component of the average price, and the seasonal component of the number of apartments sold.

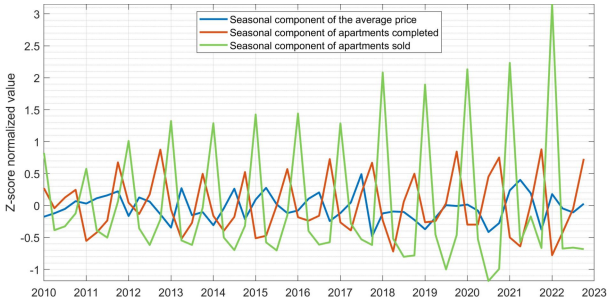


Figure 4: The seasonal components of the three considered market parameters

CORRELATION ANALYSIS RESULTS

In Figures 5 to 7, the scatter of values in three dimensions is depicted, showcasing the relationships among the three considered market parameters. The original data, although normalized, as well as the determined seasonal components and trends, are illustrated separately. Figure 5 displays the original data,

Figure 6 presents the deseasonalized data (trend), and Figure 7 shows the data in the form of the seasonal component. This three-dimensional visualization facilitates a comprehensive understanding of the intricate interdependencies between the variables, highlighting both their individual and combined effects. Through separating the original normalized data from the seasonal components and trends, the analysis provides a nuanced view of the data's underlying patterns and dynamics.

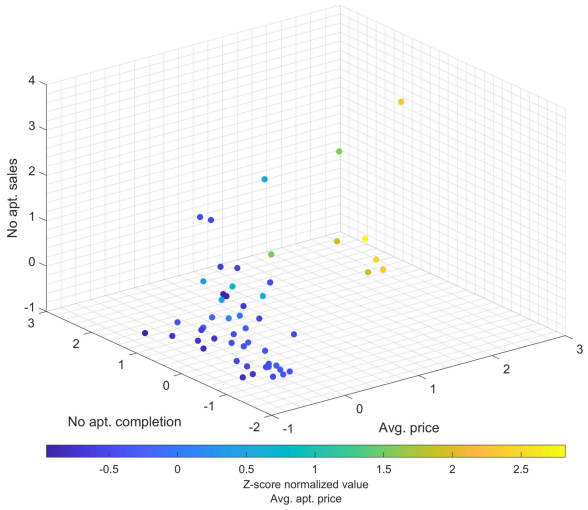


Figure 5: Scatter plot illustrating the dynamics between variables. Concerns to original data

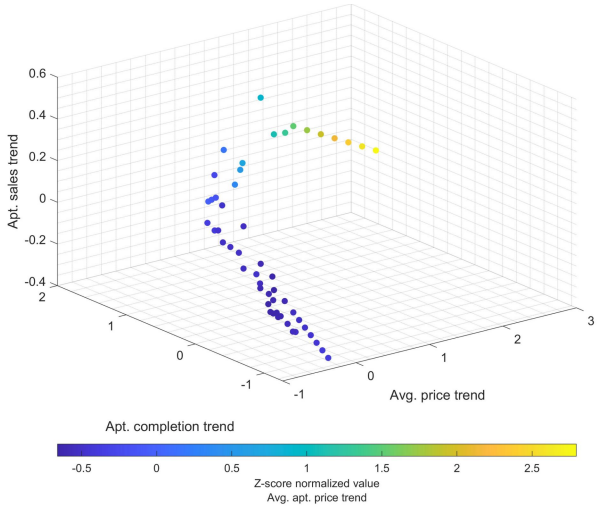


Figure 6: Scatter plot illustrating the dynamics between variables. Relates to the trend

To quantify the similarity in the considered market parameters, correlational analyses were employed, including Pearson's linear correlation analysis as well as Kendall's and Spearman's nonlinear correlation analyses. Pearson's correlation coefficient is a measure of the linear relationship between two variables, quantifying the degree to which they move together. It ranges from -1 to 1, where 1 means a perfect positive linear relationship, -1 means a perfect negative linear relationship, and 0 indicates no linear relationship. The calculation involves the covariance of the variables divided by the product of their standard deviations, essentially assessing how well a linear equation can describe the relationship between the two variables. Kendall's tau is a non-parametric measure used to assess the ordinal association between two measured quantities. It evaluates the strength and direction of the relationship between two variables by comparing the ranks of their data points. Kendall's tau considers the number of concordant and discordant pairs of data points, focusing on the consistency of the ordering in the data pairs across the entire dataset. Spearman's rank correlation coefficient, also a non-parametric measure, assesses how well the relationship between two variables can be described using a monotonic function. It ranks the data points for each variable and then calculates Pearson's correlation coefficient on these ranks. This method is particularly useful when the relationship between variables is not linear but still increases or decreases consistently. Both Kendall's and Spearman's methods do not require normality of the data distributions and are more robust against outliers than Pearson's correlation, making them suitable for ordinal data or when the assumptions of Pearson's correlation are not met.
In Tables 1 to 3, the values of both linear and nonlinear correlation coefficients are presented, calculated separately for the original data, trends, and cycles, across combinations of the variables under consideration.

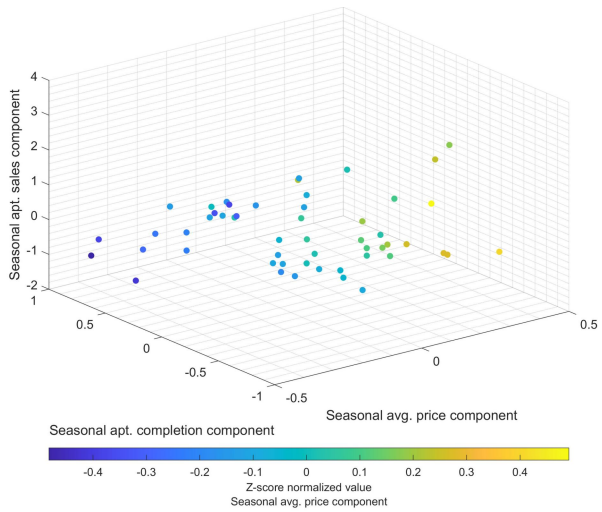


Figure 7: Scatter plot illustrating the dynamics between variables. Concerns seasonal components

It is apparent that the number of apartments completed and theirs average price (variables V1 and V3), when analyzed as original time series, exhibit statistically significant correlations (with a significance level set at 5%), evidenced by a Pearson correlation coefficient of 0.66 and a Spearman correlation coefficient of 0.54. However, these variables demonstrate considerably

weaker correlations, approximately 0.4, when only their seasonal components are analyzed. A strong correlation is observed between number of apartments completed and price (variables V1 and V3), as well as the number of apartments sold and theirs price (variables V2 and V3), when only their trends are considered, with Pearson correlation coefficients exceeding 0.8. Conversely, no correlation is evident between the number of apartments completed and sold (variables V2 and V1) in both the original and seasonal data, and between he number of apartments sold and theirs price (variables V2 and V3) in the original and seasonal data, as determined by the Spearman and Kendall methods.

Table 1: Comprehensive compilation of correlation coefficient values for the original data

Parameters	Correlation coefficients		
	Pearson	Spearman	Kendall
V1&V3	0.67	0.54	0.36
V2&V1	0.09	-0.12	-0.08
V2&V3	-0.02	0.03	0.01

Table 2: Comprehensive compilation of correlation coefficient values for the trend

Parameters	Correlation coefficients		
	Pearson	Spearman	Kendall
V1&V3	0.82	0.74	0.59
V2&V1	0.52	0.63	0.40
V2&V3	0.86	0.88	0.72

Table 3: Comprehensive compilation of correlation coefficient values for the seasonal components

Parameters	Correlation coefficients		
	Pearson	Spearman	Kendall
V1&V3	-0.42	-0.43	-0.31
V2&V1	-0.02	0.03	0.02
V2&V3	-0.37	-0.28	-0.19

WAVELET COHERENCE ANALYSIS RESULTS

As part of the research, wavelet coherence analysis (Grinsted et al., 2004) was performed on a combination of the considered real estate parameters. Wavelet coherence is an advanced data analysis method that enables the understanding of complex relationships between two signals or time series in the time and frequency domains. It facilitates the investigation of how the interdependence between variables changes over time and identifies periods in which the signals exhibit strong synchronization within specific frequency bands. This method is based on wavelet transformation, which decomposes the signal into frequency components over various time periods, allowing for the analysis of local signal properties. Wavelet coherence is expressed as a value between 0 and 1, where values close to 1 indicate strong coherence, denoting a high degree of interdependence between the time series in a given frequency range, while values close to 0 suggest a lack of such relationship. One of the key advantages of wavelet coherence is its ability to detect and illustrate phase and amplitude relationships between time series that may not be visible using traditional statistical methods. This enables a more detailed interpretation of the dynamics of relationships between variables, providing insight not only into whether variables are correlated but also how and when these correlations occur.

Scalograms depicting the outcomes for all nine variable combinations are presented in Figures 8 to 16. The arrows and colors on the charts are instrumental in interpreting the relationships between signals. The arrows indicate the phase relationship direction and nature between two signals. Arrows pointing to the right suggest the signals are in phase, implying synchronous movements of the variables. Conversely, arrows pointing to the left indicate the signals are in antiphase, where one variable reaches peaks as the other reaches troughs. Vertically oriented arrows show that one variable leads or lags the other by a quarter of a cycle. The colors on the chart reflect the coherence degree between signals across different frequency bands; warm colors denote high coherence, signifying a strong dependency between the time series, while cool colors imply low coherence, indicating a weak or absent dependency. Areas of statistical significance are marked on the chart, highlighted by outlines within the color areas, indicating statistically significant coherence between the signals and affirming the reliability of the detected relationships. Areas outside these demarcated boundaries may not be statistically significant, suggesting caution should be taken in interpretation.

Within the here considered data, a period of one quarter signifies fluctuations that recur on a quarterly basis. A period of four quarters, equivalent to one year, indicates annual cycles within the data. Larger period values suggest longer cycles, which may manifest over several years, revealing patterns of change that extend beyond the immediate temporal frame.

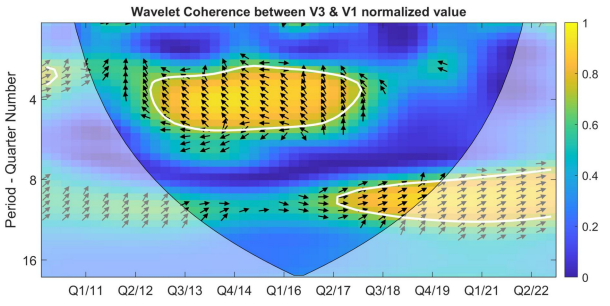


Figure 8: Wavelet coherence scalogram for V1 and V3 original data

The coherence scalogram analysis concerning the relationship between the number of apartments completed and theirs average price (variables V1 and V3, see Figures 8-10), taking into account the seasonal

component, reveals the presence of a strong and statistically significant anti-phase (observe arrows pointing to the left in Figures) correlation with an annual shift, but this is limited exclusively to the period from the second quarter of 2012 to the third quarter of 2018. After this period, a desynchronization was observed, which gave way, allowing for re-synchronization from the fourth quarter of 2019. In the context of classical correlation, the correlation coefficient did not exceed 0.44, whereas the scalogram highlights specific periods of intense coherence, approaching a value of 1. The reintegrated annual correlation observed after 2019 is not reflected in the original data, where a correlation with a mild two-year delay appears. Data corresponding to trends show phase-aligned (observe arrows pointing to the right in Figures) dependencies in biennial cycles starting from the first quarter of 2016 and annual cycles starting from the first quarter of 2021. Anti-phase consistency in the period from the third quarter of 2013 to the first quarter of 2016 is observed for annual cycles, regardless of the signal decomposition. It is noteworthy that the seasonal component has been in anti-phase since the end of 2019, while the trends and original data are in phase.

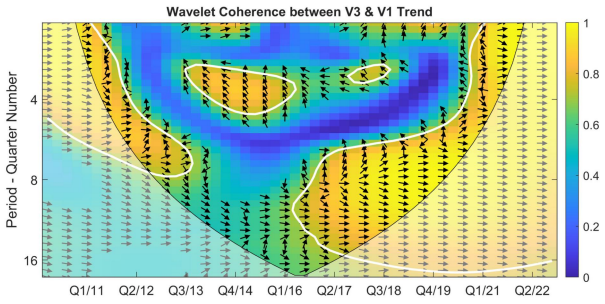


Figure 9: Wavelet coherence scalogram for V1 and V3 trend data

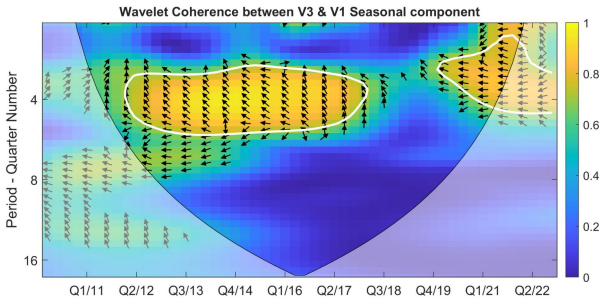


Figure 10: Wavelet coherence scalogram for V1 and V3 seasonal components

The correlational relationships for the pair of variables V1 and V2 (the number of apartments sold and theirs price, see Figures 11-13), when original data are considered, closely resemble the relationship between V3 and V1 (the number of apartments completed and theirs price). Differences in delays stem from the sequence of data processing, thus, these differences can be overlooked. The relationships for the seasonal components of these variables also show similarities. However, for the pair of variables V2 and V1, there exist slight correlations (about 0.6) within the biennial cycle in phase, post-Q2/17, yet they are not statistically significant. Conversely, the trends show significant differences. There is no strong correlation in phase starting from Q1/16 for the biennial cycle, nor in the annual cycle post-2019, as observed with variables V1 and V3. There is, however, a short period of correlation in 2013-2014 in anti-phase. Moreover, a quarter delay is observed around 2017-2018.

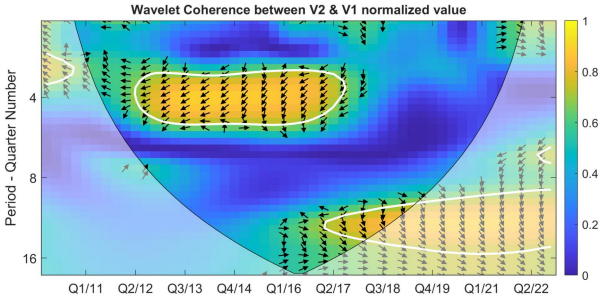


Figure 11: Wavelet coherence scalogram for V1 and V2 original data

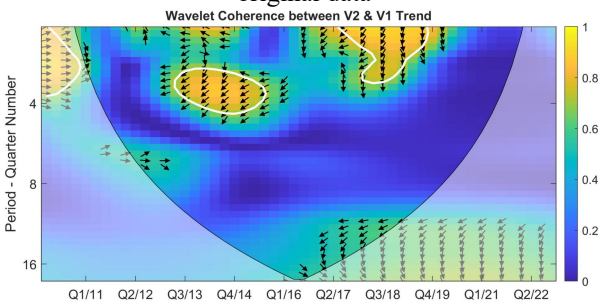


Figure 12: Wavelet coherence scalogram for V1 and V2 trend data

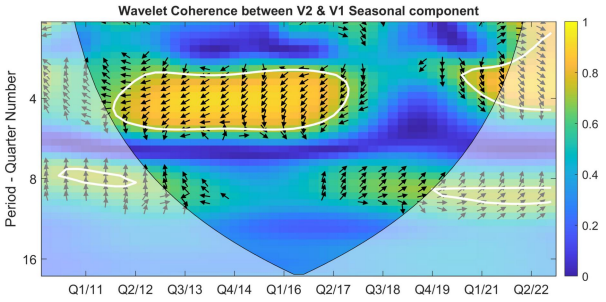


Figure 13: Wavelet coherence scalogram for V1 and V2 seasonal components

The relationships between the pair of variables V2 and V3 (the number of apartments completed and sold, see Figures 14-15) demonstrate a markedly different behavior. Wavelet analysis reveals strong annual correlations for both the original data and the seasonal components throughout the entire dataset. However, regarding trends, there is no annual correlation before the year 2013, but then biennial cycle correlations become apparent, and no correlations are observed post-Q3/2018. For this pair, the benefits of the wavelet method become particularly apparent, as Pearson,

Spearman, or Kendall correlations for the original data and cycles failed to identify any correlations, whereas the wavelet method enabled the detection of strong and statistically significant relationships shifted by a quarter period (observe downward arrows in Figures).

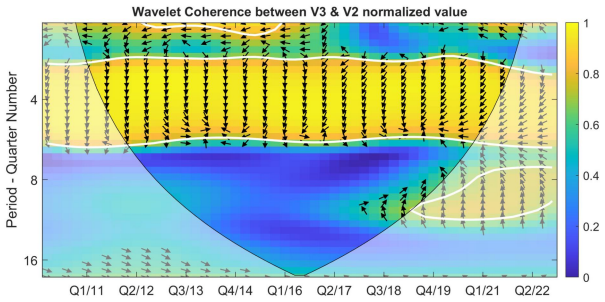


Figure 14: Wavelet coherence scalogram for V2 and V3 original data

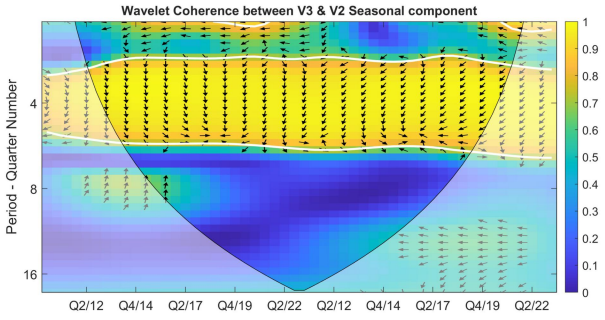


Figure 15: Wavelet coherence scalogram for V2 and V3 seasonal components

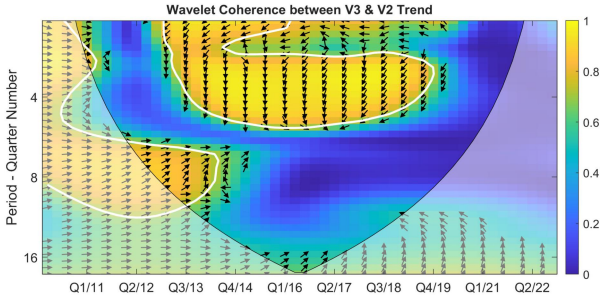


Figure 16: Wavelet coherence scalogram for V2 and V3 trend data

SUMMARY

Wavelet coherence analysis facilitates the identification of intervals during which a strong correlation exists between time series at distinct frequencies, thereby uncovering regular patterns or market trends. This approach allows for the monitoring of how fluctuations in the supply of new apartments impact their prices, an essential aspect of grasping the dynamics within the real estate market. Consequently, investors and analysts are equipped to more accurately interpret current market dynamics and forecast future pricing trends for apartments in the primary market.

The research undertaken has not only validated the utility of wavelet analysis in examining the interrelations among the price, availability, and sales volume of apartments in the primary market but has also highlighted the variability in correlation occurrences. Notably, these correlations were predominantly observed between 2012 and 2017, during which cyclical patterns were analyzed and found to be in anti-phase between the number of apartments made available for occupancy (V3) and the average price of apartments (V1). A notable cessation of correlation was detected from the end of 2017 through the beginning of 2019, particularly when price was a factor. Conversely, the relationship between the volume of apartments made available and those sold demonstrated consistent correlation throughout the entire period under review, trend exclusions notwithstanding.

REFERENCES

Ben Zeev, N., Pappa, E., and Vicondoa, A. (2017). Emerging economies business cycles: The role of commodity terms of trade news. *Journal of International Economics*, *108*, 368–376. https://doi.org/10.1016/j.jinteco.2017.07.008

Caunedo, J. (2020). Aggregate fluctuations and the industry structure of the US economy. *European Economic Review*, *129*, 103567. https://doi.org/10.1016/j.euroecorev.2020.103567

Chang, K. L. (2019). Are cyclical patterns of international housing markets interdependent? *Economic Modelling*, *88*, 14–24. https://doi.org/10.1016/j.econmod.2019.09.002

Devaney, S., and Xiao, Q. (2017). Cyclical co-movements of private real estate, public real estate and equity markets: A cross-continental spectrum. *Journal of Multinational Financial Management*, *42–43*, 132–151. https://doi.org/https://doi.org/10.1016/j.mulfin.2017.10.002

Fan, Y., Yang, Z., and Yavas, A. (2019). Understanding real estate price dynamics: The case of housing prices in five major cities of China. *Journal of Housing Economics*, *43*(April 2018), 37–55. https://doi.org/10.1016/j.jhe.2018.09.003

Fernández Muñoz, S., and Collado Cueto, L. (2017). What has happened in Spain? The real estate bubble, corruption and housing development: A view from the local level. *Geoforum*, *85*(March 2015), 206–213. https://doi.org/10.1016/j.geoforum.2017.08.002

Gabrovski, M., and Ortego-Marti, V. (2019). The cyclical behavior of the Beveridge Curve in the housing market. *Journal of Economic Theory*, *181*, 361–381. https://doi.org/10.1016/j.jet.2019.03.003

Ghysels, E., Plazzi, A., Valkanov, R., and Torous, W. (2013). Forecasting real estate prices. In *Handbook of Economic Forecasting* (Vol. 2, pp. 509–580). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53683-9.00009-8

Grinsted, A., Moore, J. C. Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series, Nonlin. Process. Geophys., 11, 561566.

Ionașcu, E., de La Paz, P. T., and Mironiuc, M. (2019). The Relationship between Housing Prices and Market Transparency. Evidence from the Metropolitan

European Markets. *Housing, Theory and Society*, *38*, 42–71.

Jones, C. A., and Trevillion, E. (2022). Macroeconomy and Real Estate Cycles. In C. A. Jones & E. Trevillion (Eds.), *Real Estate Investment : Theory and Practice* (pp. 43–61). Springer International Publishing. https://doi.org/10.1007/978-3-031-00968-6_3

Łaszek, J., & Olszewski, K. (2018). Regional Development of Residential and Commercial Real Estate in Poland and the Risk of Real Estate Cycles. *Barometr Regionalny. Analizy i Prognozy*, *1*, 41–51.

Łaszek, J., Olszewski, K., and Augustyniak, H. (2017). A model of housing demand - an analysis from the point of view of owners, owners-investors and investors. *Kwartalnik Nauk o Przedsiębiorstwie*, *43*, 69–77. https://doi.org/10.5604/01.3001.0010.4681

Łaszek, J., Olszewski, K., and Waszczuk, J. (2016). Monopolistic Competition and Price Discrimination as a Development Company Strategy in the Primary Housing Market. *Critical Housing Analysis*, *3*, 1. https://doi.org/10.13060/23362839.2016.3.2.286

Mach, Ł. (2019). Measuring and assessing the impact of the global economic crisis on European real property market. *Journal of Business Economics and Management*, *20*(6), 1189–1209. https://doi.org/10.3846/jbem.2019.11234

Mach, Ł., and Rącka, I. (2018). An analysis of the impact and influence of the global economic crisis on the housing market in European post-communist countries. *Proceedings of the 32nd International Business Information Management Association Conference (IBIMA)*, 2573–2584.

Mandler, M., and Scharnagl, M. (2022). Financial cycles across G7 economies: A view from wavelet analysis. *The Journal of Economic Asymmetries*, *26*, e00277. https://doi.org/10.1016/j.jeca.2022.e00277

Pyhrr, S., Roulac, S., and Born, W. (1999). Real Estate Cycles and Their Strategic Implications for Investors and Portfolio Managers in the Global Economy. *Journal of Real Estate Research* , *18*(1), 7–68.

Tomal, M. (2019). The impact of macro factors on apartment prices in Polish counties: A two-stage quantile spatial regression approach. *Real Estate Management and Valuation*, *27*(4), 1–14. https://doi.org/10.2478/remav-2019-0031

Tsai, I. C. (2019). Dynamic price–volume causality in the American housing market: A signal of market conditions. *North American Journal of Economics and Finance*, *48*, 385–400. https://doi.org/10.1016/j.najef.2019.03.010

Wang, B. (2021). How Does COVID-19 Affect House Prices? A Cross-City Analysis. *Journal of Risk and Financial Management*, *14*(2). https://doi.org/10.3390/jrfm14020047

Disclaimer: The paper presents the personal opinions of the authors and does not necessarily reflect the official position of the Narodowy Bank Polski.

AUTHOR BIOGRAPHIES

DARIA WOTZKA received M.Sc. degree in Computer Science from the Technische Universität Berlin, Germany, the Ph.D. and habilitation degree in Electrical Engineering from the Opole University of Technology, Poland. She is a lecturer and research fellow at the Opole University of Technology, Poland. Her research interests include data mining, modeling, and simulation. Her e-mail address is d.wotzka@po.edu.pl.

ŁUKASZ MACH received Ph.D. and habilitation degree in Economics and Finance from the Warsaw School of Economics, Poland. He is a Professor, researcher, and lecturer at the Faculty of Economics and Management at the Opole University of Technology, Poland. His research interests include quantitative methods in economic research, time series analysis, modeling of the residential real estate market, and the analysis of its cyclicality. His e-mail address is lmach@uni.opole.pl

PAWEŁ FRĄCZ is a full Professor, researcher, and lecturer at the Faculty of Economics at the University of Opole, Poland. His research interests include methods of time series analysis, issues in the area of modeling real estate markets, and applications of signal analysis methods in the time and frequency domain in economic sciences. His e-mail address is p.fracz@gmail.com.

MARZENA STEC received M.Sc. Degrees from University of Opole, Poland in the field of social sciences. She is an employee of the Narodowy Bank Polski. Her area of research interest is related to the real estate market. Her e-mail address is 7m.stec@gmail.com.

BARTOSZ CHORKOWY, M.Sc., Ph.D., earned his Master's degree in Economics and a Doctoral degree in Economics from the University of Opole. He is also a graduate of postgraduate studies in Information Systems Management at the Wrocław University of Science and Technology. Presently, he holds the position of Assistant Professor at the Institute of Economics and Finance at the University of Opole. His research interests encompass the determinants and effectiveness of the Polish pension system, analysis of the investment fund market, application of technical analysis methods for mitigating investment risk in the capital market, and the employment of quantitative methods in forecasting financial instrument prices. His contact email is bchorkowy@uni.opole.pl.

IRENEUSZ DĄBROWSKI is a Polish economist and lawyer, holder of a habilitation degree in economic sciences, and associate professor at the Warsaw School of Economics (Szkoła Główna Handlowa) in Warsaw, Poland. He serves as a professor, researcher, and lecturer at the Collegium of Management and Finance, within the Department of Applied Economics at the Warsaw School of Economics. His research interests are focused on the stability of systems, general equilibrium, and evolutionary economics. His e-mail address is ireneusz.dabrowski@sgh.waw.pl.

Implementierung eines Treibers zur Anbindung von Mikrocontroller-basierten Maschinen über OPC-UA an das Konfigurations- und Inbetriebnahme-Tool von SSI Schäfer

Julian Malovanij, B.Sc.

Product Development

SSI Schäfer IT Solutions GmbH
Friedenstraße 30
93053 Regensburg
E-Mail: julian.malovanij@ssi-schaefer.com

Professor Dr. Frank Herrmann

Innovationszentrum für Produktionslogistik und Fabrikplanung

Ostbayerische Technische Hochschule Regensburg
Galgenbergstraße 32
93053 Regensburg
E-Mail: frank.herrmann@oth-regensburg.de

ABSTRACT

Das Thema „Implementierung eines Treibers zur Anbindung von Mikrocontroller-basierten Maschinen über OPC-UA an das Konfigurations- und Inbetriebnahme-Tool von SSI Schäfer“ soll zu einem Gerätetreiber für das Konfigurations- und Inbetriebnahme-Tool CPM führen, welcher den Anschluss von neuen Maschinen an ebendieses ermöglicht.

Aktuell befinden sich diese noch in Entwicklung, jedoch sollen sie, von den Vorteilen des CPM-Tools profitieren, welches unter anderem den Inbetriebnahmeprozess einfacher und schneller gestalten kann.

Ziel ist somit die Anbindung dieser Maschinen durch die Spezifikation einer Schnittstelle und der Implementierung dieser auf Basis der Schnittstellen des CPM-Tools, in welches sich der Gerätetreiber nahtlos einfügen soll.

SCHLÜSSELWÖRTER

Gerätetreiber, Mikrokontroller-basierte Maschinen, SPS, OPC-UA

EINLEITUNG

Dieses Projekt befasst sich mit der Implementierung eines Gerätetreibers zur Anbindung von Maschinen mit einer neuen, eigenentwickelten SPS, dem SSI Controller eXtensible (SCX), an das Konfigurations- und Inbetriebnahme-Tool des Unternehmens SSI Schäfer IT Solutions GmbH. Dieses ist das CPM-Tool, wobei CPM für Configuration, Parametrization und Maintenance steht.

Weil das CPM bisher nur für Siemens S7 Steuerungen ausgelegt ist und die OPC-UA-Verbindungsaufrufe in der Geschäftslogik implementiert sind, ist ein einfacher Anschluss der SCX nicht möglich. Um diese Abhängigkeit von der S7-Schnittstelle und OPC-UA als Kommunikationsprotokoll zu entfernen, wurde das CPM-Tool um eine Treiberschicht erweitert, welche die Geräte für die Geschäftslogik abstrahiert. Dementsprechend ist ein Gerätetreiber über diese Schnittstelle für die neuen SCX-Steuerungen implementiert worden.

IST-SITUATION

CPM-Tool

Wie in Abbildung 1 zu sehen, basiert das CPM-Tool auf einer Client-Server Architektur, wobei der Client in Angular als Webanwendung implementiert ist und via REST sowie Websockets mit dem CPM-Server kommuniziert.

Im Rahmen der Erweiterung um eine Treiberschicht sollen alle Geräte darüber abgewickelt werden, um von spezifischen Kommunikationsschnittstellen unabhängig zu sein. Dementsprechend müssen für SCX und Mock-Implementierung jeweils ein Gerätetreiber erstellt werden.

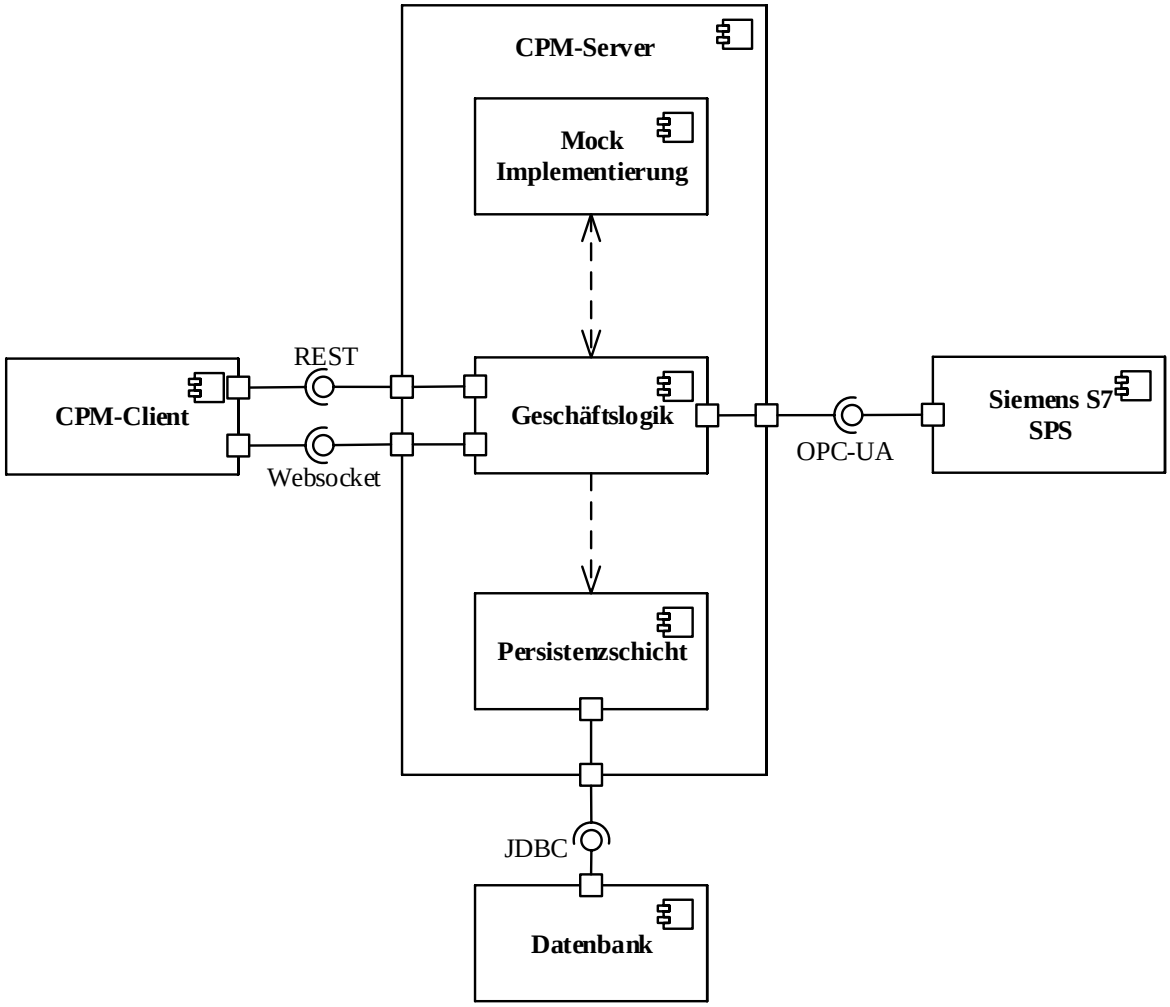


Abbildung 1: Komponentendiagramm CPM Gesamtübersicht (Ist-Situation vor Treiberschicht)

Element-Segment-Block

Das CPM-Tool legt für jede Steuerung bei der Durchsuchung des Baumes Elemente an. Diese sind die Funktionspunkte einer SPS. Da jeder Funktionspunkt aus mehreren Funktionsbausteinen bestehen kann, wie bspw. einem Standardbaustein und einem Baustein für projektspezifische Erweiterungen, sind Segmente eingeführt worden. Diese bilden die Bausteine ab und sind nur eine organisatorische Aufteilung, welche die Daten des Elementes enthält. Alle weiteren SPS-Komponenten werden als Blöcke abgebildet. Dies sind bspw. Fördertechnikelemente wie Förderer oder Plätze.

SCHNITTSTELLENSPEZIFIKATION

Da zum Zeitpunkt der Bearbeitung noch keine konkrete Schnittstellenspezifikation zur Kommunikation zwischen SCX und CPM vorhanden war, ist diese im Rahmen dieses Projekts entstanden. Als Basis dient hierfür die Siemens S7 Schnittstelle. Sie ist durch zahlreiche Abstimmungen mit den SCX- und CPM-Entwicklern entstanden und stellt, neben der Treiberschichtschnittstelle, die Basis für die Implementierung dar.

Device Discovery

Die Device Discovery hat die Aufgabe, einem zu verbindenden Gerät automatisch und ohne weitere Eingaben durch den Benutzer den passenden Gerätetreiber zuzuordnen. Hierfür gibt es zwei relevante Schnittstellen.

Die erste Schnittstelle ist zwischen der Geschäftslogik des CPMs und dem Gerätetreiber. Hierbei müssen Methoden zum Abruf der Treibereigenschaften, sowie eine Methode zur Durchführung der Discovery, angeboten werden. Die Geschäftslogik wiederum bietet eine Methode zur Registrierung des Gerätetreibers an.

Die zweite Schnittstelle ist zwischen dem Gerätetreiber und der SCX-Steuerung. Hierbei stellt das SCX seine Eigenschaften, wie bspw. die Interfaceversion, über einen OPC-UA-Knoten auf dem OPC-UA-Server der Steuerung bereit.

Session

Das CPM ist das führende System für die Datenspeicherung für alle verbundenen SPS-Steuerungen. Aus diesem Grund darf jeweils nur ein CPM-Tool zeitgleich mit einer Steuerung verbunden

sein. Dementsprechend muss das SCX die Attribute „Hostname“ und „Last Hostname“ anbieten, damit die Geschäftslogik des CPMs sich nur verbindet, wenn diese entweder leer oder mit dem gleichen CPM-Hostnamen belegt sind.

Parametrierung

Die Parametrierung ist der Kern des CPMs. Hier können die Parameter der SPS im laufenden Betrieb verändert werden. Damit das CPM-Tool erkennen kann, welche Parameter relevant sind, wird auch hierfür eine Konvention benötigt.

Control Sync
Der Control Sync ist die Durchsuchung des Informationsmodells des OPC-UA-Servers des SCX nach relevanten Parametern. Dabei wird in Elemente, welche die Funktionspunkte einer SPS abbilden, Segmente, welche die Funktionsbausteine abbilden, und Blöcke, welche alle untergeordneten Knoten darstellen, gegliedert. Bei der Durchsuchung werden Elemente gesucht. Im Anschluss werden die gefundenen Elemente vollständig aus der Steuerung ausgelesen. Die Suche bricht ab, sobald ein Knoten keinen Info-Knoten mehr hat. Die Sub-Knoten von diesem werden dann nicht mehr beachtet und der nächste Kind-Knoten des jeweiligen Eltern-Knotens wird durchsucht. Im Info-Knoten sind die relevanten Attribute enthalten. Diese spezifizieren über das Vorhandensein einer NodeID, dass ein Knoten ein Segment eines Elementes ist und enthalten auch den mechanischen Namen, welcher im CPM verwendet wird.

Die Parameter sind als Knoten mit eigenem Wert sowie den Attributen „DetailGroup“, Permission und ID angelegt. Diese Parameter befinden sich jeweils in einem ParameterSet-Knoten, welcher selbst ein Kind-Knoten eines Blocks ist. Im Rahmen des Control Syncs werden aufgefundene Parameter in die Struktur übernommen, jedoch werden die Parameterwerte nicht gespeichert. Die Werte für das CPM kommen bspw. aus einem Import von Standardwerten beim Projekt-Start.

Control Import
Der Control Import entspricht dem Control Sync mit dem Zusatz, dass bei diesem die aktuellen Parameterwerte der SPS in die CPM-Datenbank übernommen werden.

Anforderungen zwischen CPM und SCX
Damit das CPM über Änderungen der Element- und Parameterstrukturen informiert wird, stellt das SCX die Timestamps „ElementModDate“ und „ParameterModDate“ zur Verfügung. Beide lösen einen Control Sync aus.

Zusätzlich hat jedes Element einen Node-Knoten, welcher über die NodeID erreichbar ist. Dieser enthält drei Request-Done-Laufnummernsysteme. Eines zur CPM-seitigen Anforderung einer Parameterübernahme, eines für die SCX-seitige Auslösung eines Control Syncs für ein einzelnes Element, sowie eines für die SCX-seitige Anforderung aller Parameter, welche das CPM dann auf die SCX schreibt.

IO-View und IO-Check

Der IO-View zeigt den Status der Ein- und Ausgänge der SPS in Form einer „Lampe“ an. Hierfür werden vom SCX in den Ordnern „Input“ und „Output“ jeweils Boolean-Variablen bereitgestellt. Für den IO-Check bietet die Steuerung eine Request-Boolean-Variable sowie einen IO-Check-State an. Da die SCX im Falle des IO-Checks in einen Simulationsmodus geht, müssen die Ausgänge auf einen „SimulatedState“ Knoten anstelle des Wurzelknotens gesetzt werden.

Protokollierung

Bei der Protokollierung werden die Log-Einträge von der Steuerung ausgelesen und auf dem Dateisystem des CPMs gespeichert. Um dies zu ermöglichen, stellt das SCX für jedes Protokoll einen Knoten nach Benennungskonvention „Logging*“ bereit, welcher die Eigenschaften des Protokolls enthält.

Device Capabilities

Das CPM-Tool unterstützt verschiedene Funktionalitäten, welche dynamisch im Betrieb ein- und ausgeschaltet werden können. Hierfür bietet das SCX eine Capabilities-Struktur an, welche je eine Boolean-Variable für jede unterstützte CPM-Funktion enthält. Eine Übersicht darüber ist in Tabelle 1 zu sehen.

Capability:	Vom SCX unterstützt:
Parametrierung	Ja
Handbetrieb	Nein
Operationen	Nein
IO-Check	Ja
Protokollierung	Ja
Metriken	Ja
Datenmanipulation	Nein

Tabelle 1: Im SCX unterstützte Capabilities

IMPLEMENTIERUNG

Die Implementierung eines SCX-Gerätetreibers ist das eigentliche Ziel des Projekts.

Mock-Gerätetreiber

Der CPM-Server benutzt, im Rahmen von automatischen Tests, Mock-Implementierungen für OPC-UA-Steuerungen. Da diese vor der Durchführung des Projekts direkt in die Geschäftslogik eingebunden waren, wurden sie in einen eigenen Gerätetreiber migriert. Hierbei wurde die Eigenschaft, dass alle Mock-Implementierungen das OPC-UA-Verbindungs-Interface implementieren, ausgenutzt. Dabei wurde ein S7-Gerätetreiber derartig modifiziert, sodass anstelle von OPC-UA-Verbindungen jeweils eine passende Instanz einer Mock-Implementierung verwendet wird. Diese

Instanzen werden dabei einmalig angelegt und im Arbeitsspeicher gehalten, um deren eigene, transiente Datenhaltung zu ermöglichen.

Um nur einen Gerätetreiber für die fünf Mock-Implementierungen umsetzen zu müssen, erfolgt die Instanzverwaltung mit Hilfe eines Enums. Dieses enthält die URL, den Steuerungsnamen und eine Java Reflection Referenz auf die jeweilige Implementierungsklasse. Zudem enthält das Enum eine statische Methode, welche anhand der URL die korrekte Implementierung zurückgibt. Dieser Enum-Wert ist dann der Key für die Map, welche die Instanzen der Mock-Implementierungen enthält. Über eine weitere, statische Methode kann diese Logik gegenüber dem Gerätetreiber versteckt werden, sodass nur noch ein Aufruf mit der URL als Übergabeparameter notwendig ist um die Implementierung, welche das OPC-UA-Verbindungsinterface implementiert, zu erhalten. Das Verhalten entspricht daher einem OPC-UA-Verbindungsaufruf.

SCX-Gerätetreiber

Das SCX besitzt an einigen Stellen numerische Knoten-IDs anstelle der geforderten Knoten-IDs, welche einen Pfad durch den SCX-Baum beschreiben. Aus diesem Grund sind einige Optimierungen des S7-Treibers nicht möglich und es musste eine Helfer-Klasse zum Handling der potentiell numerischen Knoten-IDs angelegt werden, da das CPM selbst alle Knoten-IDs als Strings speichert.

Der Gerätetreiber des SCX hat, wie in Abbildung 2 ersichtlich, für jede implementierte CPM-Funktion einen eigenen, unabhängigen Service. Dies ermöglicht einen modularen Aufbau, da über die Device Capabilities das Vorhandensein oder Fehlen eines Service für eine CPM-Funktion indiziert werden kann.

Device Discovery

Die Device Discovery ist der Einstiegspunkt jeder Gerätetreiber-Implementierung, da sie dafür verantwortlich ist, ihre Zuständigkeit für die Steuerung zu erkennen und in diesem Fall eine Geräteinstanz an die Geschäftslogik des CPMs zurückzuliefern.

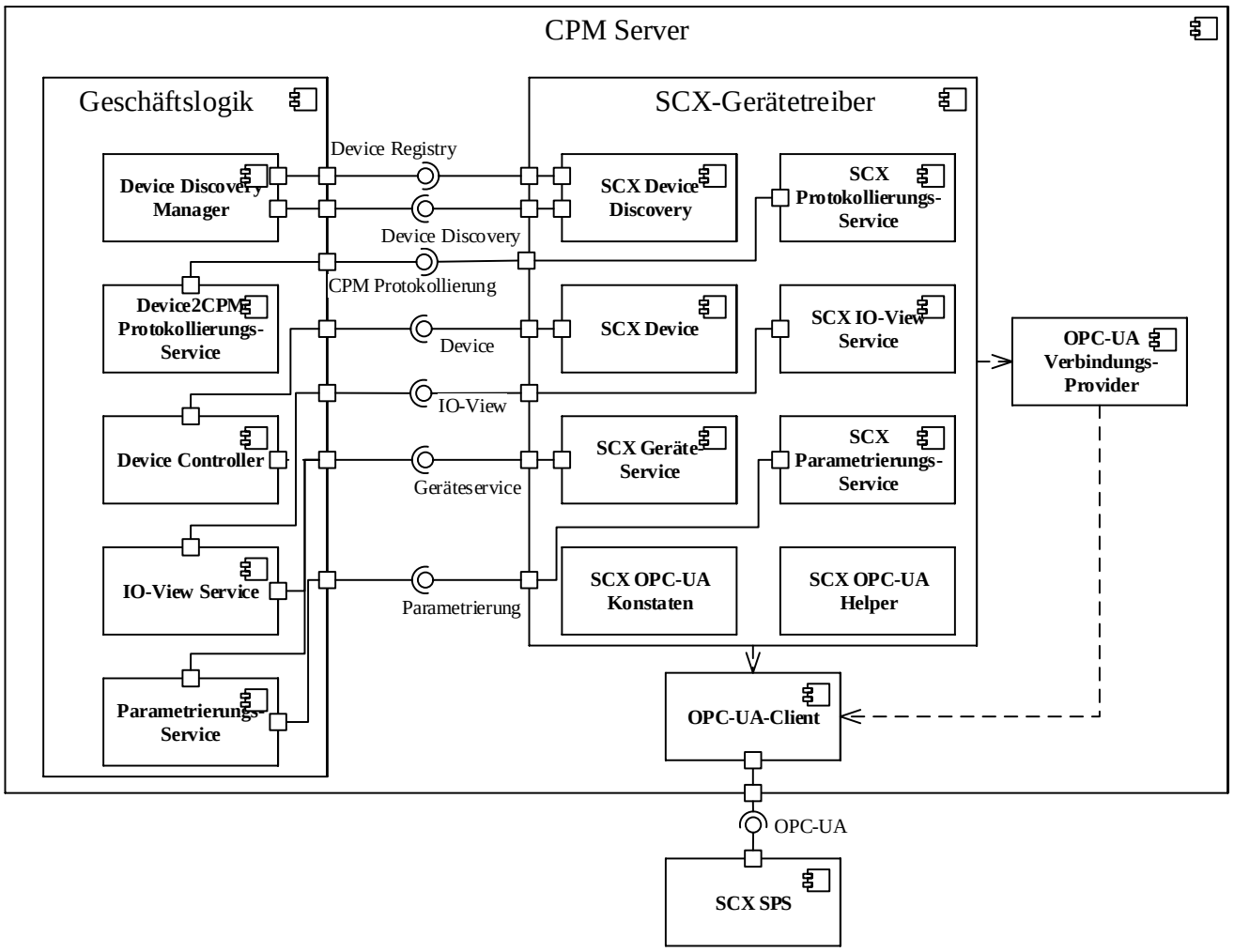


Abbildung 2: Komponentendiagramm des SCX-Gerätetreibers im Systemumfeld

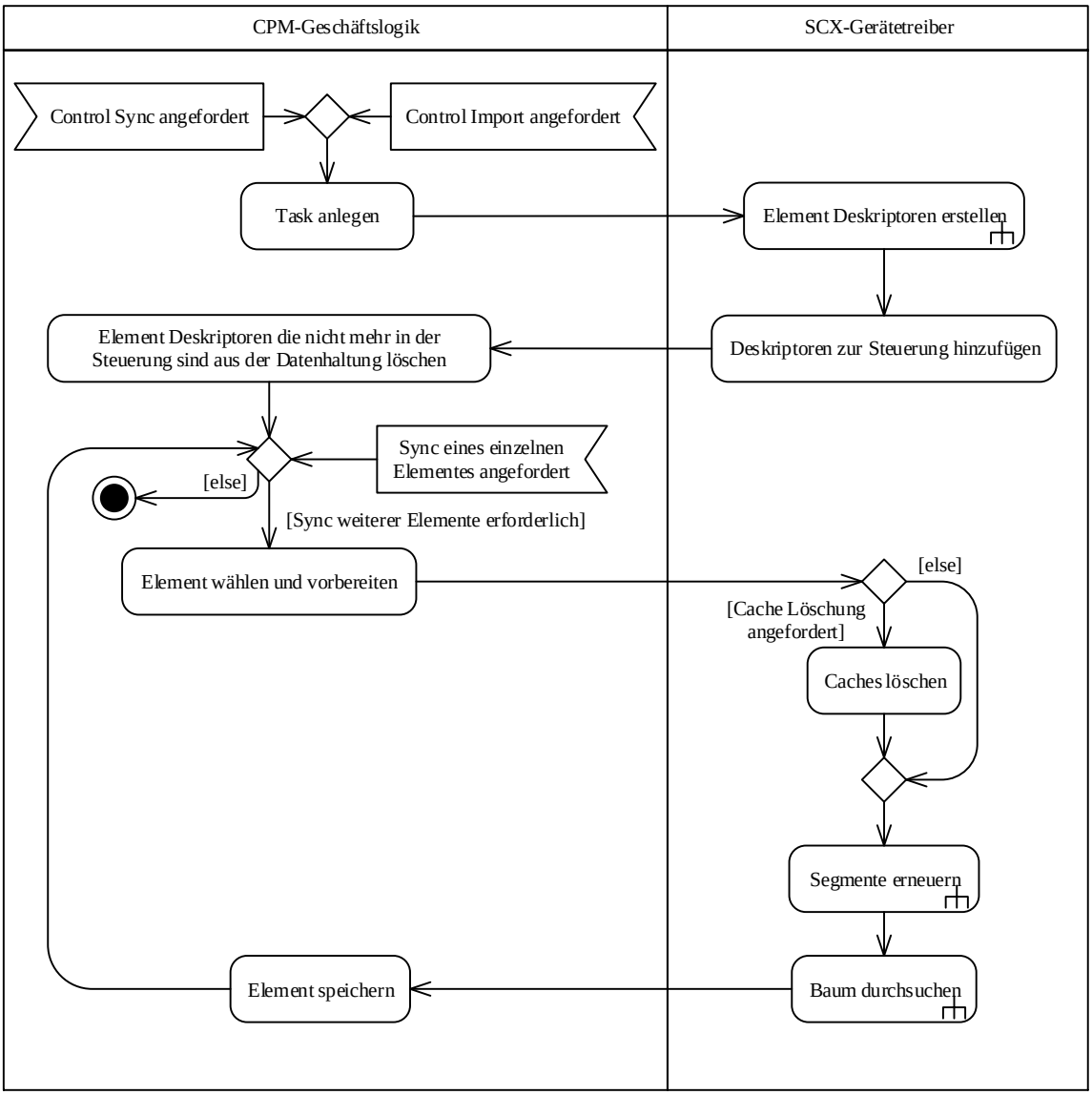


Abbildung 3: Aktivitätsdiagramm als Übersicht für den Ablauf eines Control Syncs/Imports

Um vom CPM-Tool als Gerätetreiber erkannt zu werden, muss sich die Device Discovery, neben der Implementierung des entsprechenden Interfaces, beim Device Discovery Manager der Geschäftslogik des CPMs registrieren.

Mock-Implementierung
Die Erkennung der Zuständigkeit erfolgt auf Basis der übergebenen URL.

SCX-Implementierung
Im SCX-Gerätetreiber ist die Erkennung der Zuständigkeit komplizierter als im Mock-Gerätetreiber. Hier ist ausschlaggebend, dass die DeviceHardwarePlattform, der DeviceType und die InterfaceVersion vom implementierten Treiber mit denen der Steuerung übereinstimmen.

Die Device Discovery verbindet sich hierfür mittels einer OPC-UA-Verbindung direkt mit dem SCX und versucht die benötigten Werte auszulesen. Schlägt dies fehl, wird null zurückgegeben, oder stimmen die gelesenen Werte nicht mit denen des Treibers überein, so ist dieser Gerätetreiber nicht zuständig. Andernfalls wird eine neue Geräteinstanz an die Geschäftslogik zurückgegeben.

Im Rahmen dieser Rückgabe werden weitere benötigte Informationen aus dem SCX ausgelesen. Diese sind die Device Capabilities, der Gerätename, -typ und -version.

Session

Gerät (DeviceImpl)
Die Geräteklasse ist für den Verbindungsauf- und -abbau zuständig. Dies wird größtenteils an einen OPC-UA-Connection-Provider weitergegeben, jedoch müssen alle OPC-UA-spezifischen Implementierungen wie Verbindungs-Listener in CPM-Äquivalente gekapselt werden.

Diese Klasse enthält außerdem Referenzen auf alle Services des Treibers und stellt diese dem CPM-Tool zur Verfügung.

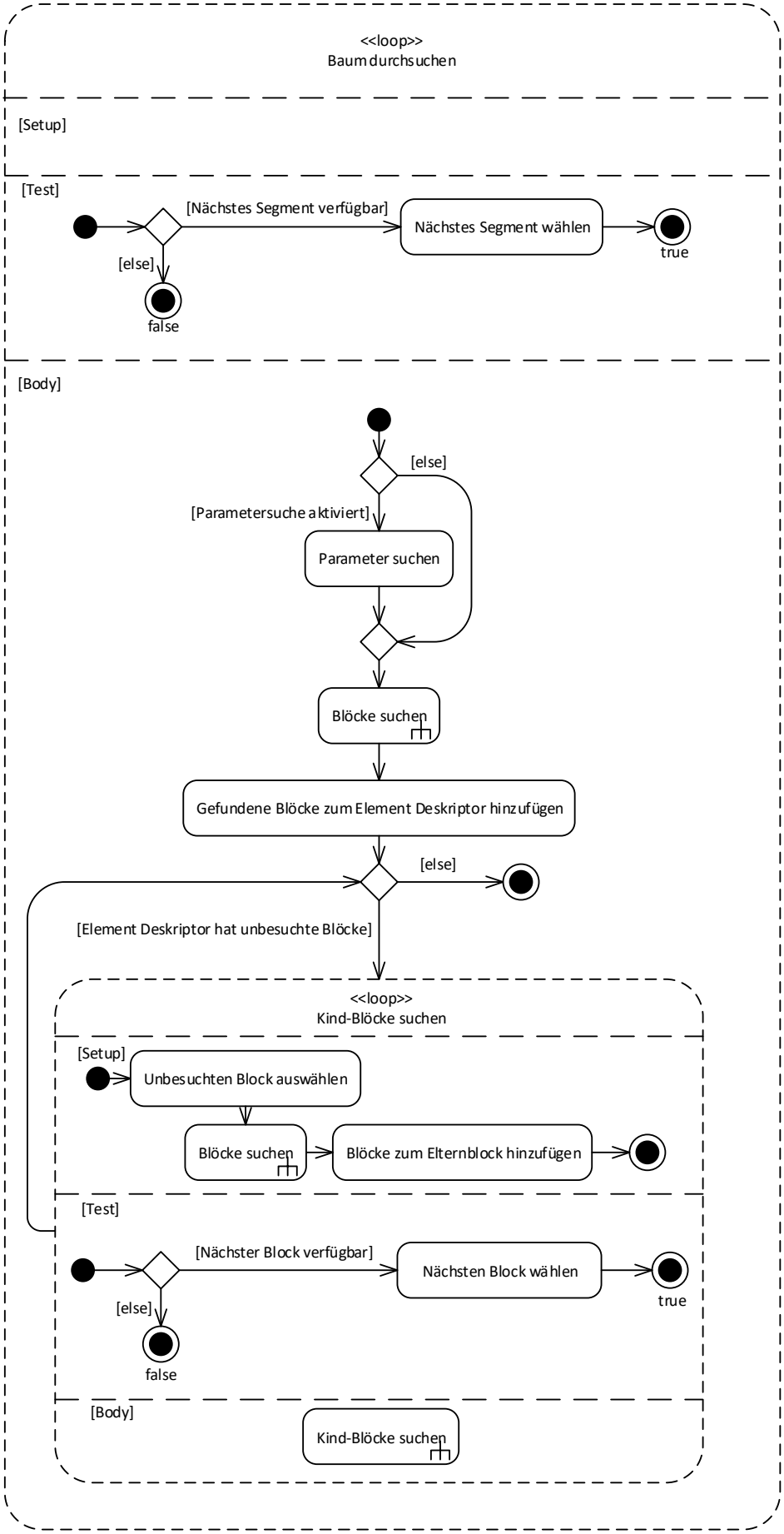


Abbildung 4: Subaktivität Baum durchsuchen

Geräteservice (DeviceServiceImpl)
Der Geräteservice bietet das Anlegen von Monitored Items für übergebene Notification-Listener an. Dies erfolgt im Rahmen der Trennung der Geschäftslogik von Kommunikationsprotokollen. Außerdem stellt der Geräteservice den Status der Steuerung sowie die Lese- und Schreibfunktionen bereit.

Parametrierung

Die Parametrierung ist die wichtigste Funktion des CPM-Tools.

Control Sync
Wie in Abbildung 3 zu sehen, erfolgt der Control Sync in zwei Schritten. . Im ersten werden alle Knoten mit einer NodeID in ihrem Info-Knoten gesucht. Dabei wird die rekursive Suche nach dem Fund des ersten entsprechenden Knoten sowie bei Fehlen eines Info-Knotens abgebrochen, ansonsten werden alle weiteren Subknoten durchsucht. Diese Knoten werden als Segment bezeichnet und besitzen neben dem mechanischen Namen und der NodeID optional einen String Control-Block. Diese Segmente werden dann auf Basis ihres mechanischen Namens und der NodeID zu Elementen zusammengefasst. Eine Liste dieser Elemente wird dann an die Geschäftslogik des CPM zurückgegeben.

Wie ebenfalls aus Abbildung 3 hervorgeht, erfolgt der zweite Schritt, nach dem Aktualisieren der CPM-internen gespeicherten Elemente, für jedes Element einzeln. Dementsprechend werden die Elemente von der Geschäftslogik an den Gerätetreiber übergeben. Dies ist nötig, da ansonsten für die Anforderung eines Control Syncs eines einzelnen Elementes eine weitere Implementierung notwendig wäre. In diesem zweiten Schritt werden die Segmente des Elements nochmals von der Steuerung gelesen. Danach wird der Baum rekursiv nach Blöcken durchsucht. Diese Suche ist in Abbildung 4 zu sehen. Die Rekursion bricht ab, sobald ein Knoten keinen Info-Knoten enthält. In einem solchen Fall wird der nächste Kind-Knoten des Eltern-Knotens betrachtet. Dabei wird ein Cache verwendet, um mehrfaches Lesen eines Knotens zu verhindern. Diese in Abbildung 5 dargestellte Blocksuche wird für alle Kind-Knoten des aktuellen Blocks rekursiv durchgeführt, sodass alle Blöcke im Baum mit einem Pfad, in welchem jeder Knoten einen Info-Knoten hat, gefunden werden können. Im Rahmen der Suche nach Blöcken wird für jeden Block auch nach Parametern gesucht. Diese sind im ParameterSet-Knoten enthalten und können von dort ausgelesen und in eine entsprechende CPM-Repräsentation umgeformt werden. Da auch hier numerische Knoten-IDs zum Einsatz kommen, müssen die ParameterSet-Knoten durch einen OPC-UA-Browse-Aufruf durchsucht werden. Um hier ebenfalls eine Beschleunigung zu erzielen, wird ein weiterer Cache verwendet.
Für jeden Parameter in einem ParameterSet-Knoten wird daraufhin zunächst der Datentyp ausgelesen. Ist dieser nicht vorhanden oder unbekannt, so wird angenommen, dass dies kein Parameter ist. Andernfalls wird der Parameterwert gelesen und im Anschluss ein Browse des Parameterknotens durchgeführt, um die benötigten Variablen DetailGroup, Permission und ID sowie ggf. das Init-Flag zu finden und daraufhin auszulesen. Der DetailGroup-String beinhaltet dabei den Detail- sowie den Group-String jeweils getrennt durch ein „@"-Zeichen. Die entsprechende Auflösung findet im Rahmen der Parametererstellung direkt im Anschluss an das Auslesen der Werte statt.

Falls ein Block keinen Namen im Info-Knoten hat, so ist dieser für das CPM „unsichtbar", sodass all seine Parameter und Kind-Blöcke an den Eltern-Block angehangen werden.

Control Import
Der Control Import ist grundsätzlich dasselbe wie der Control Sync, jedoch werden in diesem Modus auch die ausgelesenen Parameterwerte von der Geschäftslogik des CPM-Tools in die interne Datenbank übernommen. Da diese auch im Zuge des Control Syncs ausgelesen und übergeben werden, sind somit für diesen Anwendungsfall keine weiteren Implementierungs-arbeiten von Seiten des SCX-Gerätetreibers notwendig.

Anforderungen zwischen SCX und CPM
Die Request-Laufnummer wird für eine Anforderung jeweils um eins erhöht, ebenso wird die Done-Laufnummer für eine Bestätigung jeweils um eins inkrementiert. Zu beachten sind Überläufe des uInt16, da im Gerätetreiber ein Int32 verwendet wird und das Verhalten somit anders ist.

Die Init-Anforderung des CPMs erfolgt nach dem Schreiben eines Parameters mit gesetztem Init-Flag. Nach der Übernahme der Parameterwerte bestätigt die SCX die Durchführung, dies wird vom CPM jedoch nicht ausgewertet.

Die Refresh- und Update-Anforderung erfolgen SCX-seitig und werden vom CPM über Monitored Items erkannt. Nach der Ausführung des Control Syncs bzw. des Schreibens aller Parameter für dieses Element bestätigt das CPM die Ausführung.

Die Timestamps „ElementModDate" und „ParameterModDate" enthalten den letzten Änderungszeitpunkt der jeweiligen Struktur. Sie werden auch bei einem Neustart des SCX gesetzt. Falls dieser Timestamp null ist oder nach dem im CPM gespeicherten Zeitpunkt liegt, wird ein Control Sync durchgeführt. Auch hier erfolgt die Erkennung über Monitored Items.

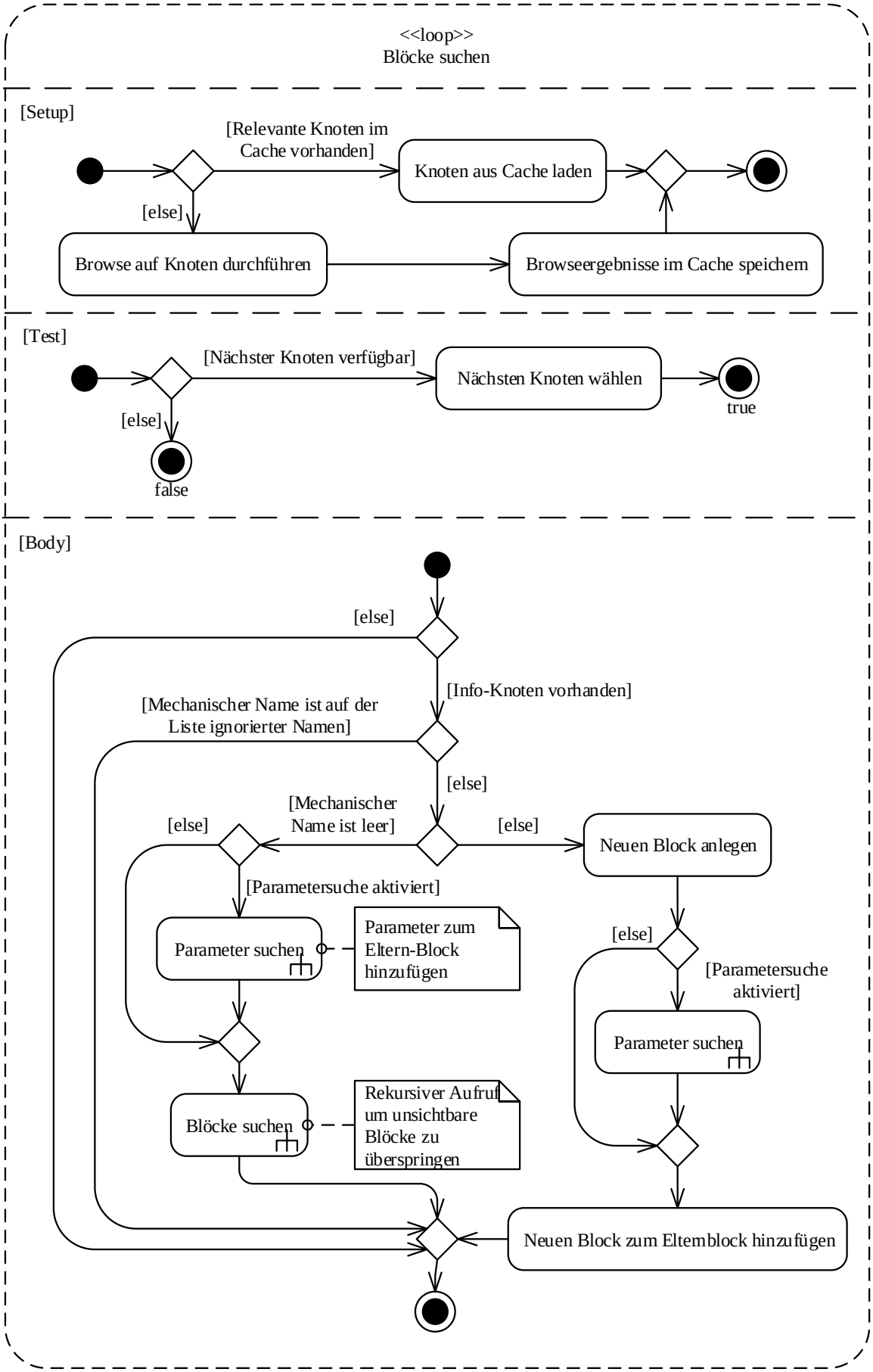


Abbildung 5: Subaktivität Blöcke suchen

IO-View

Der IO-View stellt die Ein- und Ausgänge des SCX als „Lampe“ im Frontend dar. Dafür müssen diese entweder über einen IO-Import oder einen PTC-Import eingelesen werden. Eine Statusänderung wird über Monitored Items erkannt.

IO-Import
Für den IO-Import werden die beiden Ordner „Input“ und „Output“ gebrowst. Hierbei werden alle Einträge vom Datentyp Boolean übernommen. Für Ausgänge gilt hierbei die Sonderregel, dass jeweils der „SimulatedState“-Knoten geprüft und als Ausgang registriert werden muss. Dieser ist durch eine Pfadkonkatination erreichbar.

IO-Check
Der IO-Check ist lediglich eine graphische Möglichkeit um Ein- und Ausgänge während der Inbetriebnahme zu überprüfen und deren Status zu ändern. Hierbei wird die Boolean-Variable „RequestIoCheck“ verwendet, welche während der gesamten Dauer vom CPM auf true gesetzt ist. Eine Änderung von false auf true erwirkt die SCX-seitige Initialisierung des IO-Checks und eine Änderung von true auf false bewirkt die SCX-seitige Beendigung des IO-Checks. Diese Statusänderungen kann das CPM aus der Variablen „IoCheckState“ lesen.

Der Status wird dabei durch dieselbe Methode zurückgemeldet, welche auch die Request-Variable setzt. Dies erfolgt dort im Rahmen der Prüfung, ob eine Anfrage eines IO-Checks nötig oder möglich ist und dient der beschleunigten Rückmeldung an den CPM-Client. Eine Rückmeldung einer Statusänderung erfolgt ebenfalls über ein Monitored Item.

Wie aus Abbildung 6 hervorgeht, kann ein IO-Check nur vom Zustand „Allowed“ aus gestartet werden. Der Zustand „Not-Allowed“ ist hierbei ein Trap-Zustand und alle anderen Zustände werden nach Anforderung durch das SCX gesteuert. Der Zustand „Active“ ist die SCX-Bezeichnung für den CPM-Zustand „Released“.

Protokollierung

Bei der Protokollierung wird die Speicherung des Protokolls von der Geschäftslogik des CPMs übernommen, während der Gerätetreiber für die Abholung der Log-Einträge verantwortlich ist.

Wie aus Abbildung 7 ersichtlich wird, wird bei einer Aktivierung der Protokollierung für eine Steuerung ein Browse durchgeführt, um alle Knoten zu erhalten, welche der Benennungskonvention „Logging*“ folgen. Diese werden als Protokolldefinitionen bezeichnet. Von diesen Knoten wird dann unter Ausnutzung der Pfad-Knoten-ID durch Pfadkonkatination die Knoten für den Datenbanknamen, den Protokollnamen, den Header und die Größe des Ringpuffers mit einem Aufruf gelesen.

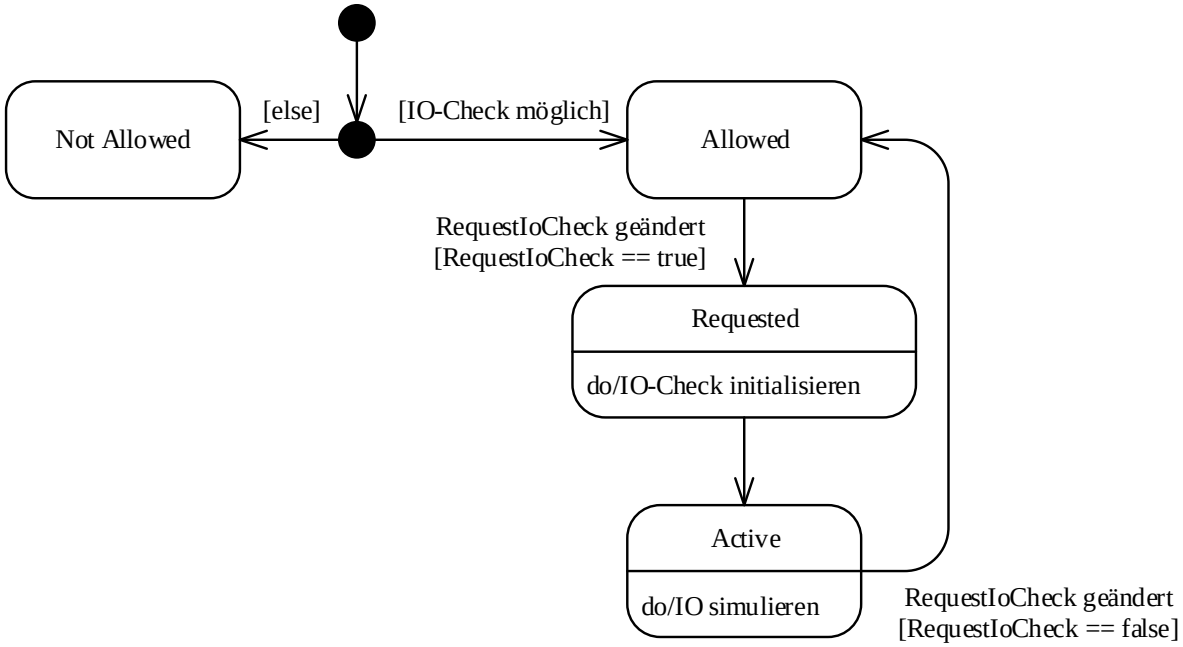


Abbildung 6: Zustandsdiagramm IO-Check

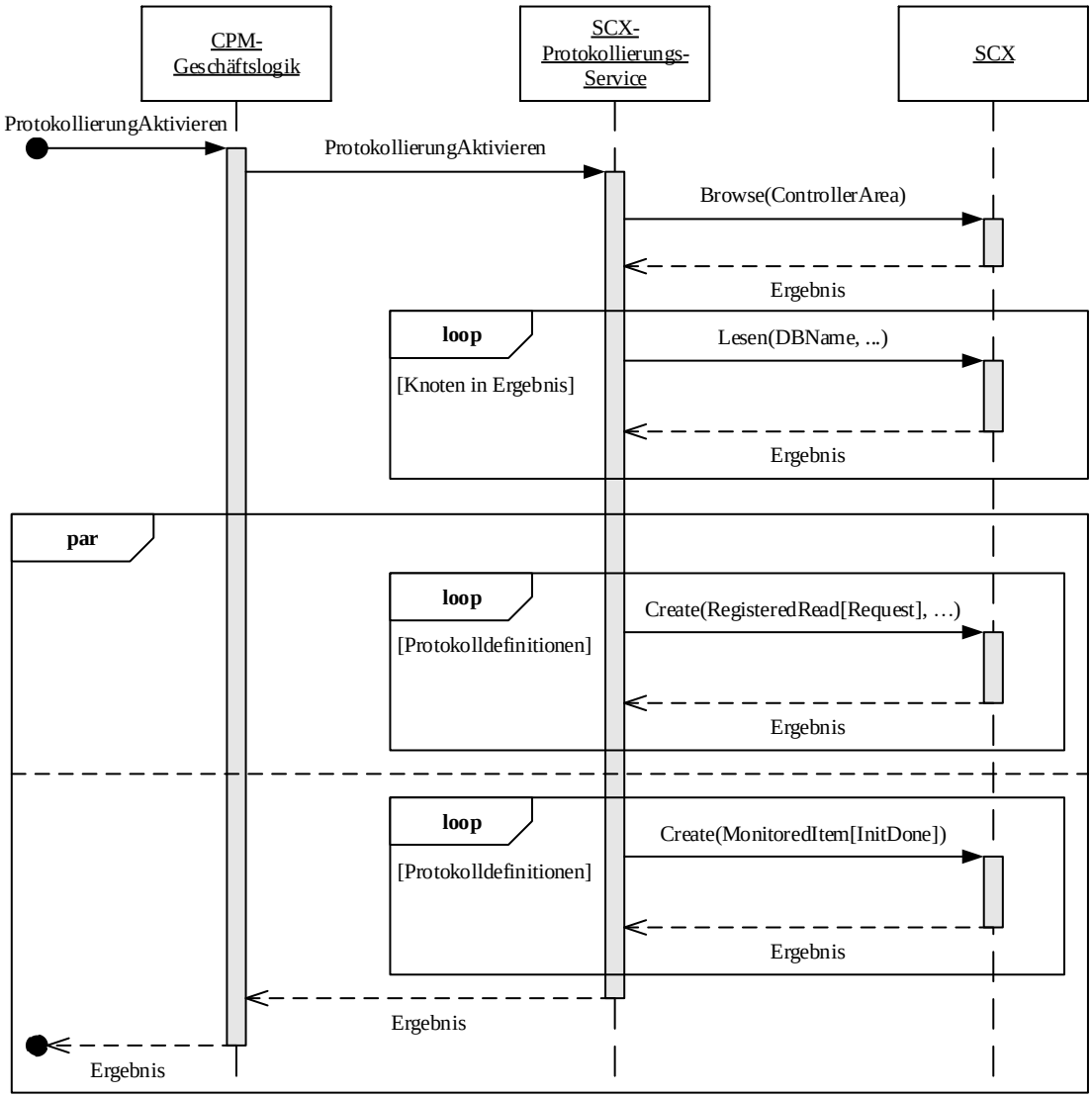


Abbildung 7: Sequenzdiagramm Protokollierungsaktivierung

Außerdem wird ein neuer Thread gestartet, welcher für jede Protokolldefinition das Request-Done-Laufnummernsystem, die Start- und Stoppindizes für den Ringpuffer, sowie die Eintragsknoten selbst als Registered Reads im SCX-OPC-UA-Server anlegt und somit veranlasst, dass diese im Arbeitsspeicher vorgehalten werden. Dies ist nötig, da bei einem Ringpuffer die Einträge schnellstmöglich abgeholt werden müssen, weil es ansonsten zu Verlusten aufgrund von Überschreibungen (Überlauf) kommen kann.

Die Abholung geschieht durch ein Monitored Item auf der Request-Laufnummer, sodass diese bei jeder Änderung ausgelöst wird. Wie in Abbildung 8 zu sehen, werden daraufhin, da das SCX einen Ringspeicher implementiert, die Knoten-IDs der abzuholenden Einträge erstellt. Nachdem die entsprechenden Log-Einträge gelesen wurden, werden sie an die Geschäftslogik übergeben. Es erfolgt eine Bestätigung über die Done-Laufnummer.

Das SCX aktiviert die Protokollierung je Protokoll durch einen Init-Flag. Treiberseitig wird dabei die Geschäftslogik mit dem Anlegen oder Abschließen einer Datei beauftragt und benötigte Monitored Items angelegt bzw. bei der Deaktivierung entfernt.

Device Capabilities

Die Device Capabilities legen die im Moment von der SCX unterstützten CPM-Funktionen fest. Da sie als Struktur angelegt sind, können sie in einem Lese-Aufruf ausgelesen werden. Für das CPM werden diese dann in ein Device Capabilities DTO übertragen. Dabei erfolgt zunächst für jede Capability die Prüfung, ob der Wert vorhanden, also ungleich null ist. Ist dies der Fall, so wird der Boolean-Wert übernommen, andernfalls wird false angenommen. Um auf Änderungen reagieren zu können wird ein Monitored Item auf die Struktur angelegt.

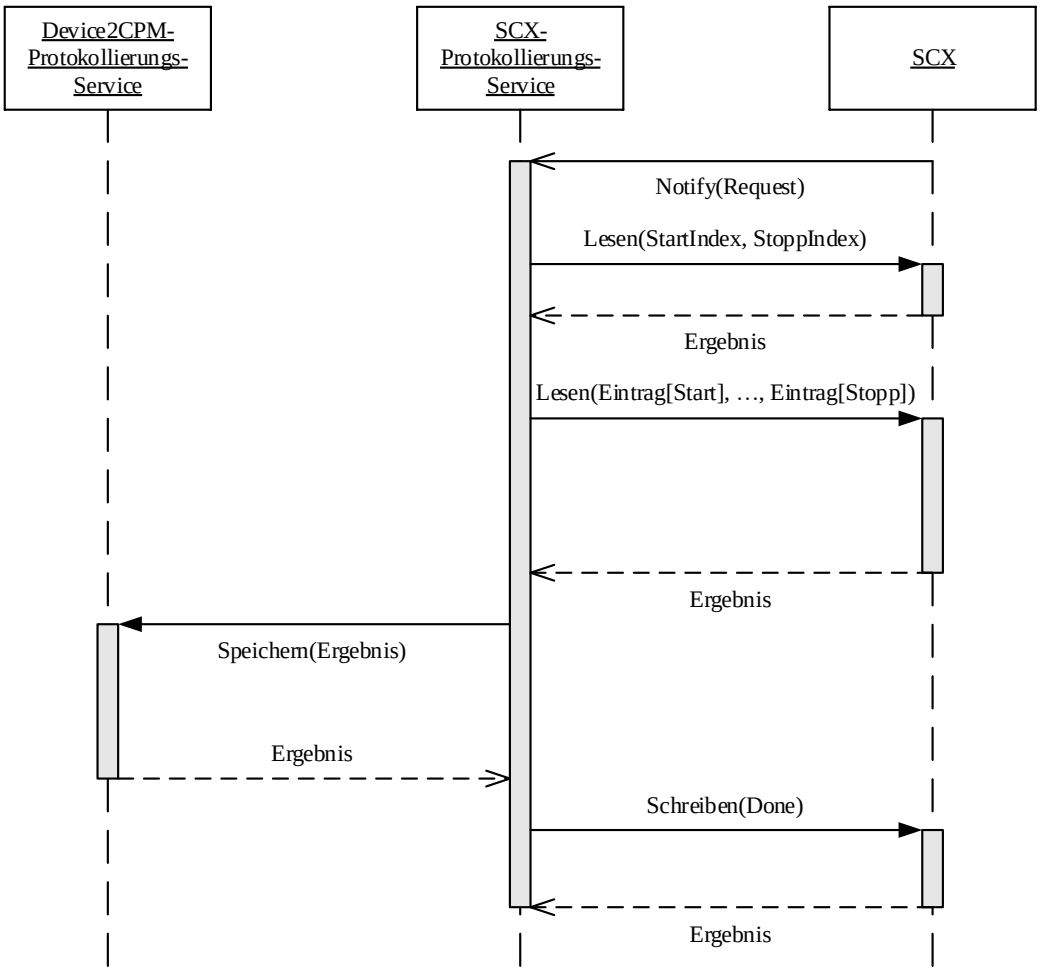


Abbildung 8: Sequenzdiagramm Protokolleintragsabholung

Bewertung der Lösung

Mock-Gerätetreiber
Diese Lösung erfüllt alle an sie gestellten Anforderungen. Zum einen sind die Mock-Implementierungen aus der Geschäftslogik entfernt worden, zum anderen sind alle OPC-UA-Aufrufe über die Treiberschicht und somit dem Mock-Gerätetreiber abstrahiert worden. Dadurch ist die Nutzung von anderen Kommunikationsprotokollen ermöglicht worden. Zudem ist durch die Package-Struktur des Treibers auch die Wiederverwendbarkeit bei Schnittstellenänderungen sichergestellt worden.

SCX-Gerätetreiber
Der SCX-Treiber erfüllt alle an ihn gestellten Anforderungen. Es ist die Schnittstelle der Treiberschicht vollständig implementiert worden, ebenso wie die Schnittstelle zum SCX selbst. Trotz der Änderungen an dieser durch die aktuell andauernde Entwicklung sind alle im Rahmen dieses Projekts angedachten Funktionen vollständig implementiert. Weil jedoch noch Zwischenlösungen in der SCX implementiert sind, welche diese Durchführung ermöglichen, ist noch viel Optimierungspotenzial für eine Weiterentwicklung der Gerätetreibers vorhanden. Ein Beispiel hierfür ist die Umstellung von numerischen auf Pfad-Knoten-IDs. Letztere ermöglichen einen direkten Zugriff auf Kind-Knoten mittels Pfadkonkatination, während bei ersteren ein Browse durchgeführt werden muss, um diese zu finden. Dabei werden auch viele weitere Knoten übertragen, was die Ausführung langsamer werden lässt und mehr Platz im Arbeitsspeicher einnimmt.

Dieses Problem besteht primär in der Parametrierung, da hier die SCX-Elemente durchsucht werden und diese in der Regel noch die numerischen Knoten-IDs nutzen. Da dies jedoch nur während eines Control Sync bzw. Control Imports passiert, ist im laufenden Betrieb davon in der Regel nichts zu merken. Jedoch kann, insbesondere bei großen Anlagen, die Dauer eines solchen Vorgangs dadurch im Vergleich zu einer Siemens S7-SPS mit dem entsprechenden, bereits optimierten Gerätetreiber, deutlich verlängert werden. Erste Messungen auf der für diese Entwicklungsarbeit zur Verfügung gestellten Test-SCX ergeben bspw. eine Control Sync Dauer von ca. 14 Sekunden, was im Vergleich zu einer größeren Test-S7-SPS mit einer Sync-Dauer von ca. fünf Sekunden einen deutlich langsameren Ablauf indiziert.

Zugriff nach oben (A)	Zugriff nach unten (B)	Zugriff nach oben erlaubt (C)	Zugriff nach unten erlaubt (D)	Ausgabe
0	0	X	X	1
0	1	X	0	1
0	1	X	1	0
1	0	0	X	1
1	0	1	X	0
1	1	0	X	1
1	1	1	0	1
1	1	1	1	0

Tabelle 2: Wahrheitswertetabelle

TEST

SCX-Test

Um den implementierten SCX-Gerätetreiber, insbesondere auf korrektes Verhalten, zu testen, wurde Mitte Dezember 2022 ein gemeinsamer Test mit dem SCX- und dem CPM-Server-Entwickler durchgeführt. Hierbei wurden, bis auf verschobene Pfade für einige Knoten, keine Fehler im Gerätetreiber gefunden.

Automatisierte Architekturtests

Das CPM verwendet automatisierte Unit- und Integrationstests, welche im Rahmen eines Continuous-Integration-Ansatzes bei jedem Build-Vorgang, und somit auch nach jeder Codeänderung, durchgeführt werden. In diese bestehende Testumgebung sollte ein neuer, ebenfalls automatisierter Architekturtest integriert werden. Dieser soll in Zukunft verhindern, dass neue Abhängigkeiten wie bspw. von Geräteschnittstellen in der Geschäftslogik entstehen.

Als Testframework wird ArchUnit verwendet. Dieses erlaubt die Definition von Architekturregeln auf Klassen- und Package-Ebene und besitzt vordefinierte Regeln für bestimmte Architekturtypen, wie bspw. für eine Schichtenarchitektur. Diese wurden im Rahmen der Regeldefinition für das CPM-Tool ebenfalls angewandt.

Ein großer Teil des Arbeitsaufwands für diese Architekturtests war die Erstellung einer neuen Kondition für ArchUnit, welche einen Verstoß verzeichnet, wenn Klassen Abhängigkeiten außerhalb der Super- oder Subpackages ihres eigenen Packages haben. Hierbei wird für jeden Zugriff geprüft, ob es ein Zugriff auf ein Superpackage ist, indem alle Subpackages des Ziels aufgelistet werden und das Package des Ursprungs darin gesucht wird. Ein Zugriff auf ein Subpackage wird genauso geprüft, nur Ursprung und Ziel sind vertauscht. Diese Erkenntnisse werden mit Hilfe eines logischen Ausdrucks in eine Aussage verwandelt, ob ein Zugriffsverstoß vorliegt oder nicht. Dieser ist in der Wahrheitswertetabelle in Tabelle 2 modelliert.

Zusätzlich zu der Wahrheitswertetabelle wird im ersten Schritt geprüft, ob sich sowohl Ursprung als auch Ziel des Zugriffs im selben Paket befindet. Wenn dies zutrifft, so ist ein Verstoß nicht gegeben.

ZUSAMMENFASSUNG

Das Projekt hat die Implementierung eines Gerätetreibers für die neue SCX-Steuerung, welche im Moment von SSI Schäfer entwickelt wird, zum Ziel. Ein solcher soll in Zukunft die Konfiguration und Inbetriebnahme der Maschinen erleichtern, indem die Funktionalität des CPM-Tools genutzt werden kann.

Der erste Schritt war die Absprache einer Schnittstelle mit dem SCX, da diese noch nicht existierte. Diese wurde dann im Verlauf der Implementierung auf Basis der Siemens S7-Schnittstelle immer wieder erweitert, sodass zum Schluss die hier beschriebene Schnittstelle entstanden ist.

Wie in Abbildung 9 zu sehen ist, wurde der SCX-Server im Laufe dieses Projekts um die Treiberschicht erweitert. Hierdurch konnte das grundlegende Designprinzip der Separation of Concerns, insbesondere im Vergleich mit der Situation in Abbildung 1, deutlich besser umgesetzt werden. Zudem ist die Umstellung der Mock-Implementierung auf die Treiberschicht sowie die Implementierung des neuen SCX-Gerätetreibers durchgeführt worden.

Der Gerätetreiber folgt wie die Geschäftslogik des CPMs dem dienstorientierten Modell verteilter Systeme. Hierbei sind die Dienste des Gerätetreibers unabhängig voneinander implementiert, was in Zukunft eine leichtere Änderung und Erweiterung ermöglicht. So können die zurzeit noch nicht im SCX enthaltenen CPM-Funktionen wie der Handbetrieb, die Datenmanipulation oder die Geräteoperationen in künftigen Versionen nachgereicht werden, sollten diese benötigt werden.

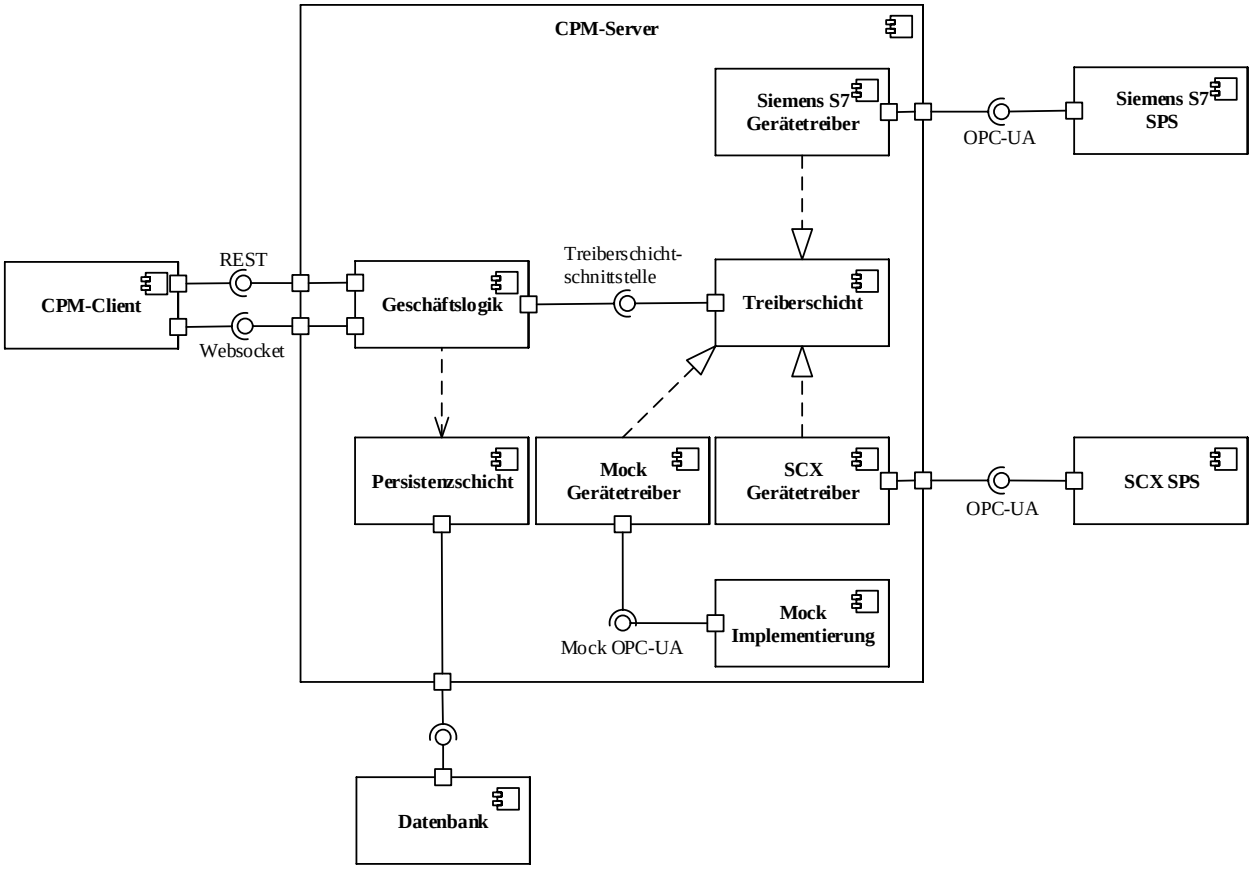


Abbildung 9: Komponentendiagramm des CPM-Tools im Überblick (Soll-Situation)

FAZIT

Die Implementierung des Gerätetreibers gestaltete sich schwieriger als zu Beginn angenommen, da sowohl in der Schnittstelle der Treiberschicht als auch in der Schnittstelle zum SCX während der Durchführung des Projekts immer wieder zu, teilweise auch konzeptionellen, Änderungen, wie bspw. bei der Protokollierung, gekommen ist. Entsprechend war eine andauernde Überarbeitung auch bereits als abgeschlossen betrachteter Dienste notwendig.

Zudem gab es Verzögerungen bei der Implementierung der vereinbarten Schnittstelle auf Seiten des SCX, welche sich aus den noch immer andauernden Lieferschwierigkeiten der Corona-Pandemie sowie der Energiekrise ergeben haben.

Trotz allem konnte die Implementierung auf Basis der aktuell gültigen Schnittstellenspezifikationen sowohl der Treiberschicht als auch des SCX bewerkstelligt werden. Dabei sind alle darin gestellten Anforderungen berücksichtigt worden und im gemeinsamen Test konnten keine funktionellen Fehler im SCX-Gerätetreiber gefunden werden.

Dementsprechend wäre dieser zur Veröffentlichung bereit, jedoch ist im Moment aufgrund der Verzögerungen noch kein Anwendungsfall in Sicht (die erste Anlage ist für 2024 anvisiert), sodass aufgrund der weiter andauernden Entwicklungen entschieden wurde, mit dieser zu warten bis sie benötigt wird. Dadurch wird verhindert, dass eine Version veröffentlicht wird, welche zum Zeitpunkt des ersten Einsatzes bereits veraltet wäre.

Dementsprechend ist die Zielsetzung des Projekts erreicht worden.

Integration eines Support-Ticketsystems in die Prozesse des Geschäftsbereichs Informatik + Systeme

Sebastian Janker, B.Sc.
Franz Laubmeier, Dipl.-Inf. (FH)

F.EE – Unternehmensgruppe
In der Seugn 20,
92431 Neunburg vorm Wald, Germany

E-Mail: sebastian.janker@fee.de
E-Mail: franz.laubmeier@fee.de

Prof. Dr. Frank Herrmann

Ostbayerische Technische Hochschule Regensburg
Innovationszentrum für Produktionslogistik
und Fabrikplanung
Galgenbergstraße 32,
93053 Regensburg, Germany

E-Mail: frank.herrmann@oth-regensburg.de

Abstract

Eine effiziente Bearbeitung von Kundenanliegen und organisierte Hilfe zu einem digitalen Produkt sind für ein modernes Unternehmen wie die F.EE GmbH unabdingbar. Ein zentraler Bestandteil für ein Support-Team beim Bereitstellen dieser Dienstleistungen ist ein Support-Ticketsystem mit Optionen zur Automatisierung von Arbeitsabläufen.
Der Geschäftsbereich Informatik + Systeme des Unternehmens F.EE GmbH verwendet zurzeit ein Ticketsystem, das aufgrund eines endenden Supportzeitraums ersetzt werden muss. Ohne Supportunterstützung dürften unter anderem Sicherheitslücken auftreten. Von den vielen am Markt verfügbaren Systemen wurde das Open-Source-Ticketsystem ausgewählt. Zur effektiven Nutzung musste es signifikant erweitert werden.

1 Einführung

Der Geschäftsbereich Informatik + Systeme des Unternehmens F.EE GmbH entwickelt und vertreibt das Enterprise-Resource-Planning (ERP)-System „FactWork", dessen Zielgruppe vor allem mittelständische Unternehmen sind [1]. Aufgrund der Vielseitigkeit und Komplexität von Fact-Work, welches sich zudem mit vielen Zusatzmodulen erweitern lässt, gibt es bei F.EE ein Support-Team, das sowohl die unternehmensinternen Benutzer, als auch die Kunden von FactWork bei Problemen unterstützt. Um diese große Zahl an Anfragen verwalten zu können, setzt das Unternehmen auf ein Ticketsystem, bei dem für jedes Anliegen ein Fall angelegt wird, sodass dieser effizient bearbeitet werden kann.

Das aktuell für den Support verwendete Ticketsystem, welches auf „Microsoft SharePoint Foundation 2013" basiert, hat das Enddatum des Standard-Supports bereits überschritten und wird daher seit dem 10. April 2018 nicht mehr offiziell unterstützt. Lediglich der erweiterte Support wird bis zum 11. April 2023 angeboten, was allerdings nur eine Übergangslösung darstellt [2]. Eine Software außerhalb des Supportzeitraums stellt für ein Unternehmen ein Problem dar, da Sicherheitslücken und Fehler auftreten können, die vom Hersteller nicht mehr behoben werden. Dadurch können hohe finanzielle Schäden und nicht zuletzt auch Imageschäden für das Unternehmen auftreten.

Um das Problem zu lösen, soll ein aktuelles Ticketsystem eingeführt und dahingehend erweitert werden, dass es in die vorhandenen Prozesse des Geschäftsbereichs eingefügt werden kann. Dafür ist es notwendig, zunächst mehrere Open-Source Projekte miteinander zu vergleichen und zu prüfen, ob diese die benötigte Grundfunktionalität zur Verfügung stellen, sodass sie sich für die unternehmensspezifischen Erweiterungen eignen.

Im Folgenden wird zunächst die Problemstellung, bestehend aus der Auswahl eines Ticketsystems und der notwendigen Anpassungen erläutert. Anschließend wird ein Lösungskonzept erarbeitet, welches bei der Realisierung umgesetzt wird. Zuletzt werden die Ergebnisse der Arbeit dargelegt und Nutzen und Rentabilität des Ticketsystems evaluiert.

2 Problemstellung

Das zentrale Problem stellt die Ablösung eines vorhandenen Ticketsystems durch ein neues System dar. Dies gliedert sich in zwei Bereiche. Zum einen ist zunächst ein geeignetes Ticketsystem auszuwählen, das die Anforderungen erfüllt, die für einen Einsatz im Geschäftsbereich Informatik + Systeme notwendig sind. Zum anderen müssen an der Software unternehmensspezifische Anpassungen vorgenommen werden, welche es ermöglichen, die Software in die vorhandenen Prozesse des Unternehmens zu integrieren.

2.1 Auswahl eines Ticketsystems

Zunächst ist ein für die Integration geeignetes Ticketsystem auszuwählen, was anhand der folgenden Kriterien vorgenommen wird. Die Software soll auf einem Windows Server betrieben werden, sodass die Kompatibilität mit Windows oberste Priorität genießt. Um hohe Lizenzkosten zu vermeiden, wird die Auswahl auf Open-Source Projekte beschränkt. Auf Grund der Tatsache, dass unternehmensspezifische Anpassungen an dem Projekt vorgenommen werden sollen, ist außerdem darauf zu achten, dass die Lizenz der Software dies zulässt. Wegen der Außenwirkung des Ticketsystems auf Kunden, wenn diese ein neues Ticket anlegen wollen, ist auch eine ansprechende Benutzeroberfläche, bestehend aus einem modernen und responsiven Design, wichtig. Um als Kandidat für das neue Ticketsystem in Frage zu kommen, muss auch die Qualität des Quellcodes den Anforderungen des Unternehmens genügen.
Aufgrund der verbreiteten Softwarekomponenten von Microsoft im Geschäftsbereich Informatik + Systeme werden bei der Recherche zu geeigneten Ticketsystemen Projekte bevorzugt, die ähnliche Softwarekomponenten verwenden. So ist zum Beispiel die Nutzung von SQL Server als Datenbank und ASP.NET als Webtechnologie im Unternehmen üblich. Obwohl sich herausstellt, dass ein Großteil der Open-Source Projekte auf PHP basiert, kann dennoch ein Projekt gefunden werden, das sowohl ASP.NET [3], als auch SQL Server verwendet. Daraufhin wird das Projekt „Quickdesk" [4], das auf Github veröffentlicht ist, genauer unter die Lupe genommen. Bei der Inspektion des Quellcodes stellt sich heraus, dass der Entwickler für sämtliche Datenbankoperationen auf „Stored Procedures" baut und diese zudem zahlreiche Möglichkeiten von SQL-Injections bieten. Da die Anzahl der Schwachstellen in den vielen Prozeduren sehr hoch ist, wäre es ein zu großer Aufwand, diese Fehler selbst zu beheben.

Ein sehr verbreitetes Ticketsystem ist „OsTicket"[5], welches als Datenbank „MySQL" verwendet. Dieses fällt vor allem durch seine Vielseitigkeit auf und nach eingehender Prüfung der Funktionen stellt sich heraus, dass es alle funktionalen Anforderungen an das neue System erfüllt. Bei den nicht-funktionalen Anforderungen offenbart sich jedoch in Form der Benutzeroberfläche ein Nachteil von OsTicket. Statt einer modernen und responsiven Benutzeroberfläche weist dieses Ticketsystem leider nur eine Webseite mit statischer Größe auf, die sich nicht an Größenänderungen des Browserfensters anpasst. Auch das Design der Webseite wirkt nicht auf dem aktuellen Stand und gibt kein sehr einladendes Gefühl an Kunden weiter, die das Support-Portal der F.EE GmbH besuchen sollen.
Ein weiterer Kandidat ist das Ticketsystem „UvDesk" [6], das mit einer modernen und ansprechenden Oberfläche glänzt, welche an das Corporate Design des Unternehmens angepasst werden kann. Dies ist ebenfalls ein Open-Source Projekt, welches unter „MIT License" auf GitHub veröffentlicht ist. UvDesk bietet ein umfassendes ereignisgesteuertes Workflow-System, mithilfe dessen alle erforderlichen Abläufe des FactWork-Supports abgebildet werden können, wie beispielsweise E-Mails an das gesamte Support-Team bei Eingang eines neuen Tickets zu senden. Als Datenbank wird hier ebenfalls MySQL genutzt. Des Weiteren basiert dieses Projekt auf dem PHP-Web-Framework Symfony [7], welches ähnlich wie Active Server Pages .NET (ASP.NET) ebenfalls auf das Konzept Model View Controller (MVC) baut und auch durch die vielen wiederverwendbaren Komponenten ein komfortables Entwickeln ermöglicht.
Die Auswahl eines Systems wird anhand der spezifizierten Kriterien getroffen. Alle angeführten Systeme erfüllen die notwendige Kompatibilität mit Windows und das Kriterium einer Lizenz, die eine Modifikation erlaubt. Die erforderliche Qualität des Quellcodes ist bei den Systemen „OsTicket" und „UvDesk" gegeben, während „Quickdesk" aufgrund der erwähnten Programmierfehler bei diesem Kriterium ausscheidet. Da die beiden verbleibenden Systeme alle funktionalen Anforderungen an die neue Software erfüllen, gilt es die Außenwirkung auf zukünftige Kunden durch die Benutzeroberfläche zu bewerten. Bei diesem Vergleich wirkt die Oberfläche von „UvDesk" wesentlich moderner und einladender, als bei „OsTicket".
Da das System „UvDesk" alle der geforderten Kriterien erfüllt, fällt die Wahl auf dieses Projekt, sodass im Folgenden die Anforderungen für die Integration dieser Software definiert werden können.

2.2 Integration des ausgewählten Systems

Um das neue Ticketsystem in die vorhandenen Prozesse zu integrieren, gilt es folgende Anforderungen zu erfüllen.

a. **Anlegen eines neuen Supportfalls im ERP-System**

Für das ausgewählte Ticketsystem ist eine Erweiterung notwendig, die für eine Kommunikation mit dem ERP-System „FactWork" sorgt. Die erstellten Tickets müssen auch in die dortige Datenhaltung eingepflegt werden, damit der zuständige Mitarbeiter des Support-Teams seinen für dieses Ticket aufgebrachten zeitlichen Aufwand auf den entsprechenden Fall buchen kann. Aus diesen Buchungen wird dann der erfasste Aufwand für einen Supportfall eines Kunden in Rechnung gestellt. Diese Übernahme eines Tickets in FactWork kann allerdings erst dann erfolgen, wenn ein eingegangenes Ticket von einem zuständigen Mitarbeiter übernommen wurde, da verschiedene Fälle auftreten können, wo dies fehlerhaft wäre. Beispielsweise könnte das Ticket einem bereits bestehenden Fall zugeordnet oder im Falle eines Spams gelöscht werden.
Ziel ist es, dass nach Übernahme eines Tickets automatisch eine Schnittstelle von FactWork angesprochen wird, über die alle benötigten Informationen für die Erstellung eines Supportfalls übermittelt werden.

b. **Anlegen eines Falls im „Bugtracking-System"**

Das Anlegen eines Falls im Bugtracking-System des Geschäftsbereichs Informatik + Systeme soll ebenfalls mithilfe des Ticketsystems möglich sein. Sollten für ein Ticket Änderungen oder Prüfungen am Code der Software FactWork notwendig sein, so wird dies im Bugtracking-System verwaltet. Daher muss die Möglichkeit geschaffen werden, für einen eingegangenen Supportfall, der eine Codeänderung oder zumindest eine Prüfung des Quellcodes erfordert, einen Fall im Bugtracking-System zu erstellen, sodass dem Ticket die entsprechende Codeänderung zugeordnet werden kann.

c. **Authentifizierung am Ticketsystem mithilfe des Domänen-Accounts**

Die Mitarbeiter des Supports-Teams sollen sich mithilfe ihres Windows-Domänen-Accounts am Ticketsystem anmelden können. Hierbei muss sichergestellt werden, dass nur die Mitarbeiter, welche diesem Team angehören, Zugangsberechtigungen zu dem eingeführten Ticketsystem erhalten. Wünschenswert wäre hier auch, dass sich die Mitarbeiter nicht bei jedem Öffnen des Ticketsystems erneut authentifizieren müssen, sondern dies mit einem sogenannten Single Sign On (SSO)-Verfahren realisiert wird, sodass für den Login automatisch der Domänen-Account verwendet wird, falls der Benutzer am PC mit diesem eingeloggt ist.

d. **Versenden einer Antwort mit Outlook**

Damit das Versenden einer Antwort auf ein Ticket durch einen Support-Mitarbeiter neben der Weboberfläche auch wie gewohnt mit Outlook möglich ist, soll umgesetzt werden, dass aus dem Ticketsystem heraus optional das Outlook-Fenster für das Versenden einer neuen E-Mail geöffnet werden kann. Hierbei soll der Betreff bereits befüllt und die letzte E-Mail des Kunden zitiert sein, damit schnell und effizient geantwortet werden kann. Dabei ist vor allem sicherzustellen, dass das Ticketsystem von dieser Antwort auf das Ticket Kenntnis erlangt, damit diese auch im Nachrichtenverlauf aufgeführt werden kann.

e. **Outlook-Add-ln für Kommunikation mit dem Ticketsystem**

Statt über die E-Mail-Adresse des Supports gehen immer wieder Anfragen direkt im persönlichen E-Mail-Postfach der Mitarbeiter des Support-Teams ein. Um diese effizient weiterverarbeiten zu können, soll ein Outlook-Add-In entwickelt werden, welches über eine Schnittstelle mit dem Ticketsystem kommuniziert. Mit diesem soll die Möglichkeit geschaffen werden, eine E-Mail direkt aus Outlook in das Ticketsystem zu übernehmen. Auf diese Weise kann das Ticket dann vom zuständigen Supportmitarbeiter übernommen werden. Auch die Funktion, das Ticket direkt selbst zu übernehmen, ist hier zu implementieren. Damit man aus Outlook heraus komfortabel zum entsprechenden Fall im Ticketsystem gelangt, soll das Add-In außerdem die Möglichkeit bieten, einen für eine E-Mail erstellten Fall direkt aus der geöffneten E-Mail aufzurufen. Es kommt auch immer wieder vor, dass Kunden nicht direkt auf die erhaltene E-Mail antworten, sodass diese nicht dem entsprechenden Ticket zugeordnet werden kann. Deswegen soll das Add-In zudem die Option bieten, die eingegangene E-Mail einem vorhandenen Ticket aus dem System zuzuordnen.

f. **Hosting des Ticketsystems mithilfe von Microsoft lnternet lnformation Services (llS)**

Die meisten Ticketsysteme, die mit PHP entwickelt wurden, sind hauptsächlich für Unix-ähnliche Betriebssysteme und vor allem für den Betrieb mit dem Webserver Apache [8] vorgesehen. Da das Ticketsystem jedoch auf einem Windows Server gehostet werden soll, ist es naheliegend, den von Microsoft mitgelieferten Webserver IIS zu nutzen. Eventuell dafür notwendige Anpassungen und das Hosting selbst durchzuführen, ist ebenfalls Aufgabe dieser Arbeit.

g. **Übergang vom Altsystem zum neuen Ticketsystem**

Ein Problem, das gelöst werden muss, stellt außerdem der Übergang vom Altsystem zum neuen Ticketsystem dar. Hierbei muss eine Lösung erarbeitet werden, damit alle laufenden Supportfälle, welche sich teilweise aufgrund hoher Komplexität auch über einen größeren Zeitraum erstrecken, lückenlos und zuverlässig bearbeitet werden können. Daher ist zu prüfen, ob eine Migration der Daten aus dem Altsystem in die neue Software mit vertretbarem Aufwand realisierbar ist, oder ob ein hybrider Betrieb die ökonomischere Lösung ist. Zu berücksichtigen gilt es hier, dass für letztere Variante bei der Verarbeitung eingehender E-Mails eine Unterscheidung notwendig ist, ob es sich um eine E-Mail handelt, die sich auf ein Ticket im neuen oder im abzulösenden System bezieht.
Für diese identifizierten Probleme wird im folgenden Abschnitt ein Lösungskonzept erarbeitet.

3 Lösungskonzept

Das Lösungskonzept beinhaltet zum einen die unternehmensspezifischen Anpassungen am Ticketsystem, zum anderen externe Komponenten und die Bereitstellung des Systems.

3.1 Anpassungen am Ticketsystem

Um Problem a zu lösen, ist zunächst die Erstellung einer Schnittstelle zum ERP-System FactWork notwendig. Für den externen Zugriff auf die FactWork Datenhaltung besteht als Schnittstelle bereits ein HTTP-REST-Server, welcher unter anderem von der mobilen App „FactWork Mobile" [9] genutzt wird, um mit dem System zu interagieren. Dieser ist um eine weitere API zu erweitern, mit deren Hilfe das Ticketsystem einen neuen Supportfall im ERP-System anlegen kann.
Damit die Tickets auch zwischen den verschiedenen Softwarekomponenten eindeutig identifizierbar sind, soll die Nummer des erstellten Supportfalls in FactWork identisch zur Nummer der Datenbankentität des Ticketsystems sein und nicht etwa in einem separaten Attribut gespeichert werden. Daher wird bei Erstellung eines Tickets zunächst mithilfe eines benutzerdefinierten Generators eine temporäre Nummer vergeben, die sich nicht mit dem Wertebereich der Nummern in FactWork überschneidet und die aktualisiert wird, sobald der Fall in FactWork erstellt wurde.
Wichtig ist, dass das Ticketsystem einen solchen Fall im System allerdings nur dann anlegt, wenn sichergestellt ist, dass es sich bei einem erstellten Ticket um einen validen neuen Supportfall handelt.
So kann ein neues Ticket beispielsweise nachträglich zu einem bestehenden Ticket hinzugefügt werden, wenn die automatische Zuordnung nicht funktioniert hat, da der Kunde nicht direkt auf die E-Mail geantwortet hat. Außerdem besteht die Möglichkeit, dass es sich bei dem Ticket um Spam handelt. Um dies zu erreichen, wird ein Supportfall im ERP-System erst dann angelegt, wenn der entsprechende Fall von einem Mitarbeiter übernommen wird.
Für eine Kommunikation mit der Representational State Transfer (REST)-API ist eine Authentifizierung des Ticketsystems gegenüber des FactWork REST-Servers notwendig. Dies erfolgt durch ein tokenbasiertes Authentifizierungsverfahren. Durch die Authentifizierung mit einem für das Ticketsystem erstellten FactWork-Account wird vom Server ein Long-lived-Token abgefragt. Auf Basis dieses Tokens kann mit einem Algorithmus, welcher zudem mit einem Public-Key Verfahren arbeitet, ein temporärer Token berechnet werden. Dieser wird als „Bearer Token" [10] zur Authentifizierung bei Anfragen an den REST-Server verwendet. Der temporäre Token besitzt eine Gültigkeit von 15 Minuten und muss nach Ablauf dieser Zeit erneut berechnet werden.
Mithilfe eines API-Aufrufs wird bei einer Übernahme durch einen Mitarbeiter aus den Daten des Tickets ein neuer Supportfall im ERP-System angelegt. Der REST-Server gibt dem Ticketsystem nach der Erstellung des Falls die vergebene Nummer zurück, sodass die temporäre Id des Tickets durch die FactWork-Nummer ersetzt werden kann.
Um Anforderung b zu erfüllen, ist ein zusätzlicher Menüpunkt in der Ticketansicht des Systems notwendig, welcher einen Dialog öffnet, mit dem ein zum Ticket gehöriger Fall im Bugtracking-System angelegt werden kann. Neben der Bezeichnung sind auch die Auswahlfelder Art, Projekt, Kategorie und Priorität anzugeben, bei denen zunächst die möglichen Optionen aus der Datenbank des Bugtracking-Systems zu ermitteln sind. Außerdem wird im System ein Link zum Ticket hinterlegt, damit aus dem Fall im Bugtracking-System schnell und komfortabel zum entsprechenden Ticket navigiert werden kann. Ebenso ist der erstellte Fall im Bugtracking-System über einen Direktlink in der Ticketansicht erreichbar, sodass auch hier schnell die zugehörigen Daten aufrufbar sind. Da die endgültige Id des Tickets erst bei Übernahme des Falls durch einen Mitarbeiter und entsprechender Anlage in FactWork feststeht, wird die Option zum Erstellen eines Bugtracking-Falls erst freigeschaltet, sobald das Ticket einem Mitarbeiter zugewiesen wurde, damit der generierte Link zum Ticket gültig bleibt.

3.2 Hosting des Ticketsystems

Das Ticketsystem soll auf einem Windows Server mithilfe des integrierten Webservers IIS gehostet werden (vgl. Punkt f). Um das Problem zu lösen, wird im IIS ein neuer Anwendungspool und eine Website benötigt. Damit der Webserver die Dateien mit PHP-Code ausführen kann, muss eine „Handlerzuordnung" für die Dateierweiterung „.php“ festgelegt werden [11]. Zudem ist der Pfad einer URL so umzuschreiben, dass dieser immer ausgehend von der Hauptdatei „index.php" ist, damit das Routing funktioniert, ohne den Dateinamen anzugeben. Dies ist mit der Funktion „URL Rewrite" des IIS zu bewerkstelligen.
Um Problem c zu lösen, muss das Ticketsystem dahingehend erweitert werden, dass eine Anmeldung mit dem Windows-Domänen-Account möglich ist. Eine Möglichkeit, dies zu realisieren, bietet der interne Dienst „PasswortManager", der bereits bei anderen Softwaresystemen die Authentifizierung übernimmt. Dies funktioniert über eine Schnittstelle, die den offenen OAuth2-Standard untertützt. Hierbei ist jedoch erforderlich, dass sich die Mitarbeiter des Support-Teams authentifizieren, wenn sie das Ticketsystem öffnen. Um die Benutzung für die Mitarbeiter noch effizienter und komfortabler zu gestalten, wird der Ansatz des Single Sign On (SSO) verwendet, damit die Benutzer automatisch eingeloggt werden, wenn sie bei Windows bereits mit ihrem Domänen-Account angemeldet sind. Für dieses Vorhaben scheidet die Authentifizierung mit dem PasswortManager und OAuth2 aus. Daher soll die Anmeldung mithilfe der Windows-Authentifizierung des IIS umgesetzt werden. Damit ein Benutzer mittels des SSO-Verfahrens automatisch am Ticketsystem eingeloggt wird, muss die Domäne, unter der das Ticketsystem erreichbar ist, unter Internetoptionen als „Lokales Intranet" oder „Vertrauenswürdige Site" definiert werden. Dies kann auf den Rechnern im Unternehmensnetzwerk in einer Gruppenrichtlinie definiert werden.
Ein durch den IIS authentifizierter Benutzer kann auf Seiten von PHP aus der globalen Variable $_SERVER abgerufen werden. Ist ein Benutzer authentifiziert, so ist dessen Benutzername im Attribut „REMOTE_USER" gespeichert. Um einen Benutzer mithilfe dieser Authentifizierungsmethode einzuloggen, wird das vorhandene Verfahren des „Symfony Security Bundle" verwendet und die sogenannte Firewall dahingehend abgeändert, dass der Inhalt von „REMOTE_USER" als Benutzername für die Authentifizierung verwendet wird [12]. Des Weiteren muss ein benutzerdefinierter User-Provider [13] erstellt und zur Verwendung für den Login-Prozess angegeben werden. Dieser ermittelt anhand des Benutzernamens den authentifizierten Benutzer aus der Datenbank und gibt diesen zurück, damit er vom Symfony Security Login-Prozess weiterverarbeitet wird und schließlich angemeldet ist. Um aufwändige Konfigurationsarbeiten zu vermeiden, wird für diesen Benutzernamen außerdem eine REST-Abfrage an den PasswortManager gesendet, um die im Active Directory (AD) hinterlegten Benutzerdaten zu erhalten. Mit diesen Daten wird ein vorhandener Benutzer aktualisiert oder gegebenenfalls ein neuer Benutzer in der Datenbank des Ticketsystems angelegt. Um sicherzustellen, dass sich nur berechtigte Mitarbeiter am System anmelden können, kann in einer Konfigurationsdatei eine AD-Gruppe angegeben werden, sodass sich ausschließlich deren Mitglieder am Ticketsystem authentifizieren können.

3.3 Integration von Outlook

Problemstellung d wird gelöst, indem das Ticketsystem um eine Schaltfläche erweitert wird, die es erlaubt, in der Ticketansicht ein neues Outlook-Element zu öffnen, bei dem bereits die notwendigen Daten vorausgefüllt sind, sodass der Mitarbeiter des Support-Teams lediglich den Antworttext ergänzen muss. Dies ist mit dem URI-Schema „mailto“ [14] umsetzbar. Damit die E-Mail trotz des Versands mit einer Drittsoftware im Datenbestand des Ticketsystems vorhanden ist, wird die Verarbeitung eingehender Tickets von UvDesk so erweitert, dass der Betreff der Antwort, dessen Syntax bereits vom Altsystem vorgegeben ist, ausgewertet wird und die Antwort so der enthaltenen Ticket-Id zugeordnet werden kann. Damit die E-Mail auch im eingehenden Postfach des Ticketsystems erscheint, wird beim Öffnen des Outlook-Fensters die Blind Carbon Copy (BCC) Zeile bereits mit der E-Mail-Adresse des Systems vorausgefüllt.
Um Problem e zu lösen wird ein VSTO Outlook Add-In erstellt, welches ein effizientes Abarbeiten von E-Mails, die direkt an Mitarbeiter gesendet werden, ermöglichen soll. Diese Erweiterung stellt sowohl im Startmenü, als auch im Menüband bei einer geöffneten E-Mail die in Problem e erwähnten Funktionen zur Verfügung (vgl. Abbildung 1).

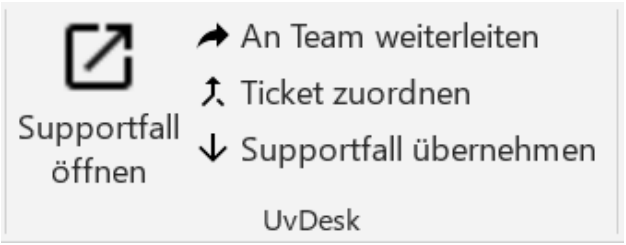


Abbildung 1: Outlook VSTO-Add-In

Für die Umsetzung der spezifizierten Anforderungen an die Outlook-Erweiterung wird das Ticketsystem um eine API erweitert, welche benötigte Funktionen zur Verfügung stellt. Das Öffnen eines zu einer E-Mail gehörigen Supportfalls wird durch das Ermitteln des entsprechenden Links über die API und Anschließendem Öffnen der Ticketansicht im Browser umgesetzt. Dabei erfolgt die Identifikation des richtigen Tickets über eine in der E-Mail-Kopfzeile gesetzte „References"-Eigenschaft oder der im Betreff befindenden Ticketnummer. Für die Funktionen „An Team weiterleiten" und „Supportfall übernehmen" wird

Kopfzeile und Nachricht der E-Mail inklusive Anlagen zunächst unter Verwendung der C# Bibliothek „Outlook Redemption" in das EML-Format konvertiert und so an eine Schnittstellenfunktion übergeben. Für beide Funktionen wird jeweils aus den übergebenen E-Mail-Daten ein Ticket erstellt, wobei beim Übernehmen des Supportfalls zusätzlich der entsprechende Mitarbeiter als Bearbeiter des Tickets gesetzt wird. Um auch das Zuweisen einer E-Mail zu einem bestehenden Fall im System zu ermöglichen, erscheint zunächst ein Fenster, das die bestehenden Fälle auflistet, die von der erstellten API beschafft werden. Nach Auswahl des Ziels der Zuweisung wird der Vorgang mithilfe einer weiteren API-Funktion durchgeführt.

3.4 Übergang zum neuen System

Um den in Problem g beschriebenen Übergang vom Altsystem zum neuen Ticketsystem ohne Ausfälle umzusetzen, ist eine Möglichkeit zu schaffen, dass auch die bestehenden Tickets weiterhin bearbeitet werden können. Eine Migration der Daten in das neu eingeführte Ticketsystem hat den Vorteil, dass kein Parallelbetrieb notwendig ist und daher keine Unterscheidung bei eingehenden E-Mails getroffen werden muss, ob es sich um einen Fall im neuen System oder um eine Antwort auf ein bestehendes Ticket im Altsystem handelt. Diese Lösung birgt allerdings das Problem, dass alle Links zu den Supportfällen, die in FactWork oder im Bugtracking-System hinterlegt sind, ungültig werden und daher eine Aktualisierung erfordern. Des Weiteren ist eine zusätzliche Konvertierungsroutine zu entwickeln, die über eine API des SharePoints alle notwendigen Daten ermittelt und aus diesen die Supportfälle inklusive der Anhänge erstellt. Daher erweist sich ein temporärer Parallelbetrieb der Systeme als ökonomischere Lösung des Problems. Dabei wird übergangsweise eine Funktion entwickelt, die anhand des Betreffs prüft, ob sich die E-Mail auf einen bestehenden Fall aus dem Altsystem bezieht und diese dann als EML-Datei in ein Verzeichnis auf dem Hosting-Server ablegt, aus der das Altsystem eingehende E-Mails bezieht. Sobald der Parallelbetrieb beendet ist, kann diese Funktion abgeschaltet werden.
Es gilt auch den Fall zu beachten, dass ein Kunde nicht direkt auf eine E-Mail antwortet, sodass die automatische Zuordnung zu einem Supportfall fehlschlägt und folglich ein neues Ticket im System erstellt wird. Hierfür ist eine Anpassung des Ticketsystems nötig, die dafür sorgt, dass zu jedem erstellten Ticket die originale E-Mail als EML-Datei gespeichert wird. Diese kann dann in Outlook geöffnet, mithilfe eines Add-Ins zu einem Supportfall im Altsystem hinzugefügt und in UvDesk schließlich gelöscht werden.

4. Realisierung

Als Basis für eine Evaluierung der bestehenden Funktionalitäten dient ein Testsystem, auf dem das ausgewählte Ticketsystem zusammen mit der dafür notwendigen Open-Source-Datenbank MariaDB bereitgestellt wird. Um eine praxisnahe Testumgebung zu schaffen, sind zudem realistische Testdaten notwendig. Die abschließende Prüfung der Anforderungen ergibt, dass alle benötigten Grundfunktionen gegeben sind und daher die unternehmensspezifischen Anpassungen umsetzbar sind. Dabei wird sichergestellt, dass bei der Wahl des Ticketsystems keine gravierende Fehlentscheidung getroffen wurde, die zu vermeidbaren zusätzlichen Kosten führt.
Da es sich bei dem ausgewählten System um ein Open-Source-Projekt handelt, das von der breiten Community stets weiterentwickelt und um neue Funktionen erweitert wird, erscheinen in regelmäßigen Abständen neue Versionen des Ticketsystems. Daher ist es im Interesse des Unternehmens, das eingesetzte Ticketsystem stets auf dem aktuellsten Stand zu halten. Um den aktualisierten Quellcode und die unternehmensspezifischen Anpassungen beim Update möglichst einfach zusammenfügen zu können, wird beim gesamten Entwicklungsprozess darauf geachtet, die eigenen Änderungen möglichst gut abzugrenzen und in eigene Dateien auszulagern.
Wie unter Abschnitt 2 beschrieben muss das ausgewählte Ticketsystem dahingehend angepasst werden, dass es in die Prozesse der vorhandenen Softwarearchitektur integriert werden kann. Hierfür ist zu realisieren, dass für Tickets automatisch Supportfälle im ERP-System FactWork erstellt werden. Für die geplante Kommunikation mit dem ERP-System FactWork muss in Absprache mit dem zuständigen Team eine geeignete Schnittstelle zum Anlegen neuer Supportfälle mithilfe des Factwork REST-Servers erstellt werden. Ein Testsystem, auf dem FactWork mit einer anonymisierten Demodatenbank und der REST-Server installiert sind, soll dazu dienen, die Implementierung der API-Aufrufe seitens des Ticketsystems und das dafür benötigte Authentifizierungsverfahren zu testen. Da die Nummer eines Tickets nach der Erstellung eines Supportfalls in FactWork identisch zur FactWork-Nummer sein soll, muss die Datenbankentität dahingehend angepasst werden, dass der Primärschlüssel nicht mehr automatisch von der Datenbank generiert wird, sondern temporäre Nummern von einem Generator vergeben werden, die nicht mit dem Wertebereich der FactWork-Nummern kollidieren. Um die Daten konsistent zu halten, ist für alle Fremdschlüssel, die auf das Ticket verweisen, ein „On Update Cascade" notwendig. Da das verwendete Framework Doctrine ORM dies allerdings nicht unterstützt, werden die Fremdschlüssel durch das Ausführen einer „Stored Procedure" nach einem Update des Datenbankschemas angepasst.

Eine weitere Anforderung stellt das Anlegen von Fällen im Bugtracking-System dar. Auch hierfür ist eine Demodatenbank auf einem weiteren Testserver notwendig, aus der zunächst die für das Formular notwendigen Daten ermittelt werden und dann ein entsprechender Fall angelegt werden kann. Um aus einem Ticket effizient einen Fall im Bugtracking-System erstellen zu können, wird das Ticketsystem um einen Dialog erweitert, der das Formular für die notwendigen Daten zur Verfügung stellt. Dies ist bereits mit den Daten vorausgefüllt, die anhand des zugehörigen Tickets ermittelt werden können, sodass auch dieser Vorgang mit möglichst wenig Aufwand umsetzbar ist. Für die Überprüfung der Berechtigung zum Anlegen eines solchen Falls werden die im Projekt bereits vorhandenen Privilegien durch ein weiteres ergänzt, sodass dies für Berechtigungsgruppen individuell konfigurierbar ist.
Die Entwicklung des Outlook-Add-Ins ist eine weitere Anforderung, die zu realisieren ist. Die vom Visual Studio Tools for Office (VSTO) Add-In benötigte Funktionalität stellt eine API zur Verfügung, um die das Ticketsystem zu erweitern ist. Damit das angefertigte Menüband neben dem Startmenü auch im Menü einer geöffneten Nachricht hinzugefügt wird, muss dieses kopiert werden, da das Menü nicht für beide Orte gleichzeitig konfiguriert werden kann. Um redundanten Code zu vermeiden, wird die Logik in einen wiederverwendbaren Dienst ausgelagert. Die Authentifizierung gegenüber der API des Ticketsystems wird unter Verwendung von „Default Credentials" [15] realisiert, welche vom angemeldeten Windows Account bezogen werden. Die Prüfung, ob der Benutzer für die gewünschte Funktion eine Berechtigung besitzt, wird seitens der erstellten REST API mithilfe des Rechtesystems von UvDesk durchgeführt.

5. Evaluation

Abschließend wird Nutzen und Rentabilität der Integration des neuen Ticketsystem bewertet.

5.1 Sicherheitsaspekt

Da das Unternehmen hohe Priorität auf Informationssicherheit legt, ist es wichtig, verwendete Software stets auf einem aktuellen Stand zu halten und dadurch potenzielle Sicherheitslücken zu vermeiden. Solche Lücken sind für Unternehmen sehr problematisch, da durch einen Ausfall eines Systems sowohl sehr hohe Finanz- als auch Imageschäden entstehen können. Der gesamte finanzielle Schaden für das Unternehmen ist sehr schwer abzuschätzen, da der Imageschaden auf lange Sicht ebenfalls finanzielle Einbußen zur Folge hat, die schwer zu bemessen sind.
Im Falle eines Ausfalls des Ticketsystem im Geschäftsbereich Informatik + Systeme könnten die Mitarbeiter Supportfälle nicht oder nur sehr eingeschränkt bearbeiten. Zunächst müssten eingehende Tickets manuell aus dem Postfach geholt und mit allen Kollegen besprochen werden, um zu vermeiden, dass zwei Mitarbeiter denselben Fall bearbeiten. Auch wäre der bisherige Schriftverkehr zu einem bestehenden Ticket nicht einsehbar. Durch diese Einschränkungen ist mindestens der doppelte Zeitaufwand im Vergleich zum Normalfall notwendig. Allein durch die 7 Mitarbeiter, welche momentan mit dem Altsystem arbeiten, entsteht bei einer 8-stündigen Schicht ein Mehraufwand von $7 \cdot 4h = 28h$ pro Tag.
Das bisher verwendete Ticketsystem basiert auf „Microsoft SharePoint Foundation 2013", welcher das Enddatum des Standard-Supports mit dem 04. Oktober 2018 bereits überschritten hat und sich nun im erweiterten Support befindet, der am 11. April 2023 endet. Daher soll dieses nun zeitnah durch ein aktuelles System ersetzt werden.

5.2 Lizenzkosten

Dass das vorhandene Ticketsystem durch ein aktuelleres ersetzt werden muss, wurde bereits festgestellt. Nicht zuletzt auch wegen der zusätzlichen Kosten, die auf das Unternehmen zukommen, wenn außerhalb des Support-Zeitraums Fehler und Sicherheitslücken seitens Microsoft behoben werden müssen. Es stellt sich allerdings die Frage, warum man sich dafür entschieden hat, statt eines Upgrades auf eine neuere SharePoint Version auf die Integration eines gänzlich anderen Systems zu setzen. Die momentan verwendete Microsoft SharePoint Foundation 2013 ist eine kostenlose Variante, die Microsoft für die aktuellen SharePoint Versionen nicht mehr anbietet, wodurch für eine neue Version hohe Lizenzkosten auf das Unternehmen zukämen (vgl. Tabelle 1). Das Unternehmen F.EE legt Wert darauf, dass alle internen Daten im Haus gespeichert sind, daher kommt nur eine On Premise Lösung in Frage. Für den Weiterbetrieb der SharePoint Lösung sind eine Server Lizenz und für jeden Mitarbeiter (aktuell ca. 20) eine Client Access License (CAL) erforderlich, die vom Unternehmen erworben werden müssen.

Artikel	Preis
SharePoint Server 2019	7085,55 €
SharePoint Server 2019 CAL	2549,20 €
SQL Server Standard 2019	850,62 €
SQL Server CAL 2019	3.947,60 €
Summe	14.432,97 €

Tabelle 1: Angebot für Sharepoint 2019

Aufgrund der Tatsache, dass für das neue Ticketsystem auf ein Open-Source-Projekt gesetzt wird, können für das Unternehmen die in Tabelle 1 aufgeführten Lizenzkosten für eine neue SharePoint Version eingespart werden.
Aus Gründen der Kompatibilität ist beim Update auf die neue SharePoint-Version auch ein aktuellerer SQL Server notwendig, sodass auch hierfür Lizenzkosten

vom Unternehmen getragen werden müssen. Viele Neuerungen und Fehlerkorrekturen bewerkstelligt zudem die Open-Source- Community, die das Ticketsystem weiterentwickelt, sodass interne Entwicklungsarbeit eingespart werden kann.

5.3 Zeitersparnis

Als Evaluationsumgebung soll ein Testsystem dienen, auf dem eine vollständige Installation des Ticketsystems durchgeführt wurde. Testpersonen stellen dabei die zukünftigen Anwender dar, da diese bereits mit den Abläufen vertraut sind. Um ein möglichst realistisches Ergebnis zu erhalten, werden für die Simulation der Bearbeitung von Supportfällen mithilfe des neuen Ticketsystems echte Kundenanliegen herangezogen, die in anonymisierter Form in die Testdatenbank übernommen werden.

5.3.1 FactWork Support-Team

Das neue Ticketsystem bietet viele Funktionen, die eine Bearbeitung der Anliegen von Kunden effizienter und dadurch kostengünstiger machen. Neben individuellen und komplexeren Kundenanfragen gibt es auch häufig auftretende, wiederkehrende Probleme und Unklarheiten, bei denen die Kunden von den Mitarbeitern des Support-Teams unterstützt werden. Diese können dank des neuen Ticketsystems durch vorbereitete Antworten mit geringerem Zeitaufwand bearbeitet werden. Unterstützt wird dies zudem von einem umfangreichen Workflow-System [16], wodurch viele Arbeitsabläufe automatisiert und somit den Mitarbeitern redundante Arbeiten abgenommen werden können. Es gibt eine Vielzahl an Anfragen, die unter die Kategorie „Handhabung und Bedienung von FactWork" fallen, bei denen es möglich ist, eine vorbereitete Antwort zu verwenden. Beispielsweise benötigt ein Kunde Unterstützung beim Anlegen von neuen FactWork-Benutzern. Hier kann der bearbeitende Mitarbeiter eine vorbereitete Antwort zum Thema „Anlegen eines FactWork-Benutzers" an den Kunden senden, die genau beschreibt, wie in diesem Fall vorzugehen ist. Dadurch wird vermieden, dass der Mitarbeiter die Anleitung zunächst im FactWork Handbuch nachschlagen und dann eine ausführliche Antwort verfassen muss, um dem Kunden zu helfen.
Eine weitere Neuerung stellt das Ticketformular für Kunden dar, welches das Ticketsystem standardmäßig mitbringt. Neben dem Eingang von Supportanfragen im E-Mail-Postfach melden auch viele Kunden ihre Probleme telefonisch. Daher kann zum einen Zeit gespart werden, wenn sich ein Teil der Kunden für den Weg des Ticketformulars entscheidet, zum anderen wird durch das Führen des Kunden durch das Formular sichergestellt, dass alle zur Bearbeitung notwendigen Angaben enthalten sind, sodass keine zusätzlichen Rückfragen notwendig sind.
Da sich das System zurzeit noch nicht im Produktivbetrieb befindet, können aktuell noch keine Daten aus dem realen Alltag erhoben werden. Hierfür wird die bereits beschriebene Testumgebung genutzt. Da die eingehenden Tickets sehr unterschiedlich sind und oft auch sehr individuelle Probleme beinhalten, kann nur bei Supportfällen, die sich für eine Automatisierung anbieten, eine zeitliche Einsparung erzielt werden. Die Simulation mithilfe des Testsystems ergab eine durchschnittliche Einsparung von 3 Minuten pro Fall. Eine Auswertung der Anzahl aller Supportfälle bringt für das vergangene Jahr 2021 eine Gesamtzahl von 2972 Tickets hervor, woraus sich eine eingesparte Zeit von 8916 Minuten, also rund 149 Stunden pro Jahr berechnet.

5.3.2 IT-Systemhaus

Neben der Ablösung des Ticketsystems der FactWork Supportabteilung soll das neu eingeführte System auch im IT-Systemhaus der F.EE GmbH zum Einsatz kommen. Dort wurde bisher noch kein Ticketsystem eingesetzt und die Kundenanfragen wurden direkt an die E-Mail-Adresse eines Mitarbeiters gesendet. Das hat zur Folge, dass bei einer Abwesenheit eines Mitarbeiters ein Fall nicht durch einen anderen Kollegen übernommen werden kann, bzw. dieser sich erst aufwendig in das Anliegen einarbeiten muss, ohne Zugriff auf den bisherigen Schriftverkehr zu haben. Durch das Ticketsystem werden die Kundenanfragen organisierter, für alle berechtigten Mitarbeiter zugänglich und sind vor allem effizient abzuarbeiten.
Da in dieser Abteilung bis zum aktuellen Zeitpunkt kein Ticketsystem eingesetzt wurde, kann bei der Evaluierung mithilfe des Testsystems eine deutlich höhere Zeiteinsparung pro Fall festgestellt werden. Aufgrund der besseren Aufteilung auf alle zuständigen Mitarbeiter können die Tickets wesentlich effizienter abgearbeitet werden, als dies der Fall ist, wenn Supportfälle bei den einzelnen Mitarbeitern persönlich eintreffen. Auch ein aufwändiges Einarbeiten in den bisherigen Schriftverkehr ist nicht nötig, da alle Nachrichten des Falls im System einsehbar sind. Bei den Messungen konnte eine durchschnittliche Zeitersparnis von 10 Minuten pro Fall erzielt werden. Für die 1169 bearbeiteten Fälle des IT-Systemhauses sind daher pro Jahr 11690 Minuten, also ca. 195 Stunden weniger Zeitaufwand zu verzeichnen.

6 Zusammenfassung

Für ein Unternehmen, das täglich eine Vielzahl von Kundenanfragen bearbeitet, ist es unerlässlich, ein gut funktionierendes Ticketsystem zu verwenden, welches eine effiziente und strukturierte Abarbeitung der eingehenden Fälle ermöglicht. Aufgabe dieser Arbeit ist die Integration eines besagten Systems in die Prozesse des Geschäftsbereichs Informatik + Systeme der F.EE GmbH. Die zentrale Problemstellung hierbei ist neben der Auswahl eines geeigneten Systems die Integration der vorhandenen Softwarekomponenten, mit denen das Ticketsystem kommunizieren muss. Die Einführung eines neuen Systems ist notwendig, da das bisher verwendete System veraltet ist und nicht mehr den

aktuellen Sicherheitsstandards entspricht. Falls also bei der Software Sicherheitslücken oder Fehler auftreten, werden diese nicht vom Hersteller behoben, was sehr hohe finanzielle Schäden als auch Imageschäden zur Folge haben kann.
Zur Lösung dieses Problems ist ein aktuelles Ticketsystem einzuführen, das regelmäßige Updates veröffentlicht. Daher werden verschiedene Open-Source Projekte anhand von spezifizierten Kriterien miteinander verglichen, um das für die Integration am besten geeignete System zu ermitteln. Durch das erarbeitete Lösungskonzept, das die definierten Anforderungen an das neue System erfüllen soll, wird das ausgewählte Ticketsystem dahingehend angepasst, dass es in die vorhandenen Prozesse des Geschäftsbereichs Informatik + Systeme integriert werden kann.
Dadurch können potenzielle Sicherheitslücken durch den Einsatz veralteter Software vermieden werden. Aufgrund der Tatsache, dass die neue Lösung auf einem Open-Source Projekt aufbaut, können zudem hohe Lizenzkosten für das Unternehmen eingespart werden. Außerdem kann durch die Möglichkeit der Automatisierung wiederkehrender Arbeitsabläufe, welche das neue Ticketsystem bietet, eine effizientere Bearbeitung der Anfragen gewährleistet werden.

Literatur

[1] F.EE GmbH. „Die Unternehmenssoftware aus der Praxis für die Praxis." (2022), Adresse: https://www.factwork.de/ (besucht am 31.01.2022).

[2] Microsoft Corporation. „SharePoint Foundation 2013 (Supportzeiträume)." (2022), Adresse: https://docs.microsoft.com/lifecycle/products/sharepoint-foundation-2013 (besucht am 31.01.2022).

[3] Microsoft Corporation. „ASP.NET, Free. Cross-platform. Open source. A framework for building web apps and services with .NET and C#." (2022), Adresse: https://dotnet.microsoft.com/apps/aspnet (besucht am 02.02.2022).

[4] Saineshwar Bageri. „QuickDesk." (2022), Adresse: https://github.com/saineshwar/QuickDesk (besucht am 29.01.2022).

[5] Enhancesoft. „OsTicket." (2022), Adresse: https://www.osticket.com (besucht am 29.01.2022).

[6] Webkul Software Pvt Ltd. „UVdesk Open Source." (2022), Adresse: https://www.uvdesk.com/opensource (besucht am 29.01.2022).

[7] Symfony SAS. „Symfony." (2022), Adresse: https://www.symfony.com (besucht am 29.01.2022).

[8] Apache Software Foundation. „HTTP Server Project." (2022), Adresse: https://httpd.apache.org/ (besucht am 10.02.2022).

[9] F.EE GmbH. „Auch unterwegs optimal vernetzt." (2022), Adresse: https://www.factwork.de/factwork/produktion-zeiterfassung/factwork-mobile.html (besucht am 31.01.2022).

[10] SmartBear Software. „Bearer Authentication." (2022), Adresse: https://swagger.io/docs/specification/authentication/bearer-authentication/ (besucht am 02.02.2022).

[11] Microsoft Corporation. „Configuring Step 1: Install IIS and PHP." (2022), Adresse: https://docs.microsoft.com/iis/application-frameworks/scenario-build-a-php-website-on-iis/configuring-step-1-install-iis-and-php (besucht am 10.02.2022).

[12] Symfony SAS. „Security." (2022), Adresse: https://symfony.com/doc/current/security.html#remote-users (besucht am 10.02.2022).

[13] Symfony SAS. „User Providers." (2022), Adresse: https://symfony.com/doc/current/security/user_providers.html#creating-a-custom-user-providers (besucht am 10.02.2022).

[14] M. Duerst, L. Masinter und J. Zawinski. „The 'mailto' URI Scheme." (2010-10), Adresse: https://datatracker.ietf.org/doc/html/rfc6068 (besucht am 10.02.2022).

[15] Microsoft Corporation. „CredentialCache. DefaultCredentials Eigenschaft." (2022), Adresse: https://docs.microsoft.com/dotnet/api/system.net.credentialcache.defaultcredentials (besucht am 10.02.2022).

[16] Webkul Software Pvt Ltd. „Workfow." (2022), Adresse: https://www.uvdesk.com/features/workflow (besucht am 11.02.2022)

Autoren

Sebastian Janker (B.Sc.) studierte von 2018 bis 2022 Informatik an der OTH Regensburg im Rahmen eines dualen-Studiums bei der Firma F.EE. Seitdem ist er bei dem Unternehmen als Softwareentwickler beschäftigt.

Franz Laubmeier (Dipl.-Inf. (FH)) schloss 1992 sein Informatikstudium an der Fachhochschule Regensburg ab. Seitdem ist er bei der Firma F.EE beschäftigt. Anfangs als Softwareentwickler, seit 1998 als Leiter der Softwareentwicklung und nach der Gründung des Geschäftsbereichs Informatik + Systeme im Jahr 2000 als Leiter dieses Bereichs.

Professor Dr. Frank Herrmann wurde in Münster, Deutschland, geboren und studierte Informatik und Mathematik an der RWTH Aachen, wo er 1989 ein Diplom in Informatik verliehen bekam. Während seiner Zeit am Fraunhofer Institut IITB in Karlsruhe promovierte er 1996 über Ressourcenplanungsprobleme. Von 1996 bis 2003 arbeitete er bei der SAP AG in verschiedenen Funktionen, zuletzt als Direktor. Im Jahr 2003 wurde er Professor für Produktionslogistik an der Ostbayerische Technische Hochschule Regensburg. Er forscht an Algorithmen und Optimierungsmodellen für die operative Produktionsplanung und -steuerug.

Einsatz von Gamification zur Steigerung der Akzeptanz von digitalen Gesundheitsanwendungen (DiGA) - Ergebnisse einer qualitativen Studie

Lena Ulrich
Wirtschaftsinformatik
Hochschule für Technik und Wirtschaft Berlin
Treskowallee 8
10318 Berlin

Prof. Dr. Birte Malzahn
Wirtschaftsinformatik
Hochschule für Technik und Wirtschaft Berlin
Treskowallee 8
10318 Berlin
E-Mail:
birte.malzahn@htw-berlin.de

ABSTRACT

Digitale Gesundheitsanwendungen (DiGA) unterstützen u. a. die Therapie von Krankheiten oder Verletzungen und können unter bestimmten Bedingungen auf Rezept verordnet werden (Bundesinstitut für Arzneimittel und Medizinprodukte, 2023; Friesendorf & Lüttschwager, 2021). Für den Erfolg einer Therapie mittels DiGA ist jedoch eine beständig hohe Eigenmotivation der Patient*innen erforderlich. Der vorliegende Beitrag untersucht, welche Faktoren die Nutzerakzeptanz von DiGA beeinflussen, und insbesondere auch, welchen Einfluss Gamification in diesem Zusammenhang ausübt. Hierfür wurde ein theoretischer Bezugsrahmen aufgestellt, der mittels einer stichpunktbezogenen qualitativen Datenerhebung überprüft wurde. Die Ergebnisse zeigen, dass u. a. die wahrgenommene Nützlichkeit einen Einfluss auf die Nutzungsabsicht ausübt. Gamification erhöht zwar den Spaß an der Nutzung, dieser übt aber keinen starken Einfluss auf die Absicht zur Nutzung der untersuchten DiGA aus. Der Beitrag zeigt mögliche Konsequenzen der Ergebnisse für Forschung und Praxis auf

SCHLÜSSELWÖRTER

Gamification, Digitale Gesundheitsanwendungen (DiGA), Nutzerakzeptanz

1 EINLEITUNG

Digitale Gesundheitsanwendungen (DiGA) unterstützen u. a. die Therapie von Krankheiten oder Verletzungen (Bundesinstitut für Arzneimittel und Medizinprodukte, 2023). Wenn eine DiGA in das DiGA-Verzeichnis aufgenommen wurde, kann sie in Deutschland auf Rezept verordnet werden (Friesendorf & Lüttschwager, 2021). Bei allen medizinischen Behandlungen ist entscheidend, dass die Therapie konsequent durchgeführt wird (Hielscher, 2023). Da bei DiGA wenig Kontrollmöglichkeit für die Ärzt*innen hinsichtlich der Durchführung der Therapie besteht, müssen DiGA andere Ansätze verfolgen, um das Interesse an der Therapie bei den Patient*innen langfristig aufrecht zu erhalten (Hielscher, 2023). In einer im Jahr 2022 durchgeführten Befragung bewerteten Hausärzt*innen verordnete DiGA als nützlich und berichteten von positiven Versorgungseffekten (Wangler & Jansky, 2023). Als verbesserungsfähig wurden jedoch u. a. die Interaktivität und der Einsatz von Gamification-Elementen angesehen (Wangler & Jansky, 2023). Gamification-Elemente stellen eine Möglichkeit dar, die Einstellung der Nutzer*innen zu einer DiGA und damit zur Behandlung zu beeinflussen (Hielscher, 2023).

So ist möglich, dass Nutzer*innen ihre Therapie aufgrund der spielerischen Elemente effektiver durchführen als bei einer reinen Ausführung gesundheitsfördernder Übungen (Hielscher, 2023). Kurzfristig kann ein Anreiz z. B. durch erreichte Meilensteine sogar stärker zu zukünftigen therapeutischen Aktivitäten motivieren als das Therapieziel selbst (Hielscher, 2023). Ist die Akzeptanz einer DiGA dagegen zu gering, kann dies trotz therapeutischer Wirksamkeit einer DiGA zu einer geringen Nutzung bei den Patient*innen führen (Schlieter et al., 2024). Gamification kann ein vielversprechender Ansatz sein, um die Nutzung von DiGA nachhaltig zu fördern (Schlieter et al., 2024). Es mangelt jedoch noch an Wissen und Erfahrungen über Wirkungsweisen und Langzeiteffekte von Gamification-Elementen (Schlieter et al., 2024).

Der vorliegende Beitrag untersucht anhand eines theoretischen Bezugsrahmens, welche Faktoren die Nutzerakzeptanz von DiGA beeinflussen. Dabei liegt ein Fokus auf der Untersuchung des Einflusses von Gamification auf die Nutzerakzeptanz. Der Beitrag ist wie folgt aufgebaut: Zunächst werden die begrifflichen Grundlagen erläutert und ein kurzer Einblick in frühere Forschungsarbeiten zu Gamification in E-Health Anwendungen gegeben. Anschließend werden die Ergebnisse einer durchgeführten Marktstudie zur Integration von Gamification-Elementen in DiGA dargelegt.

Für die anschließende empirische Studie wurden zwei Prototypen entwickelt und eine qualitative Datenerhebung durchgeführt. Der Aufbau der Studie wird in Kapitel 4 beschrieben. Die Überprüfung der aufgestellten Hypothesen des theoretischen Bezugsrahmens wird in Kapitel 5 aufgezeigt. Die Arbeit schließt mit einem Fazit und Ausblick.

2 GRUNDLAGEN

2.1 Gamification

Gamification bezeichnet den Einsatz von Designelementen, die charakteristisch für Spiele sind, in Anwendungen, die nicht Teil eines Unterhaltungsspiels sind (Deterding et al., 2011). So werden z. B. Produkte oder Informationssysteme mit Spiel-Design-Elementen angereichert (Blohm & Leimeister, 2013), um u. a. die Produktivität und die Zufriedenheit von Nutzer*innen zu erhöhen (Stieglitz, 2015). Gängige Elemente sind u. a. das Erwerben von Punkten, Bestenlisten, Auszeichnungen, Abzeichen, Herausforderungen, das Erreichen von Leveln oder das Anzeigen des Fortschritts (Hamari et al., 2014). Gamification wird abgegrenzt vom Konzept der Serious Games. Dieser Begriff bezeichnet Spiele, die einem ernsten bzw. produktiven Ziel dienen, beispielsweise 3D-Computerprogramme, in denen Brände möglichst effektiv gelöscht werden sollen (Stieglitz, 2015).

2.2 Digitale Gesundheitsanwendungen

DiGA sind Medizinprodukte der Risikoklasse I (geringes Risiko) oder IIa (mittleres Risiko), die u. a. folgende Eigenschaften erfüllen (Bundesinstitut für Arzneimittel und Medizinprodukte, 2023): Die Anwendung unterstützt die Erkennung, Überwachung, Behandlung oder Linderung von Krankheiten oder die Erkennung, Behandlung, Linderung oder Kompensierung von Verletzungen oder Behinderungen. Der medizinische Zweck muss dabei wesentlich durch die digitale Hauptfunktion erreicht werden.

DiGA ermöglichen eine Versorgung von Patient*innen „remote“ und ohne notwendige menschliche Interaktion, potentiell rund um die Uhr (Friesendorf & Lüttschwager, 2021). Damit bieten sie u. a. eine kostengünstige Chance, der Unterversorgung strukturschwacher Regionen sowie dem Fachkräftemangel entgegenzuwirken (Friesendorf & Lüttschwager, 2021).

Der Erfolg des Konzepts ist jedoch maßgeblich davon abhängig, ob die DiGA verschrieben werden können („Apps auf Rezept“), denn nur dann ist ein ausreichender Absatz wahrscheinlich (Friesendorf & Lüttschwager, 2021). Hierfür ist die Aufnahme einer DiGA in das DiGA-Verzeichnis notwendig, was in Deutschland seit 2020 erfolgen kann (Friesendorf & Lüttschwager, 2021). Das jeweils aktuelle DiGA-Verzeichnis ist über die Seiten des Bundesinstituts für Arzneimittel und Medizinprodukte einsehbar (Bundesinstitut für Arzneimittel und Medizinprodukte, 2024). Essenziell für den Erfolg ist des Weiteren die Akzeptanz der Anwendungen bei Patient*innen und auch bei der Ärzteschaft, die die DiGA verordnen muss (Friesendorf & Lüttschwager, 2021).

2.3 Technology Acceptance Model (TAM)

Für die Untersuchung der Nutzerakzeptanz von IT-Anwendungen haben sich das TAM (Davis, 1986) und dessen Weiterentwicklungen (z. B. Venkatesh & Davis (2000), Venkatesh et al. (2003), Mailizar et al. (2021)) bewährt. Folgende fünf Konstrukte wurden für die Studie dieses Beitrags übernommen:

- Wahrgenommene Nützlichkeit: Dieses Konstrukt bezieht sich darauf, wie sehr ein Individuum glaubt, dass die Nutzung eines bestimmten Systems seine / ihre Leistung verbessert (Davis, 1986).
- Leichtigkeit der Nutzung: Dieses Konstrukt beschreibt das Ausmaß, zu dem ein Individuum davon ausgeht, dass die Nutzung eines bestimmten Systems ohne körperliche oder geistige Anstrengung möglich ist (Davis, 1986).
- Subjektive Norm: Aus der Theory of Reasoned Action (TRA) (Fishbein & Ajzen, 1975) übernommen; beschreibt die Wahrnehmung einer Person, dass die meisten Menschen, die für sie wichtig sind, denken, dass sie ein bestimmtes Verhalten ausüben oder nicht ausüben sollte (Venkatesh & Davis, 2000).
- Wahrgenommener Spaß: Ausmaß, in dem die Benutzung eines bestimmten Systems an sich als vergnüglich empfunden wird, unabhängig von Auswirkungen auf die Leistung, die sich aus der Nutzung des Systems ergeben (Davis et al., 1992; Venkatesh, 2000).
- Nutzungsabsicht: Angelehnt an das Konstrukt Verhaltensabsicht der TRA (Fishbein & Ajzen, 1975); beschreibt die Absicht eines Individuums, ein bestimmtes System zu benutzen (Davis et al., 1989). Die Absicht eines Individuums, sich in einer bestimmten Weise zu verhalten, ist laut TRA maßgeblich für ein resultierendes Verhalten (Davis et al., 1989).

2.4 Gamification in E-Health-Anwendungen

Der Einsatz von Gamification im Gesundheitsbereich wird seit einigen Jahren erforscht (Sardi et al., 2017). Sardi et al. (2017) zeigen in ihrer Metastudie, dass Forschungsarbeiten v. a. kurzfristige Effekte von Gamification untersucht haben. Hurmuz et al. (2022) analysieren die Nutzung einer gamifizierten E-Health-Anwendung bei älteren Personen auf Basis des Technology Acceptance Models und zeigen u. a., dass der Spaß an der Nutzung die wahrgenommene Nützlichkeit beeinflusst und dass die wahrgenommene Nützlichkeit die Intention beeinflusst, die Anwendung weiterhin zu benutzen. Damaševičius et al. (2023) weisen in ihrer Metastudie darauf hin, dass einige Studien positive Effekte von Gamification / Serious Games feststellen, andere dagegen keine Effekte nachweisen können. Sie fordern daher, die Wirkungsweisen weiter zu untersuchen, insbesondere auch bei unterschiedlichen Bevölkerungsgruppen.

3 MARKTSTUDIE ZUM EINSATZ VON GAMIFICATION IN DIGITALEN GESUNDHEITS-ANWENDUNGEN

In einer Marktstudie wurden die im September 2023 im DiGA-Verzeichnis (Bundesinstitut für Arzneimittel und Medizinprodukte, 2024) dauerhaft (24) und vorläufig aufgenommenen (25) DiGA auf das Vorhandensein von Gamification-Elementen überprüft (Ulrich, 2023). Die Analyse erfolgte dabei hauptsächlich über die Informationen, die über die Webseite der Hersteller bzw. der jeweiligen DiGa verfügbar waren. Für einige DiGA waren die Informationen auf der Webseite des Anbieters jedoch nicht ausreichend, um sie bezüglich integrierter Gamification-Elemente zu untersuchen, so dass für diese DiGA keine entsprechende Analyse möglich war (Ulrich, 2023).
Bei sieben DiGA konnten Gamification-Elemente ermittelt werden (Ulrich, 2023):

- Smoke Free: Punkte, Trophäen, Herausforderungen, Minispiel
- Kaia (Rückentraining): Fortschrittsbalken
- Mawendo (Physiotherapie): Fortschrittsbalken, Level
- NeuroNation MED (Training bei kognitiven Beeinträchtigungen): Punkte
- NichtraucherHelden-App: Status
- re.flex (Therapie bei Kniearthrose): Fortschrittsbalken
- Vitadio (Therapie bei Typ2-Diabetes): Herausforderungen

Insgesamt zeigte sich, dass Gamification bei den untersuchten DiGA nicht sehr ausgeprägt waren. Die DiGA mit den meisten Gamificationelementen war die DiGA SmokeFree.

4 AUFBAU DER STUDIE

4.1 Forschungsmethodik und Hypothesen

Da es sich bei Gamification von DiGA um ein relativ neues Forschungsgebiet handelt, wurde eine qualitative Forschungsmethodik gewählt. Anhand von Hypothesen wurde ein theoretischer Bezugsrahmen aufgestellt, in den Konstrukte und Beziehungen aus bestätigten Modellen der Akzeptanzforschung (s. Kapitel 2.3) übernommen wurden (Ulrich, 2023). Das Modell der Studie ist in Abbildung 1 veranschaulicht.
H1: Die wahrgenommene Nützlichkeit hat einen positiven Einfluss auf die Nutzungsabsicht (Venkatesh, 2000), (Venkatesh & Davis, 2000).
H2: Die Leichtigkeit der Nutzung hat einen positiven Einfluss auf die Nutzungsabsicht (Venkatesh, 2000), (Venkatesh & Davis, 2000).
H3: Die Leichtigkeit der Nutzung hat einen positiven Einfluss auf die wahrgenommene Nützlichkeit (Davis, 1986), (Venkatesh & Davis, 2000).
H4: Die subjektive Norm hat einen positiven Einfluss auf die Nutzungsabsicht (Venkatesh & Davis, 2000).
H5a: Der Einsatz von Gamification erhöht den wahrgenommenen Spaß an der Nutzung einer DiGA (Borrás-Gené et al., 2019) (Lounis et al., 2014).
H5b: Der wahrgenommene Spaß hat einen positiven Einfluss auf die Nutzungsabsicht (Davis et al., 1992) (Borrás-Gené et al., 2019).

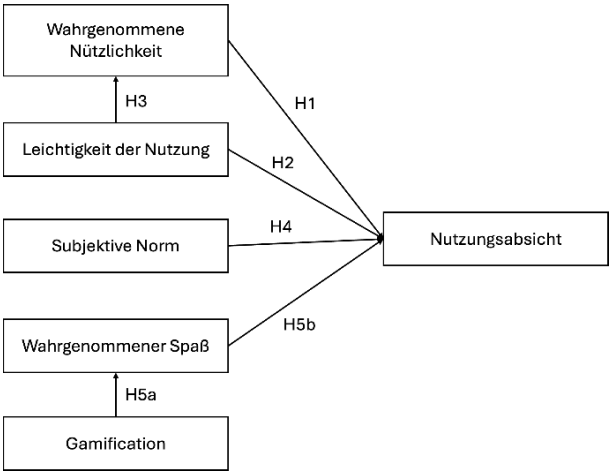


Abbildung 1 Modell der Studie (Eigene Darstellung)

Anhand von Interviews in Kombination mit einem Prototypentest durch Probandinnen wurden qualitative Daten erhoben (Ulrich, 2023), die anschließend mittels skalierender Strukturierungstechnik (Mayring, 2015) ausgewertet wurden. Die Ergebnisse wurden anschließend dem vorab erstellten Bezugsrahmen gegenübergestellt (Ulrich, 2023).

4.2 Entwicklung der Prototypen

In Zusammenarbeit mit der akquinet tech@spree GmbH und der G. Pohl-Boskamp GmbH & Co. KG wurde eine DiGA für Belastungsinkontinenzbeschwerden für die Zielgruppe biologisch anatomischer Frauen ab 50 Jahren entwickelt (Ulrich, 2023). Für die Studie wurden zwei Prototypen dieser DiGA mittels der Software Figma (Figma GmbH, 2024) erstellt: Prototyp A beinhaltete vier Gamification-Elemente: Punkte, Trophäen, Status und Fortschrittsanzeige. Prototyp B enthielt keine Gamification-Elemente. Jeweils ein Beispiel für die beiden Prototypen ist in Abbildung 2 zu sehen.

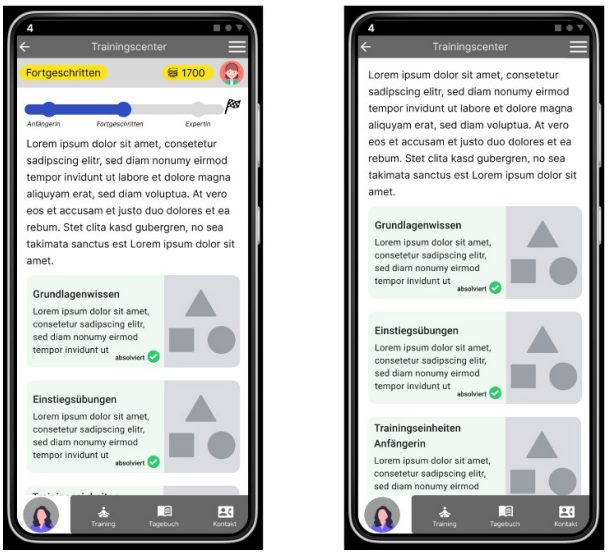


Abbildung 2 Prototypen mit / ohne Gamification (Screenshots)

4.3 Durchführung der Studie

Die Datenerhebung wurde zwischen dem 28.07.2023 und dem 19.08.2023 einzeln mit insgesamt 14 Frauen im Alter von 48 bis 72 Jahren im Rahmen einer qualitativen Studie durchgeführt. Die Probandinnen wurden vorab per Zufallsprinzip in zwei Testgruppen eingeteilt: Gruppe A arbeitete mit Prototyp A. Für Gruppe B wurde der Prototyp B ohne Gamification-Elemente eingesetzt.
Zu Beginn erhielt jede Teilnehmerin eine kurze Einführung in das Thema, um sicherzustellen, dass der Zweck der DiGA verstanden wurde. Anschließend wurden den Teilnehmerinnen mittels eines Interviewleitfadens einige allgemeine Fragen über die Erwartungen und Befürchtungen bezüglich der DiGA sowie Fragen zur allgemeinen Nutzungsabsicht der Anwendung gestellt. Die Antworten wurden mit Hilfe eines Mobiltelefons aufgezeichnet. Im Anschluss erhielten die Probandinnen ein Google Pixel Mobiltelefon, auf dem der für sie zufällig zugeteilte Prototyp abspielbar war. Die Probandinnen führten dann mehrere Aufgaben in einer festgelegten Reihenfolge mit dem zugeteilten Prototyp aus. Anschließend wurden den Probandinnen mittels Leitfadens Fragen zu den Konstrukten wahrgenommene Nützlichkeit, Leichtigkeit der Nutzung, wahrgenommener Spaß, subjektive Norm und zur Nutzungsabsicht (ex post) gestellt. Das Gespräch wurde ebenfalls aufgezeichnet. Die Fragen des Leitfadeninterviews zu den Konstrukten wurden auf Basis früherer Forschungsarbeiten (s. Kapitel 2.3) erarbeitet. Da es sich im Gegensatz zu diesen Arbeiten nicht um eine quantitative, sondern um eine qualitative Studie handelt, wurden die Fragen entsprechend angepasst. Beispielsweise wurde die auf einer Skala zu bewertende Aussage „Es macht mir Spaß, das System zu benutzen.“ in die offene Frage „Inwieweit hat die Nutzung der App dir Spaß bereitet?“ umformuliert (Ulrich, 2023).

4.4 Auswertung der Studie

Die Äußerungen der Probandinnen wurden transkribiert. Zur qualitativen Analyse der Interviews wurde die Software MAXQDA (VERBI – Software. Consult. Sozialforschung. GmbH, 2024) genutzt. Bei der Analyse wurde die skalierende Strukturierungstechnik (Mayring, 2015) verwendet, um die Aussagen der Probandinnen in deduktive Kategorien zu codieren, die aus dem Modell dieser Studie abgeleitet wurden. Anschließend wurden die Inhalte bezüglich der untersuchten Konstrukte auf einer Ordinalskala bewertet (Ulrich, 2023). Es wurden die folgenden drei Ausprägungen verwendet: "+" für positive Aussagen, "o" für neutrale Aussagen und "-" für negative Aussagen (s. auch Kapitel 5.2).

5 ERGEBNISSE

5.1 Auswertung der qualitativen Daten

Tabelle 1 zeigt die Bewertung der Konstrukte nach Durchführung des Tests. Die Ergebnisse der beiden Gruppen zeigen bei den Konstrukten wahrgenommene Nützlichkeit, Leichtigkeit der Nutzung, subjektive Norm und Nutzungsabsicht (ex post) keine gravierenden Unterschiede auf. Eine deutliche Abweichung ist jedoch beim Konstrukt wahrgenommener Spaß erkennbar. Dieses Konstrukt wird von den Probandinnen der Gruppe A (mit Gamification) deutlich besser bewertet als von den Probandinnen der Gruppe B (ohne Gamification).

Gruppe A: mit Gamification

Kategorie	P2	P3	P4	P7	P9	P10	P14
Wahrgenommene Nützlichkeit	+	+	+	+	o	+	+
Leichtigkeit der Nutzung	+	+	o	+	+	+	+
Wahrgenommener Spaß	+	+	o	+	o	+	+
Subjektive Norm	+	+	+	+	o	+	+
Nutzungsabsicht (ex post)	o	+	+	+	o	+	+

Gruppe B: ohne Gamification

Kategorie	P1	P5	P6	P8	P11	P12	P13
Wahrgenommene Nützlichkeit	+	+	+	+	-	+	+
Leichtigkeit der Nutzung	+	o	+	o	+	+	+
Wahrgenommener Spaß	+	o	o	-	o	o	-
Subjektive Norm	+	+	+	+	+	+	+
Nutzungsabsicht (ex post)	+	+	+	o	o	+	+

Tabelle 1 Bewertung der Konstrukte nach Durchführung des Tests (Eigene Darstellung)

5.2 Überprüfung der Hypothesen

Die Überprüfung aller Hypothesen außer H5a erfolgte für beide Gruppen zusammen. Es wurde für jede dieser Hypothese ermittelt, zu welchem Anteil die Bewertung der beiden Konstrukte der jeweiligen Hypothese über-

einstimmen (z. B. positiv-positiv, neutral-neutral, negativ-negativ). Bei einem Anteil von über 75 Prozent an Übereinstimmung wurde die Hypothese als stark bestätigt angesehen. Ein Anteil von 51 bis 75 Prozent wurde als Bestätigung der Hypothese angesehen. Bei einem Anteil von 26 bis 50 Prozent wurde die Hypothese als schwach bestätigt gewertet. Auf eine Darstellung der Ergebnisse mittels Kreuztabellen wurde in diesem Beitrag aus Platzgründen verzichtet.

H1: Die wahrgenommene Nützlichkeit hat einen positiven Einfluss auf die Nutzungsabsicht.
Mehrere Probandinnen bestätigen den Zusammenhang: *„Der Nutzen ist, dass es mir besser geht. Also ich nutze die App, damit ich eine Verbesserung in meinem Alltag, in meinem Leben habe (...).“ (Prob. 10, Pos. 31)*
„(...) Ich denke, der Leidensdruck muss groß genug sein, damit man das macht. (...)“ (Prob. 9, Pos. 12)
„Ja, ich würde das nutzen, weil es mir was bringen würde. (...).“ (Prob. 10, Pos. 65)
Bei 11 von 14 Probandinnen (79 Prozent) stimmt die Bewertung der Konstrukte überein. Die Hypothese kann somit als stark bestätigt angesehen werden.

H2: Die Leichtigkeit der Nutzung hat einen positiven Einfluss auf die Nutzungsabsicht.
Zwei Probandinnen stellen den Zusammenhang zwischen *Leichtigkeit der Nutzung* und *Nutzungsabsicht* her, verweisen dabei jedoch auch auf den Aspekt der Nützlichkeit.
„Ja, ich würde das nutzen, weil es mir was bringen würde. Sie ist für mich einfach und leicht zu bedienen. Und dadurch würde ich sie schon nutzen, ja. (...).“ (Prob. 10, Pos. 65)
„(...)Das soll einfach nur einfach sein, bedienerfreundlich und funktionieren. Mehr nicht.“ (Prob. 13, Pos. 44)
Bei neun Probandinnen (64 Prozent) stimmt die Bewertung der Konstrukte überein. Die Hypothese kann somit als bestätigt angesehen werden.

H3: Die Leichtigkeit der Nutzung hat einen positiven Einfluss auf die wahrgenommene Nützlichkeit.
Bei neun Probandinnen (64 Prozent) stimmt die Bewertung der Konstrukte überein. Die Hypothese kann somit bestätigt werden.

H4: Die subjektive Norm hat einen positiven Einfluss auf die Nutzungsabsicht.
Insgesamt stimmt bei 11 Probandinnen (79 Prozent) die Bewertung der Konstrukte überein. Die Hypothese kann somit als stark bestätigt angesehen werden.

H5a: Der Einsatz von Gamification erhöht den wahrgenommenen Spaß an der Nutzung einer DiGA.
Zur Auswertung dieser Hypothese wurden alle positiven Bewertungen des *wahrgenommenen Spaßes* mit 1 bewertet, neutrale Bewertungen mit 0 und negative Bewertungen mit -1. Insgesamt zeigt sich eine deutliche Abweichung der durchschnittlichen Bewertung des *wahrgenommenen Spaßes* bei den beiden Gruppen: Für Gruppe A (mit Gamification) ergibt sich ein Durchschnittswert von 0,71, für Gruppe B (ohne Gamification) ein Durchschnittswert von -0,14.
„(...) Ein bisschen Freude ist auch mit dabei, wenn man sieht, man erreicht Pokale, man erzielt Geld, man kann sich ein neues Profil kaufen und so. Es ist nicht verkehrt.“ (Prob. 7, Pos. 20)
Es gibt jedoch auch gegenteilige Aussagen: *„Eher welchen, die sehr technikaffin sind, die viel mit solchen Belobigungen und so arbeiten. Für jemanden, der das sonst nicht so in seinem Alltag hat, finde ich es befremdlich.“ (Prob. 9, Pos. 44)*
Auch wenn Gamification nicht alle Probandinnen anzusprechen scheint, kann geschlussfolgert werden, dass der Einsatz von Gamification durchschnittlich zu einer Erhöhung des *wahrgenommenen Spaßes* führt. Aufgrund der deutlichen Abweichung der Ergebnisse in den beiden Gruppen wird die Hypothese als stark bestätigt gewertet.

H5b: Der wahrgenommene Spaß hat einen positiven Einfluss auf die Nutzungsabsicht (ex post).
Mehrere Probandinnen verneinen einen Zusammenhang zwischen den Aspekten.
„(...) Ja, natürlich ist so ein Belohnungssystem dahinter. So Konfetti und Hurra, ich habe noch eine Trophäe jetzt und ich kann mir noch einen Avatar kaufen und so. Das finde ich, dass ist mir persönlich ein bisschen zu verspielt. Weil es ja kein Spiel letztendlich ist. (...)“ (Prob. 4, Pos. 38)
„*Dass ich wieder 100 Punkte habe, das ist der Unterhaltungswert, den ich für eine Gesundheitsapp nicht brauche*.“ (Prob. 9, Pos. 40)
Insgesamt stimmt bei sieben Probandinnen (50 Prozent) die Bewertung der Konstrukte überein. Die Hypothese kann somit nur als schwach bestätigt angesehen werden.

5.3 Diskussion der Ergebnisse

Die Hypothesen des Modells können anhand der qualitativen Daten – wenn auch mit unterschiedlicher Stärke - als bestätigt angesehen werden. In Übereinstimmung mit früheren Studien zeigt u. a. die wahrgenommene Nützlichkeit einen starken Einfluss auf die Nutzungsabsicht (ex post). Interessanterweise hat sich die Hypothese, die für die Untersuchung des Einflusses von Gamification am wichtigsten ist, nur schwach bestätigt. Zwar weisen die vorliegenden Daten darauf hin, dass Gamification den wahrgenommenen Spaß durchschnittlich deutlich erhöht. Jedoch konnte nur ein schwacher Einfluss des wahrgenommenen Spaßes auf die Nutzungsabsicht (ex post) festgestellt werden. Die Ergebnisse deuten darauf hin, dass eine DiGA und das damit verbundene gesundheitliche Problem von einigen Probandinnen als eine ernste Angelegenheit angesehen werden, bei der ein Spaßfaktor als unangemessen empfunden wird. Als Konsequenz ließe sich ableiten, dass der Einsatz von Gamification nicht unbedingt zu einer stärkeren Nutzung von DiGA führt und somit in diesem Kontext eventuell sogar verzichtbar wäre.

Bei der Interpretation der Ergebnisse ist jedoch zu berücksichtigen, dass der Test der DiGA und die Befragung der Probandinnen nur zum Zeitpunkt der initialen Nutzung stattgefunden hat. Es ist denkbar, dass sich bei einer Untersuchung der Nutzungsabsicht nach verschiedenen Nutzungszeiträumen (Venkatesh & Davis, 2000) andere Ergebnisse ergeben würden. Eine Studie zeigt, dass nach einiger Zeit ein Gewöhnungseffekt an Gamificationelemente auftreten kann, der zu positiven Auswirkungen führen kann (Rodrigues et al., 2022). Die Langezeitwirkung des Einsatzes von Gamification bei DiGA sollte somit einer näheren Untersuchung unterzogen werden. Ob sich die Ergebnisse auf andere Anwendungskontexte übertragen lassen, muss in weiteren Studien untersucht werden. Die Zielgruppe der untersuchten DiGA waren Frauen ab 50 Jahren. Es ist denkbar, dass das Alter oder das Geschlecht der Probandinnen einen Einfluss auf die Ergebnisse haben (Koivisto & Hamari, 2014). Die Ergebnisse könnten zudem mit der Art und dem Einsatzgebiet der DiGA variieren. Auch das Forschungsdesign führt zu Limitationen der Aussagekraft der Ergebnisse. Aufgrund der geringen Teilnehmerzahl qualitativer Studien sind keine weitergehenden statistischen Auswertungen möglich. Die qualitative Studie kann jedoch als Vorstudie wichtige Hinweise darauf geben, welche Aspekte in einer quantitativen Studie weiter untersucht werden sollten.

6 FAZIT UND AUSBLICK

Die Ergebnisse der Studie deuten darauf hin, dass u. a. die wahrgenommene Nützlichkeit einen maßgeblichen Einfluss auf die Nutzungsabsicht der Probandinnen hat. Gamification übt zwar einen Einfluss auf den wahrgenommenen Spaß der Nutzung aus. Jedoch scheint der wahrgenommene Spaß zum Zeitpunkt der Untersuchung nur einen untergeordneten Einfluss auf die Nutzungsabsicht auszuüben. Weitere quantitative Studien sind notwendig, um die Auswirkungen von Gamification auf die Nutzung von DiGA zu untersuchen. Neben der Berücksichtigung von weiteren Nutzungszeitpunkten sollten auch andere DiGA und weitere Zielgruppen untersucht werden. Für Anbieter von DiGA ist eine kontinuierliche Nutzungsabsicht bzw. Nutzung durch die Anwender*innen unerlässlich, um dauerhaft am Markt bestehen können. Sie sollten daher weiter die Einflussfaktoren auf die Nutzungsabsicht von DiGA untersuchen.

LITERATUR

Blohm, I., & Leimeister, J. M. (2013). Gamification: Gestaltung IT-basierter Zusatzdienstleistungen zur Motivationsunterstützung und Verhaltensänderung. WIRTSCHAFTSINFORMATIK, 55(4), 275–278. https://doi.org/10.1007/s11576-013-0368-0

Borrás-Gené, Martínez-Núñez, & Martín-Fernández. (2019). Enhancing Fun Through Gamification to Improve Engagement in MOOC. Informatics, 6(3), 28. https://doi.org/10.3390/informatics6030028

Bundesinstitut für Arzneimittel und Medizinprodukte. (2023). Das Fast-Track-Verfahren für digitale Gesundheitsanwendungen (DiGA) nach § 139e SGB V - Ein Leitfaden für Hersteller, Leistungserbringer und Anwender. https://www.bfarm.de/SharedDocs/Downloads/DE/Medizinprodukte/diga_leitfaden_aenderung_markiert.pdf?blob=publicationFile

Bundesinstitut für Arzneimittel und Medizinprodukte. (2024). DiGA-Verzeichnis. https://diga.bfarm.de/de

Damaševičius, R., Maskeliūnas, R., & Blažauskas, T. (2023). Serious Games and Gamification in Healthcare: A Meta-Review. Information, 14(2), 105. https://doi.org/10.3390/info14020105

Davis, Fred. D. (1986). A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Result. MIT Sloan School of Management.

Davis, Fred. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Management Science, 35(8), 982–1003.

Davis, Fred. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and Intrinsic Motivation to Use Computers in the Workplace. Journal of Applied Social Psychology, 22(14), 1111–1132.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining „Gamification“. Proceedings of the 15th International Academic MindTrek Conference, 9–15. https://doi.org/10.1145/2181037.2181040

Figma GmbH. (2024). Figma. https://www.figma.com/de/

Fishbein, M., & Ajzen, I. (1975). Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley.

Friesendorf, C., & Lüttschwager, S. (2021). Digitale Gesundheitsanwendungen: Assessment der Ärzteschaft zu Apps auf Rezept. Springer.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification Work? —A Literature Review of Empirical Studies on Gamification. Proceedings of the 47th Hawaii International Conference on System Sciences, 3025–3034.

Hielscher, B. (2023, September 1). Therapieadhärenz DiGA und Therapietreue. Healthcare Digital. https://www.healthcare-digital.de/diga-und-therapietreue-a-e0d555f245b1b076c0539e2b7b5a0f92/

Hurmuz, M. Z., Jansen-Kosterink, S. M., Hermens, H. J., & Van Velsen, L. (2022). Game not over: Explaining older adults' use and intention to continue using a gamified eHealth service. Health Informatics Journal, 28(2), 146045822211060. https://doi.org/10.1177/14604582221106008

Koivisto, J., & Hamari, J. (2014). Demographic Differences in Perceived Benefits from Gamification. Computers in Human Behavior, 35, 179–188. https://doi.org/10.1016/j.chb.2014.03.007

Lounis, S., Pramatari, K., & Theotokis, A. (2014). Gamification is all about Fun: The Role of Incentive Type and Community Collaboration. ECIS 2014 Proceedings, 1–15.

Mailizar, M., Burg, D., & Maulina, S. (2021). Examining University Students' Behavioural Intention to Use E-Learning during the COVID-19 Pandemic: An Extended TAM Model. Education and Information Technologies, 26(6), 7057–7077. https://doi.org/10.1007/s10639-021-10557-5

Mayring, P. (2015). Qualitative Inhaltsanalyse: Grundlagen und Techniken (12. Auflage). Beltz Verlag.

Rodrigues, L., Pereira, F. D., Toda, A. M., Palomino, P. T., Pessoa, M., Carvalho, L. S. G., Fernandes, D., Oliveira, E. H. T., Cristea, A. I., & Isotani, S. (2022). Gamification Suffers from the Novelty Effect but Benefits from the Familiarization Effect: Findings from a Longitudinal Study. International Journal of Educational Technology in Higher Education, 19(1). https://doi.org/10.1186/s41239-021-00314-6

Sardi, L., Idri, A., & Fernández-Alemán, J. L. (2017). A systematic review of gamification in e-Health. Journal of Biomedical Informatics, 71, 31–48. https://doi.org/10.1016/j.jbi.2017.05.011

Schlieter, H., Kähling, M., Hickmann, E., Fürstenau, D., Sunyaev, A., Richter, P., Breitschwerdt, R., Thielscher, C., Gersch, M., Maaß, W., Reuter-Oppermann, M., & Wiese, L. (2024). Digitale Gesundheitsanwendungen (DiGA) im Spannungsfeld von Fortschritt und Kritik: Diskussionsbeitrag der Fachgruppe „Digital Health" der Gesellschaft für Informatik e. V. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 67(1), 107–114. https://doi.org/10.1007/s00103-023-03804-2

Stieglitz, S. (2015). Gamification – Vorgehen und Anwendung. HMD Praxis der Wirtschaftsinformatik, 52(6), 816–825. https://doi.org/10.1365/s40702-015-0185-6

Ulrich, L. (2023). Untersuchung des Einsatzes von Gamification zur Steigerung der Wirksamkeit einer digitalen Gesundheitsanwendung [Masterarbeit]. HTW Berlin.

Venkatesh, V. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. Information Systems Research, 11(4), 342–365.

Venkatesh, V., & Davis, Fred. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Management Science, 46(2).

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, Fred. D. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3).

VERBI – Software. Consult. Sozialforschung. GmbH. (2024). MAXQDA. https://www.maxqda.com/de/

Wangler, J., & Jansky, M. (2023). Digitale Gesundheitsanwendungen (DiGA) in der Primärversorgung – Erfahrungen und Beobachtungen von Hausärzt*innen hinsichtlich der Anwendung von DiGA. Prävention und Gesundheitsförderung, 18(4), 483–491. https://doi.org/10.1007/s11553-022-00988-4

The Impact of Object-Centric Process Mining on Business Efficiency: A Case Study Approach

Neža Pintarič
University of Ljubljana Kardeljeva ploščad 17 1000 Ljubljana np7755@student.uni-lj.si

Frank Morelli
Pforzheim University Tiefenbronner Straße 65 75175 Pforzheim frank.morelli@hs-pforzheim.de

Amira Elkanawati
University of Ljubljana Kardeljeva ploščad 17 1000 Ljubljana

Anton Manfreda
University of Ljubljana Kardeljeva ploščad 17 1000 Ljubljana anton.manfreda@ef.uni-lj.si

Przemyslaw Radziszewski
Kraków, Poland p.radzizewski@accenture.com

KEYWORDS

Object-centric process mining (OCPM), process mining, business process optimization, business efficiency, Celonis, business process management, data modelling

ABSTRACT

Although traditional process mining techniques have enabled substantial enhancements in process transpa-rency and optimization, they frequently prove inadequate for capturing the intricacies of actual business processes, which are characterized by multiple interacting entities. OCPM addresses this limitation by allowing a more nuanced and multi-dimensional analysis of processes. By integrating both traditional and object-centric approaches, the study provides a comparative analysis of the performance and utility of each method, highlighting the specific benefits of OCPM in improving process transparency and operational efficiency. The findings reveal that OCPM offers a more comprehensive understanding of complex business operations, which is crucial for enhancing decision-making and achieving better performance outcomes. Additionally, the study outlines the challenges and opportunities of implemen-ting OCPM in real-world settings, emphasizing its practical implications for business process optimization.

INTRODUCTION

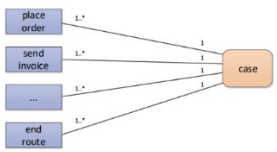
The continuously changing landscape of modern business processes requires a commitment to continuous improvement in operational efficiency and effectiveness. To remain competitive and adapt to fast changing market needs, organizations are increasingly relying on innovative methodologies and technologies. In this context, process mining has emerged as a crucial tool, enabling organizations to extract valuable insights from event logs and enhance their workflows (W. Van Der Aalst et al., 2012). The conventional approach to process mining involves analyzing the sequence of activities that comprise a particular business process, thereby providing a snapshot of how tasks are executed. Notwithstanding, the complex nature of contemporary business operations, which often encompass a multitude of interrelated entties, necessitates a more nuanced approach.

The emergence of object-centric process mining represents a big leap forward in this field as it enables detailed analysis of business processes at granular level from several aspects. By meeting interactions between different objects like orders, invoices, and deliveries it provides a wider view of process dynamics than before. (Bolt et al., 2016).

In recent years, PM has undergone a significant evolution, progressing from traditional approaches to more advanced techniques that can effectively address the complexities inherent to modern business environ-ments. One of the most promising developments in this field is object-centric process mining (OCPM). In contrast to the traditional PM approach, which is concerned with a single case, OCPM permits the examination of processes that involve multiple interacting objects and it gives an augmented view on interactions between processes.

The traditional approach to process mining is based on the premise that each event log is associated with a single case notion. This reductionist approach frequently proves inadequate for capturing the complex, multifaceted nature of real world processes, where an event may be associated with multiple cases or objects. To illustrate, an order handling process may entail the participation of multiple related entities, including orders, items, packages, and route (W. M. P. Van Der Aalst, 2019).

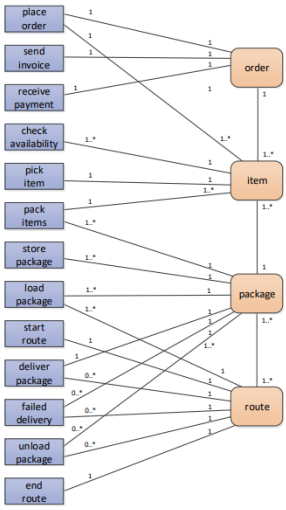
Figure 1: Concept of a single case notion in classical process mining



Source: W. M. P. Van Der Aalst, (2019)

This assumption results in the flattening of event data, which can obscure the true complexity of the process and result in the loss of crucial information. The generated models may not accurately reflect the intricacies of the actual processes, which may result in incomplete or misleading insights. (W. M. P. Van Der Aalst, 2019).

Figure 2: Overview of the interconnections between activities and object types in everyday life



Source: W. M. P. Van Der Aalst, (2019)

The objective of classical PM is to identify the sequence of events within log cases. Two principal issues may impinge upon the process of analysis. In real-world scenarios, events often pertain to multiple cases (convergence) or recur multiple times within a single case (divergence). The traditional PM approach, which assumes that each event belongs to a single, unique case, is ill-equipped to handle these complexities. Convergence results in the repetition of events across multiple cases, which increases the number of events and distorts the true nature of the process behavior. Conversely, divergence gives rise to the formation of artificial loops within the process model, whereby repeated events within the same case are erroneously interpreted as occurring in a sequential rather than simultaneous manner. Such misinterpretation can result in process models that are either overly or insufficiently fit for purpose, thereby reducing their accuracy and utility (W. M. P. Van Der Aalst, 2019).

Complex procedures are consolidated into single-case event logs, which results in the loss of crucial relational data. In cases where multiple objects are involved in a process, their interactions have the potential to exert a noteworthy influence on the eventual outcome. The application of traditional PM techniques often reduces data complexity, resulting in the intricate contextual relationships between different objects being lost. The aforementioned flattening process has the potential to result in inaccurate analyses and suboptimal decision-making, as it fails to take into account the nuances of the process dynamics (W. M. P. Van Der Aalst, 2019).

A further significant drawback of traditional PM is its lack of scalability. As the quantity of event data increases, traditional PM algorithms can become computationally intensive, which may result in performance bottlenecks. This is particularly problematic in large organizations with high-frequency processes and extensive event logs, where the issue is compounded. The necessity for the adequate handling and processing of large datasets is of paramount importance, and traditional PM techniques frequently prove inadequate in this regard. The single-case notion serves to exacerbate this issue, requiring the undertaking of separate analyses for each case type, as opposed to allowing for a holistic analysis that considers all relevant entities and their interactions. This can result in the repetition of computations and an increase in processing time, which further constrains the scalability of traditional PM approaches (Adams & van der Aalst, 2021).

In addition to the aforementioned methodological limitations, traditional PM also encounters difficulties in integrating disparate data sources. In the contemporary business environment, data is often generated from many sources, including transactional systems, Internet of Things (IoT) devices, and social media. The traditional PM techniques, which are primarily designed to handle structured event logs from transactional systems, frequently prove inadequacy for the effective integration and analysis of these heterogeneous data sources. This limitation constrains the capacity of organizations to obtain extensive insights from their process data (W. M. P. Van Der Aalst, 2019).

In conclusion, although conventional PM has yielded notable benefits in enhancing process efficiency and compliance, it is constrained by a number of critical limitations. These include its reliance on single-case event logs, the loss of relational information through data flattening, inadequacy in handling concurrency and multiobject interactions, scalability challenges, and difficulties in integrating diverse data sources. The advancement of methodologies is essential to address the aforementioned limitations. One promising avenue is the development of object-centric process mining, which can provide a more inclusive and accurate view of complex business processes (Adams et al., 2022).

The main aim of this paper is to study the influence of object centric process mining on business efficiency, especially in relation to Order to Cash (O2C) and Purchase to Pay (P2P) processes. The O2C process is essential for businesses because it covers the entire sequence of events, from receiving orders to collecting payments. The optimization of this procedure enables organizations to enhance cash flow, decrease operational expenses, and enhance customer satisfaction, as stated by Van Der Aalst et al. (2012). The research seeks to address the following questions:

1. *How can object-centric process mining improve the efficiency and effectiveness of the Order to Cash process compared to traditional process mining methods?*
2. *What are the specific benefits and challenges encountered when implementing object-centric process mining in a real-world context using Celonis?*

This paper distinguishes itself by integrating key theoretical concepts related to Object-Centric Process Mining (OCPM), providing a solid foundation for developing a mockup dashboard tailored to analyze interconnected processes. It builds upon state-of-the-art methodologies, including advancements in event log standards like OCEL and tools for multi-object process visualization. By leveraging these foundations, the OCPM analysis aims to demonstrate the potential of OCPM in improving process transparency, optimizing workflows, and enabling better decision-making.

The creation of this paper was guided by the seven principles of Design Science Research (DSR), ensuring a structured and iterative approach to its development and evaluation. While adhering to the core tenets of DSR, this work embraces flexibility, as suggested by Hevner et al. (2004), to focus on practical outcomes rather than rigid procedural adherence. By balancing systematic exploration with pragmatic design, this paper under-scores the role of OCPM in addressing real-world challenges and highlights the potential for further research and implementation of such innovative IT artifacts.

OBJECT CENTRIC PROCESS MINING

Introduction to Object Centric Process Mining

OCPM is a significant step forward in the field of process mining. It overcomes constraints of traditional techniques by focusing on the interactions and relationships between multiple objects within complex business processes. The predominant methodologies employed in PM are largely based on the analysis of single-case event logs, wherein each event is associated with a specific object or case. This approach is frequently inadequate for capturing the multifaceted nature of real-world processses, particularly in contexts such as Enterprise Resource Planning (ERP) systems, where events are associated with a range of objects, including orders, items, invoices, and shipments. The multifaceted nature of these interactions necessitates a more comprehensive approach to accurately reflect the operational dynamics of these systems (Adams, et al., 2022).

The fundamental innovation of OCPM resides in its capacity to process object-centric event data, which links events with multiple objects in lieu of a singular case. This change in thinking permits a more broad and detailed examination of processes. Object-centric event logs (OCELs) represent an extension of traditional logs, incorporating references to multiple objects for each event. This approach allows for the preservation of the complex interdependencies inherent in the processes, which would otherwise be lost. To illustrate, an event in a procurement process may relate to an order, several items within that order, and corresponding invoices. This captures the true nature of the process dynamics, supporting advanced analyses and insights into complex business environments (van der Aalst, 2019).

At the foundation of OCPM is the concept of object-centric event logs. Unlike traditional case-centric event logs that capture events at the process instance level, object-centric event logs store information about the lifecycle of individual business objects and their relationships. This data structure allows OCPM techniques to model the complex interactions between different entity types within a process. (Berti et al., 2023)
Object-centric event logs typically include the following key elements:

- Events: Object-centric process mining deals with discrete events representing actions in a system or process, like order approval or payment. Each event is unique, timestamped, and may have attributes. Events are categorized into types.
- Event Types: Events are grouped based on their nature, such as Order Created or Invoice Sent. Each event type represents a specific action in the process.
- Objects: Objects in object-centric process mining represent entities involved in events, like products or orders. Objects have attributes that can change over time.
- Object Types: Each object belongs to a type, like Product or Invoice.
- Timestamps: The timing of when each event occurred
- Qualifier: Provides additional context or qualifiers related to objects or their relationships. (Berti et al., 2023)

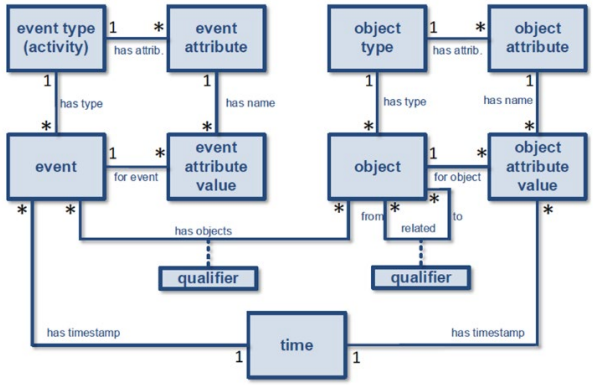
The diagram below represents an enhanced data model that supports object-centric process analysis. It extends the previous model by introducing the concept of “Object Type” and relationships between different objects.

The relationships include:

- An Event has an Event Type and can have multiple attributes stored in Event Attribute and Event Attribute Value.
- An Object has an Object Type and can have multiple attributes stored in Object Attribute and Object Attribute Value.
- An Event is associated with one or more Objects through the “has objects” relationship, allowing events to be linked to multiple objects.
- Objects can have relationships with other Objects, such as “from” and “to” relationships, representing transitions or flows between objects.

- The Qualifier entity provides additional context or qualifiers for the relationships between objects.
- Each Event and Object state change has an associated timestamp captured in the Time entity.

Figure 3: OCPM Relationships



Source: Van Der Aalst (2023)

This meta data model enables advanced object-centric process analysis capabilities, such as tracing the end-to-end lifecycle of different object types, analyzing object state changes and transitions, identifying bottlenecks or inefficiencies in object handling, performing root cause analysis based on object attributes and relationships, and conducting comparative analysis across different object variants or segments.

Advantages of object-centric process mining

Object-Centric Process Mining (OCPM) offers significant advantages over traditional process mining (PM) by addressing the complexity and interconnected nature of modern business processes. One key benefit is its single data extraction process, which captures all relevant objects and their interactions upfront. Unlike traditional PM, which often requires repeated data extraction to integrate new sources or answer additional questions, OCPM's unified approach saves time, reduces redundancy, and ensures consistent data representation for all analyses (Adams & van der Aalst, 2021).

OCPM also excels in modeling relationships between diverse entities, a limitation of traditional PM. By explicitly representing connections and dependencies—for example, between suppliers, transport vehicles, inventory, and customers in a supply chain—OCPM helps organizations understand how changes or delays in one area affect others, enabling better decision-making (Aalst, 2023).

Furthermore, OCPM offers a three-dimensional view of processes, considering interactions, attributes, and temporal aspects. This contrasts with traditional PM's linear, event-focused perspective. The multidimensional approach uncovers insights and optimization opportunities across organizational domains, providing a comprehensive understanding of complex workflows (Adams & van der Aalst, 2021).

Challenges of object-centric process mining

OCPM represents a significant advancement in process mining, offering deeper insights and capturing complex interactions among multiple entities. However, it comes with notable challenges, including data integration, scalability, computational complexity, data quality, and practical implementation (Adams & van der Aalst, 2021).

Integrating diverse datasets with varying formats and structures is a complex, error-prone process, particularly with large, heterogeneous data. Scalability remains a hurdle as processing Object-Centric Event Logs (OCELs) demands significant computational power, especially in large-scale environments like global supply chains. Additionally, the computational complexity of analyzing multiple objects and their interactions can create bottlenecks, requiring advanced techniques such as machine learning and parallel processing to ensure performance. Data quality is another critical concern, as inconsistent or incomplete data can lead to flawed models, necessitating resource-intensive preparation and validation. Practical implementation challenges include transitioning from traditional process mining to OCPM, training staff, and demonstrating tangible value to stakeholders (Adams & van der Aalst, 2021).

Overcoming these hurdles through innovation and robust methodologies is essential for OCPM's adoption and its potential to drive process optimization and operational excellence (Adams & van der Aalst, 2021).

Visualization of Interconnected Processes: Celonis Process Sphere

Celonis Process Sphere marks a significant evolution in the field of process mining by fully embracing object-centric process mining. Leveraging data from multiple enterprise systems, the Process Sphere uses sophisticated algorithms and machine learning techniques to visualize and analyze end-to-end processes in a three-dimensional format. This innovative tool enables organizations to identify inefficiencies, bottlenecks, and compliance issues with unparalleled accuracy, facilitating more effective decision making and process optimization (*Celonis Moves from 2D to 3D Process Mining during Celosphere 2022 - Techzine Europe*, n.d.).

Developed in collaboration with RWTH Aachen University's PADS group, the Process Sphere builds on foundational research, including the OCEL standard and OCPM tools, ensuring robust and reliable analysis. Unlike traditional two-dimensional process mining tools, the Process Sphere's 3D perspective provides a more comprehensive view of complex workflows. Its innovative subway map-like interface simplifies complexity, allowing users to eplore interdependencies and uncover hidden patterns.

By integrating process analytics with an intuitive visualization, the Process Sphere facilitates better decision-making and operational optimization, helping organizations address root causes of challenges effectively (*Post-2022-Process-Sphere.Pdf*, n.d.)

CASE STUDIES AND OCPM-ANALYSIS

Introduction of Business Cases

WoodCorp Inc., a German manufacturer of wooden pallets and crates for shipping, faces significant operational challenges, particularly with on-time delivery and volume compliance. The company's raw materials, mainly wood, are stored in both automated and non-automated warehouses. Production starts before or after order confirmation, using both push and pull strategies. With a current delivery rate of just 64.46%, WoodCorp suffers financial losses due to delays. Despite its large network of warehouses and partners like UPS, DHL, and FedEx, the company faces ongoing delivery issues, resulting in annual losses of approximately €11.8 million.

WoodCorp is looking to gain a more in-depth understanding of its order to cash process and identify ways in which it can improve its on-time delivery performance. By outlining a clear roadmap for improving on-time delivery performance, the WoodCorp case aims to demonstrate the tangible benefits of PM and how it can be further optimiyed using OCPM.

The current system is disjointed, leading to fragmented data across procurement, finance, and operations departments. This fragmentation results in duplicate data entries, inconsistent information, and prolonged processing times. The impact on business is substantial. Woodcorp's ability to manage procurement efficiently is compromised, causing increased operational costs, supplier dissatisfaction, and missed opportunities for bulk purchasing discounts. Delays and errors in processing purchase orders and payments have strained relationships with suppliers, potentially disrupting the supply chain's reliability. Moreover, the absence of real-time data and analytics capabilities hampers decision-making and strategic planning.

OCPM Analysis

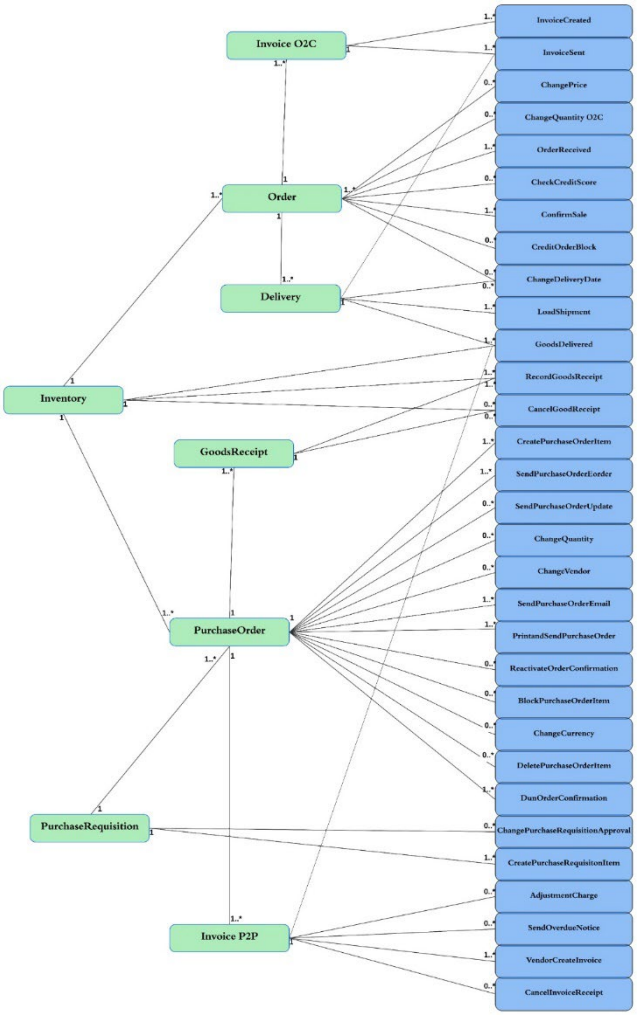
This chapter demonstrates the use of object-oriented process mining by linking two related business processes: the Order to Cash (O2C) process and the Purchase to Pay (P2P) process. We aim to demonstrate the capabilities of object-oriented process mining when dealing with multiple interacting object types in different but related business functions

The first step in conducting an object-centric process mining analysis entails the creation and utilization of an object-centric event log. These logs differ from traditional event logs because they allow for the tracking of multiple object types within a single process. Every event in an object-centric event log is associated with one or more objects, which capture their interactions over time. This structure facilitates the analysis of complex dependencies between different entities. For the purpose of this analysis, the object types identified in the integrated O2C and P2P processes are:

- Order - refers to the customer's request for goods. As such, it includes the following attributes: Order ID, Product Type, Ordered Quantity, Customer Market, Customer ID, Order Date, Order Value and Unit Price.
- Delivery - the logistical process of transporting the goods ordered by the customer from the warehouse to the customer's location. The delivery object contains valuable information: Delivery ID, Delivery Company, Delivered Quantity, Delivered Date, Promised Date, Delivery Status (e.g., tolerance met), Delivery Unit (e.g., kg) and Warehouse Type.
- Invoice O2C - a financial transaction that occurs after the successful delivery of goods. Details related to the invoice process are included in this object: Invoice ID, Invoice Value (derived from the order value), Invoice Date and Customer Country.
- Purchase Requisition - is a company-created document that requests the purchase of goods. The following attributes are required for this object: Purchase Requisition ID, Plant, Unit, Quantity, Document Type, Material Number, Vendor, and Item Number.
- Purchase Order - a formal document issued by a buyer to a seller, indicating the types, quantities, and agreed prices for products. It includes the following details: Purchase Order ID, Purchase Requisition ID, Document Type, Order Type, Created Date, Vendor, Currency, Item Number, Material Number, Order Quantity, Order Unit, Net Price, Plant and Storage Location.
- Invoice P2P - issued after the delivery of the goods or services and their receipt by the buyer. Therefore, the below information is required: Invoice ID, Purchase Order ID, PO Line, Amount, Currency, Company Code and Vendor.
- Goods Receipt - document used to record the receipt of goods or services by the buyer. The following attributes are pertinent for this object: Goods Receipt ID, Purchase Order ID, Item Number, Vendor, Plant and Storage Location.
- Inventory - represents the physical stock of goods held by the organization. It is a critical object type that connects both the Purchase-to-Pay and Order-to-Cash processes, as changes in inventory levels directly impact the organization's ability to fulfil orders and meet customer demand. Inventory ID, Purchase Order ID, Order ID, Material Number.

Object to object relationships are defined as part of object types. They show the expected relationships of objects to each other in business processes. The diagram below shows the relationships between our objects. Each of these objects represents a crucial stage in the procurement lifecycle, from initial purchase requisitions to final payment processing.

Figure 4: Identified objects, events, and relationships



Source: Own Work

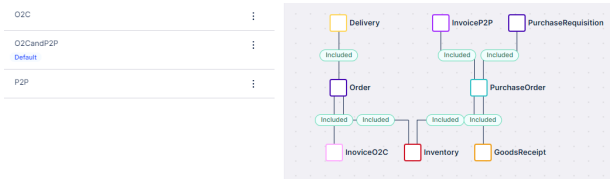
OCPM focuses on discrete events that represent distinct actions that occur at specific times. Unlike traditional PM, where events are linked to a single instance, OCPM links events to multiple types of objects, enabling multi-dimensional analysis of complex processes. These event-to-object (E2O) links convey how events affect or are affected by different objects, revealing rich interdependencies within the system. The diagram below displays all the events that have been detected and their correlation with our objects. All of our events involve their unique id and timestamp, and additional attributes related to each event.

After identifying Object Types, Events and Relationships the next step is to build an object-oriented data model. The process for building an object-oriented data model is focusing on the key steps of data extraction, transformation, and population. We will utilize the Celonis tool to implement this process. The initial step involves connecting to the source system and extracting raw data. Celonis offers pre-built extraction methods for popular systems such as SAP ECC and Oracle EBS that can pull data directly into the OCPM database. However, for this paper's purposes, we have manually uploaded the files. Although this approach works for demonstration purposes, it is much more efficient to connect to the database if possible, so that data can be dynamically refreshed and updated.

When the data is retrieved, we start modelling the business scenario by defining our object and event types, which we introduced in the previous section. The subsequent step entails implementing SQL transformations to utilize the raw extracted data and populate the appropriate object and event tables within the OCPM schema. This involves carefully defining the relevant tables and mapping them to appropriate objects. For example, for the O2C process, we used case and activity tables that were mapped to the relevant objects associated with the process. The P2P process, on the other hand, relied on several tables, such as the EBAN table containing information for purchase orders, which was mapped to the Purchase Order object.

Once the core model is populated with objects, events, and their relationships, we can define perspectives and generate event logs. These perspectives help to narrow down the focus to specific parts of the process, making it easier to analyze. For example, we could create a perspective for the O2C process that includes objects such as orders, invoices, and deliveries, along with events such as order creation and goods delivery. The event logs are then created within these perspectives, providing a sequence of events linked to specific objects. This is key to process analysis in Celonis.

Figure 5: Building a perspective that combines the two processes



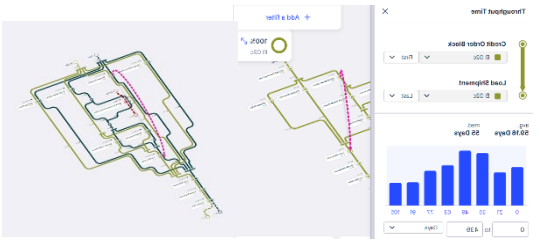
Source: Own Work

The final phase involves the publication and execution of the model within the OCPM data pool. Once the model has been fully configured and populated with data, it can be published and made accessible for analysis. When dealing with dynamic data and therefore databases, to keep everything up to date, data jobs are scheduled to update the model regularly with added information from the source systems. This ensures that the data reflects the most recent business operations, enabling organizations

to monitor their processes in real-time, identify inefficiencies, and make necessary enhancements. There were several limitations to the process discovery and analysis of process variation and object behavior using OCPM in this study, which are described later in this article. These limitations constrained our ability to fully leverage the potential benefits of OCPM. Although OCPM offers considerable flexibility in visualizing and analyzing the relationships between diverse objects, the immediate advantages of this transformation may not be evident in all scenarios, particularly in cases of limited scale or data availability, such as the present instance.

Nevertheless, the study has indicated that OCPM offers a more adaptable and comprehensive understanding of interrelated processes than conventional PM techniques. By allowing multiple objects to be analyzed simultaneously, OCPM can link two or more processes in a way that traditional PM would struggle to accomplish effectively. This flexibility's value depends mainly on the specific needs of the business and the data's complexity. Before implementing OCPM, organizations should carefully assess their business case, as the benefits of this approach are context dependent.

Figure 4:Connecting events from different objects to calculate throughput time

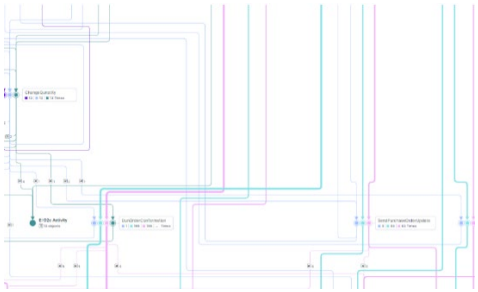


Source: Own Work

Unlike traditional process mining's step-by-step approach, OCPM focuses on individual cases or objects, enabling organizations to uncover concealed bottle-necks and inefficiencies that may otherwise go unnoticed. It provides a holistic view of multiple objects and their relationships, which allows analysts to gain insight into the interdependencies that drive process performance.

In our case study, we used the Process Explorer in Celonis to examine different objects and their associated activities. For each object, we were able to select the number of activities to include in the analysis and the connections between those activities. This resulted in a 'metro map' visualization, where each station repre-sented a different activity. As discussed earlier, each activity may involve sever-al different object types, allowing us to observe various aspects that could influ-ence the performance of processes, such as the fulfilment of an order.

Figure 5: A piece of our "metro map"



Source: Own Work

Regardless, while OCPM provides a more detailed view of processes and interactions, whether it delivers immediate and tangible benefits depends heavily on the specific context of the business and the quality of the available data. Organizations with complex, multi-object processes may find OCPM invaluable in identifying and resolving inefficiencies, but smaller or simpler cases may not derive as much benefit from this approach. Therefore, it is advisable to conduct a comprehensive assessment of the organization's process complexity, data quali-ty, and specific objectives prior to selecting OCPM. The WoodCorp case demonstrated the method's ability to uncover more profound insights into process rela-tionships, but its utility in smaller-scale cases remains limited by data availabil-ity and process complexity.

EVALUATION AND DISCUSSION

Based on the foundations established previously this chapter shifts the focus to the primary objective of this study: investigating the impact of implementing object-centric process mining on business. The directions of our study have been shaped by the research questions outlined in introdcution.

This first question focuses on the comparative analysis of traditional process mining methods versus object-centric process mining, specifically in terms of their impact on the efficiency and effectiveness of the O2C process. *How can object-centric process mining improve the efficiency and effectiveness of the Order-to-Cash process compared to traditional process mining methods?* Due to certain limitations encountered during the study, this question is addressed primarily from a theoretical perspective. As previously discussed, traditional process mining is constrained by its reliance on single-case identifiers. This approach may result in a fragmented view of processes, as it typically analyses each object—such as orders, deliveries, or invoices—in isolation. In contrast, object-centric process mining, which incorporates multiple objects, provides a more comprehensive overview. The integration of diverse objects within a unified analysis enables the capture of intricate relationships and interactions between disparate events.

To illustrate, in the context of the Order-to-Cash (O2C) process, object-centric process mining permits the combination of objects, including orders, deliveries, and invoices, within a unified framework. This holistic approach allows for a more accurate identification of inefficiencies and provides clearer insights into the specific locations within the process where such issues occur. Traditional process mining, on the other hand, conducts separate analyses for each object, limiting its ability to identify such bottlenecks or violations. As a result, object-centric process mining presents a significant advantage in offering a more de-tailed and interconnected view of the O2C process.

Another key benefit of using object-centric data model is its ability to create more accurate and detailed process maps that reflect the intricate realities of business operations. Unlike traditional models that may oversimplify processes, object-centric models capture the full complexity of interactions between various entities, offering a more faithful representation of the actual process landscape. This detailed mapping is crucial for organizations looking to optimize their processes, as it provides a clearer picture of where improvements can be made.

A further significant advantage of object-centric process mining is its ability to reduce system dependencies. By unifying processes and data into a standardized format, organizations can create a consistent and reliable process model that is less reliant on specific systems or technologies. This standardization not only facilitates integration and scalability across different platforms but also ensures that business insights are derived from a holistic view of the process, rather than being constrained by the limitations of individual systems. This approach ultimately leads to more robust and flexible process management, enabling organizations to adapt more readily to changing business environments and technological advancements.

Focusing on the relationships between objects within a process allows for a more nuanced and comprehensive understanding of process dynamics. This approach enhances the visibility of these relationships and significantly improves decision-making capabilities. For example, by analyzing the time it takes to deliver goods to customers using different postal providers, organizations can identify patterns where certain providers consistently result in delayed deliveries. Armed with this insight, companies can proactively choose alternative providers for cases where timely delivery is critical, thereby addressing the root cause of customer dissatisfaction due to late deliveries. This capability to quickly pinpoint and address inefficiencies highlights the value of object-centric process mining in improving customer satisfaction and operational efficiency.

OCPM provides a comprehensive view of process dynamics through the analysis of object relationships, enhancing decision-making. This approach improves efficiency by identifying and addressing insufficiencies and bottlenecks more effectively. Additionally, it simplifies system integration and provides a unified, accurate representation of complex processes.

The second research question examines the benefits and challenges of imple-menting OCPM in a real-world context, using the Celonis tool. *What are the specific benefits and challenges encountered when implementing object-centric process mining in a real-world context using Celonis?* As the adoption of OCPM represents a significant shift from traditional process mining methodologies, understanding these benefits and challenges is critical for organizations considering this approach.
One of the primary challenges involves obtaining access to relevant data and ensuring its quality. In this study, limitations were encountered in acquiring sufficient and appropriate data to construct a robust object-centric data model, which constrained the ability to derive granular insights into the detailed interac-tions between different events and objects within the process. Data quality issues such as inconsistencies, missing values and incomplete records further complicated the modelling process, highlighting the need for rigorous data governance practices in OCPM implementations.

Further, implementing OCPM demands a considerable investment of time and resources. Unlike traditional process mining, which often focuses on a single case identifier, OCPM requires a profound understanding of the various objects that make up a business process and the complex interactions between them. This complexity requires a thorough re-evaluation of existing business processes, data structures and system configurations. Organizations accustomed to single-case process mining may find this transition challenging, as it frequently involves reconfiguring existing systems to accommodate the multi-object framework of OCPM. In addition, the need for specialized knowledge of both business processes and advanced data modelling techniques adds another layer of complexity to the implementation process.

The relative newness of OCPM presents further challenges. As an evolving field, OCPM lacks established best practices and standardized methodologies. This lack of a standardized approach can lead to significant variability in implementation outcomes, making it difficult to predict the success of OCPM initiatives. Organizations that opt to implement OCPM are essentially pioneers in the field and may encounter unforeseen obstacles during

the implementation phase. The absence of standardization also raises concerns regarding the reproducibility and scalability of OCPM solutions, particularly in diverse organizational settings. As a result, it is essential for organizations to conduct a thorough analysis to determine whether their processes would benefit from the complex implementation of OCPM, and to assess the potential risks associated with adopting such emerging technologies.

Regardless of these challenges, once OCPM is successfully implemented, it offers numerous benefits. One of the key benefits is the ability to perform data extraction only once, reducing the manual effort typically required in traditional process mining. By integrating multiple objects into a single analytical frame-work, OCPM provides a more holistic view of the process, allowing the identification of inefficiencies and bottlenecks that might be missed with a single-case approach. This integrated view not only increases productivity, but also supports more informed decision-making by providing a comprehensive understanding of process dynamics.

The implementation of object-centric process mining may encounter difficulties due to data accessibility, complexity, and the absence of standardized approaches. However, its potential advantages in terms of effectiveness, adaptability, and comprehensive process insight make it a promising approach for organizations looking to enhance their process mining capabilities. As OCPM continues to evolve, more research and practical experience must be used to develop more robust methodologies and overcome obstacles identified in this study.

Despite the valuable insights gained from this study, it is imperative to acknowledge several limitations that have significantly impacted the depth and scope of the research findings. The limitations stem from both the evolving nature of OCPM as a field, as well as the practical challenges encountered during the research.

One of the primary limitations of this research is that OCPM is still a relatively new discipline within the broader realm of process mining. Given its novelty, there is a lack of academic literature and empirical studies that examine the practical implementation of OCPM in different industries. Although theoretical frameworks and initial studies have established the foundations, the field remains underexplored, especially in terms of real-world applications. The absence of comprehensive research presents a challenge in validating the study's findings against a broader body of knowledge. Moreover, the absence of well-documented case studies and practical examples hinders the possibility of benchmarking and generalizing the outcomes across diverse business contexts.

The absence of practical examples also complicates the analysis workflow, as it is difficult to draw on already existing examples of how to address specific challenges inside the OCPM implementation.

Another important limitation relates to the practical application of OCPM in industry. OCPM has only been adopted by some companies, which means that access to reliable databases specifically dedicated to OCPM analysis is limited. In this study, we encountered the challenge of utilizing a dataset that was not specifically designed for OCPM. The data we received was originally intended for conventional process mining analysis, which typically focuses on logs of individual events rather than the multiple, interrelated objects that are central to OCPM. Moreover, the data set's limitations extended to the level of detail available for individual events. For example, critical information regarding event connections and object interactions was often insufficient, preventing a deeper exploration of process variants and behaviors. In order to maximize the benefits of OCPM, it is essential to have data that captures the full complexity of an organization's processes. This includes having rich event logs with multiple object types and detailed relationships between events, which were not available in this study. The absence of such data limited the scope of our analysis, making it dif-ficult to draw definitive conclusions about the full range of interactions within the O2C process.

Furthermore, the limitations of the Celonis software platform used during the study posed a significant challenge, given the limitations of our account option. Our access was restricted to a training account, which regrettably did not provide all the advanced functionality required for a comprehensive implementation of OCPM. This prevented us from fully exploring and utilizing the sophisticated process visualization, data modelling, and analysis tools that Celonis offers. As a result, we were unable to fully utilize OCPM's potential to visualize and analyze process flows at a granular level. The extent of our findings was limited due to the inability to explore certain process enhancements or optimizations that could have been revealed through a more sophisticated utilization of the software.

Although OCPM has significant potential for improving process transparency and uncovering inefficiencies in intricate, multi-object processes, there is still room for improvement. This study demonstrates the importance of robust data collection, advanced software functionality, and further research to fully realize its potential. Future research should prioritize addressing these limitations by utilizing more comprehensive data sets, exploring advanced software tools, and conducting comprehensive empirical analyses across diverse industries.

CONCLUSION

This article examines the potential of object-centric process mining (OCPM) to enhance the Order-to-Cash (O2C) and Purchase to Pay (P2P) process, using the Celonis process mining platform. The primary objectives of this study were to assess the efficiency of OCPM in comparison to conventional process mining methodologies, to identify the obstacles associated with its implementation, and to evaluate its potential to enhance business performance. Combining a thorough literature review with an in-depth case study, the investigation offered insights into both the advantages of OCPM and the substantial obstacles it presents.

OCPM offers a more nuanced and comprehensive analysis of complex business processes, considering multiple interconnected entities, such as orders, invoices, and deliveries. However, the process of implementing it poses numerous significant obstacles. The development and management of object-centric data models are notably complex and require advanced technical expertise that may not be readily available in many organizations. Furthermore, OCPM requires extensive data collection and significant computational resources, which can impose significant financial and logistical burdens, particularly on smaller enterprises. The steep learning curve associated with OCPM tools, such as Celonis, further complicates adoption, requiring specialized training. Furthermore, the integration of data from disparate sources presents obstacles, particularly in organizations with siloed or fragmented systems, which may result in inconsistencies in data analysis. Additionally, the detailed and complex insights generated by OCPM can result in information overload, underscoring the need for clear analytical objectives and strategic focus.

OCPM has considerable potential for improving the analysis and optimization of business processes, but the challenges identified in this study must be carefully addressed. Future research should prioritize the development of solutions to mitigate these obstacles, including more intuitive tools, enhanced training resources, and strategies for managing data complexity. Addressing these issues will be critical to making OCPM more accessible and effective across a broader range of organizations, thereby allowing its full potential to be realized.

REFERENCES

Aalst, W. van der. (2023). *Twin Transitions Powered by Event Data. Using Object-Centric Process Mining to Make Processes Digital and Sustainable*. https://publica.fraunhofer.de/handle/publica/450437

Adams, J. N., Park, G., Levich, S., Schuster, D., & van der Aalst, W. M. P. (2022). *A Framework for Extracting and Encoding Features from Object-Centric Event Data* (arXiv:2209.01219). arXiv. http://arxiv.org/abs/2209.01219

Adams, J. N., & van der Aalst, W. M. P. (2021). *Precision and Fitness in Object-Centric Process Mining* (arXiv:2110.05375). arXiv. https://doi.org/10.48550/arXiv.2110.05375

Berti, A., Koren, I., Adams, J. N., Park, G., Knopp, B., Graves, N., Rafiei, M., Liß, L., Genannt Unterberg, L. T., Zhang, Y., Schwanen, C., Pegoraro, M., Van Der Aalst, W. M. P., & Chair of Process and Data Science, RWTH Aachen University. (2023). OCEL (Object-Centric Event Log) 2.0 specification. https://www.ocel-standard.org/2.0/ocel20_specification.pdf

Bolt, A., de Leoni, M., & van der Aalst, W. M. P. (2016). Scientific workflows for process mining: Building blocks, scenarios, and implementation. *International Journal on Software Tools for Technology Transfer*, *18*(6), 607–628. https://doi.org/10.1007/s10009-015-0399-5

Celonis moves from 2D to 3D process mining during Celosphere 2022—Techzine Europe. (n.d.). Retrieved 7 August 2024, from https://www.techzine.eu/news/analytics/93727/celonis-moves-from-2d-to-3d-process-mining-during-celosphere-2022/

Https://vdaalst.com/news/post-2022-process-sphere.pdf. (n.d.). Retrieved 7 August 2024, from https://vdaalst.com/news/post-2022-process-sphere.pdf

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., Van Den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., De Leoni, M., … Wynn, M. (2012). Process Mining Manifesto. In F. Daniel, K. Barkaoui, & S. Dustdar (Eds.), *Business Process Management Workshops* (Vol. 99, pp. 169–194). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28108-2_19

Van Der Aalst, W. M. P. (2019). Object-Centric Process Mining: Dealing with Divergence and Convergence in Event Data. In P. C. Ölveczky & G. Salaün (Eds.), *Software Engineering and Formal Methods* (Vol. 11724, pp. 3–25). Springer International Publishing. https://doi.org/10.1007/978-3-030-30446-1_1

Van der Aalst, W. M. (2023). *Object-centric process mining: unraveling the fabric of real processes.* Mathematics, 11(12), 2691.

Entwicklung eines Dialog-Konzeptes für einen KI-basierten Chatbot im Hochschulbereich

Maximilian Hönig (B.Sc.)

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail: maximilian.hoenig@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail: harald.ritz@mni.thm.de

Kategorie

Abschlussarbeit

Schlüsselwörter

KI, Chatbot, Digitale Assistenten, Regeldatenbank, NLP

Zusammenfassung

Die Technische Hochschule Mittelhessen (THM) setzt einen Chatbot namens *„Winfy“* zur Beantwortung von Fragen im Kontext von Prüfungsangelegenheiten der Studiengänge B.Sc. und M.Sc. Wirtschaftsinformatik ein (vgl. Ritz/Tansel (2023) und Ludwig/Ritz (2023); URL: https://feedback.mni.thm.de/winfy/). Der Dialog zwischen diesem und dem Nutzer ist bisher statisch. Der Chatbot *„Winfy“* antwortet als FAQ-Bot nämlich direkt auf Fragen. Er stellt zum Beispiel keine Rückfragen und besitzt kein Gedächtnis (vgl. Ritz/Tansel 2023). Ziel der Masterarbeit ist es, allgemeine Möglichkeiten zur Optimierung des Chatbots zu identifizieren und umzusetzen, insbesondere aber auch die Fähigkeiten des Bots im Dialog mit dem Nutzer noch besser zu gestalten.

Eine Analyse externer Chatbots half dabei, Verbesserungspotentiale zu erkennen. Die analysierten Chatbots sollten dabei möglichst ähnlich zum Chatbot *„Winfy“* sein und stammten ebenfalls aus Hochschulen, aber auch aus Unternehmen. Die wichtigsten erkannten Potentiale waren dabei:

- Einführung/Verbesserung von Smalltalk
- Fähigkeit, Rückfragen zu stellen
- Einführung von Themen/Kategorien
- Unterstützung von Fremdsprachen
- Kürzere/prägnantere Antworten

Die meisten Punkte erscheinen selbsterklärend, aber warum sollte ein Chatbot, der nicht für das Halten von Konversationen konzipiert wurde, trotzdem in einem gewissen Maß Smalltalk beherrschen? Chatbots werden in Umfragen häufig als zu unpersönlich beschrieben. Smalltalk ermöglicht eine erste emotionale Annäherung und sollte den Bot somit persönlicher wirken lassen (vgl. Adam et al. 2021).

Für die konkrete Umsetzung ist die technische Grundlage entscheidend. Der Chatbot *„Winfy“* nutzt für die Spracherkennung Embeddings und verwendet zur Beantwortung der Fragen eine Regeldatenbank. In dieser sind Beispielfragen mit den jeweils dazugehörigen Antworten hinterlegt. Embeddings sind mathematische Repräsentationen von Wörtern oder Sätzen in Form von Vektoren. In diesem Fall werden alle Beispielfragen der Regeldatenbank in Vektoren umgewandelt. Stellt der Nutzer eine Frage, wird auch diese in einen Vektor transformiert. Anschließend wird die Beispielfrage in der Datenbank gesucht, die der Nutzerfrage am ähnlichsten ist. Im Vektorraum liegen ähnliche Begriffe oder Sätze nah beieinander. Die Ähnlichkeit zwischen zwei Sätzen, und damit zwischen ihren Vektoren, wird hier mit der Cosinus-Ähnlichkeit berechnet. Die Berechnung erfolgt, indem das Skalarprodukt der beiden Vektoren durch das Produkt ihrer Längen geteilt wird. Die hinterlegte Antwort der Beispielfrage mit der größten Ähnlichkeit wird anschließend ausgegeben.

Aufgrund der Leistungsfähigkeit von generativen Transformer-Modellen wurde die Implementierung eines solchen erwägt (vgl. Yenduri et al. 2023). Das Generieren von Antworten wurde jedoch allgemein verworfen. Der Chatbot *„Winfy“* beantwortet unter anderem Fragen über Prüfungsangelegenheiten des Studiengangs, wobei eine korrekte, in einem bestimmten Wortlaut gegebene Antwort nötig ist. Dies könnte mit generativen Modellen nicht gänzlich sichergestellt werden.

Zur Einführung von Smalltalk wurden entsprechende Regeln in der Regeldatenbank hinterlegt. Eine Bearbeitung und Aufteilung der vorhandenen Regeln ermöglicht es, feingranularere Antworten zu geben. Regeln können nun auf eine oder mehrere Kategorien verweisen, die in einer separaten Tabelle gespeichert sind. Der Nutzer kann über eine Liste eine Kategorie auswählen. In diesem Fall werden zur Beantwortung der

Frage des Nutzers nur Regeln betrachtet, die die ausgewählte Kategorie besitzen. Äquivalentes gilt für Fragenvorschläge, die der Chatbot dem Nutzer macht.

Zur Einführung von Rückfragen wurden zwei regelbasierte Vorgehensweisen implementiert, welche nachfolgend als ‚einfache' und ‚multiple' Rückfragen genannt werden. Dazu wurde den Einträgen in der Regeldatenbank ein Attribut hinzugefügt, das eben diese Rückfrage speichert. Diese muss mit ‚ja' oder ‚nein' beantwortbar sein, also z.B. „Möchten Sie wissen, wo Sie das Modulhandbuch finden?". Bei dem Vorgehen für einfache Rückfragen wird die hinterlegte Rückfrage ausgegeben, insofern die Cosinus-Ähnlichkeit nur einen unbefriedigenden Wert erreicht - oder sprichwörtlich, wenn der Chatbot sich unsicher ist. Der Nutzer kann mit ja oder nein antworten und erhält entweder die Antwort oder eine standardisierte Entschuldigung. Eine multiple Rückfrage wird hingegen gestellt, wenn mehrere Regeln eine ähnliche Cosinus-Ähnlichkeit besitzen, also mehrere Antworten in Frage kommen. In diesem Fall werden die Rückfragen in einer Nachricht gesammelt und dem Nutzer vorgestellt. Durch eine Nummerierung der Rückfragen ist der Nutzer über die Angabe der entsprechenden Zahl in der Lage sich die Frage auszuwählen, die er gerne beantwortet haben möchte. Zur Bestimmung der infrage kommenden Antworten wurde eine konfigurierbare Variable eingeführt, die die größte zulässige Abweichung von der höchsten gefundenen Cosinus-Ähnlichkeit beschreibt. Befindet sich innerhalb dieser Abweichung keine weitere Regel, wird die Antwort der Regel mit der höchsten Cosinus-Ähnlichkeit normal ausgegeben.

Die Änderungen konnten vollständig implementiert werden. Der Smalltalk und die einfachen Rückfragen ermöglichen einen natürlicheren Gesprächsverlauf, der den Bot intelligenter wirken lässt. Durch die Aufteilung vorhandener Regeln kann der Chatbot *„Winfy"* zukünftig präziser auf Fragen der Nutzer antworten. Gleichzeitig wird die Regelbasis dadurch homogener. Einzelne Regeln unterscheiden sich also weniger, und es entsteht dadurch eine Verwechselungsgefahr zwischen ihnen. Das Sprachmodell hat es somit schwerer, die Fragen korrekt zu klassifizieren. Dem entgegen wirkt die Implementierung der multiplen Rückfragen. Selbst wenn der Bot nicht in der Lage wäre, die Regeln korrekt auseinander zu halten, würde er in diesem Fall die möglichen Optionen dem Nutzer präsentieren und dieser kann die gewünschte Frage auswählen. Durch die Kategorien wird, wie bereits erwähnt, sowohl beim Beantworten von Fragen als auch beim Geben von Fragenvorschlägen die Regelbasis vorher nach der gewählten Kategorie gefiltert.

Literatur

Adam, M.; Wessel, M.; Benlian, A. (2021): AI-based chatbots in customer service and their effects on user compliance. Electron Markets, vol 31, S.427-455 URL: https://link.springer.com/article/10.1007/s12525-020-00414-7

Ludwig, S.; Ritz, H. (2023): Entwicklung einer plattformunabhängigen Chatbot-Frontend-Anwendung. Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI), Nr. 18, S.170-171, DOI: https://doi.org/10.26034/lu.akwi.2023.n18

Ritz, H.; Tansel, D. (2023): Entwicklung eines KI-basierten FAQ-Chatbots für die Hochschule im Bereich Prüfungsangelegenheiten. Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI), Nr. 17 , S.81-92 URL: https://akwi.hswlu.ch/article/view/3972

Yenduri, G.; Murugan, R.; Govardanan, C.; u.a. (2023): GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. URL: https://arxiv.org/abs/2305.10435

Analyse-Dashboard mit relevanten Kennzahlen für einen KI-basierten Chatbot im Hochschulbereich

Roberto Kakur

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
roberto.kakur@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
harald.ritz@mni.thm.de

Prof. Dr. Peter Hohmann

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
peter.hohmann@mni.thm.de

Kategorie

Abschlussarbeit

Schlüsselwörter

KPI, Kennzahlen, Kennzahlensteckbrief, Analyse-Dashboard, Chatbot, Frontend-Entwicklung, Backend-Entwicklung, Künstliche Intelligenz, Hochschule

Zusammenfassung

1966 entwickelte der deutschstämmige MIT-Professor Joseph Weizenbaum den ersten Chatbot „ELIZA“. Heutzutage sind Chatbots viel weiter in ihrer Entwicklung. Das ist zum einen der zunehmenden Digitalisierung geschuldet und dem Einsatz von Technologien wie maschinellem Lernen und Natural Language Processing. Unternehmen setzen immer mehr auf die digitalen Helfer, zum Beispiel in ihren Online-Shops, um Kundenanfragen zu beantworten und Besucher zu begrüßen. Durch diese Technologie wird ermöglicht, einfache Anliegen eigenständig zu bearbeiten oder bei komplexeren Problemen den Kunden an menschliche Betreuer weiterzuleiten. Ebenso hat das Chatbot-Tool ChatGPT vom Unternehmen OpenAI erheblich zur steigenden Bekanntheit dieser beigetragen. Diese und andere Entwicklungen im Chatbot-Bereich haben dazu geführt, dass das Marktvolumen dieser vom Jahr 2022 bis zum Jahr 2032 voraussichtlich um 23,9 Prozent steigen wird.

Auch im Hochschulbereich ist der Chatbot bereits angekommen. Ein Beispiel dafür ist der Chatbot *„Winfy“*, der an der Technischen Hochschule Mittelhessen (THM) genutzt wird und bereits vielen Studierenden hilft (URL: https://feedback.mni.thm.de/winfy/). 2021 ist dieser aus einem Master-Projekt entstanden und wird kontinuierlich weiterentwickelt. Sein Zweck besteht darin, administrative Fragen im Kontext des Prüfungsbereichs des Bachelor- und Masterstudiums der Wirtschaftsinformatik zu beantworten. Der positive Nebeneffekt ist, dass Studierende jederzeit ihre Fragen stellen können und keine Wartezeiten haben. Dies führt zu einer höheren Zufriedenheit der Studierenden. Zudem entlastet das die Bereiche, die eigentlich für solche Anliegen aufgesucht werden.

Ob ein Chatbot wirklich effektiv ist, muss herausgefunden werden. Den Verantwortlichen fehlt oft eine objektive Betrachtung der Leistung. Diesen Überblick kann ein Analyse-Dashboard geben, indem relevante Kennzahlen (Key Performance Indicators, KPIs) in diesem aufgezeigt werden, aber auch Werkzeuge zur Erweiterung des Chatbot-Wissens bereitgestellt werden.

Im Zentrum dieser Arbeit steht die Identifizierung und Auswahl relevanter Kennzahlen für den Chatbot „Winfy“ an der THM, die sowohl die technische Leistung des Chatbots als auch die Nutzerzufriedenheit abbilden. Um dies umzusetzen, wurden zunächst Kennzahlen gesammelt. Die Sammlung erfolgte mithilfe von Literatur und Online-Quellen. Darauf wurde für jede Kennzahl ein Kennzahlensteckbrief erstellt – dieser beschreibt die einzelnen Kennzahlen im Detail, einschließlich ihrer Zielsetzung, Datenquellen und Darstellungsformen. Dieser Steckbrief bietet eine detaillierte Übersicht über die ausgewählten Kennzahlen und ermöglicht eine strukturierte Analyse der Chatbot-Leistung und Nutzer-zufriedenheit.

Zu den relevanten Kennzahlen gehören unter anderem die Fallback-Rate, die Intent-Scores, die Sentiment-Analyse und der Confidence Score. Diese Kennzahlen zeigen, wie häufig der Chatbot eine unpassende Antwort gegeben hat (Fallback-Rate), wie gut die Intention der Frage erkannt wurde (Intent Scores), welche Stimmung in der formulierten Frage zum Ausdruck kommt (Sentiment-Analyse) und wie zuverlässig die Antwort des Chatbots eingeschätzt wird (Confidence Score).

Neben der Auswahl relevanter Kennzahlen wurde eine Anforderungsanalyse für den Prototyp des Analyse-

Dashboards durchgeführt. Es wurden funktionale und nichtfunktionale Anforderung des Analyse-Dashboards ermittelt.

Die funktionalen Anforderungen des Analyse-Dashboards umfassen die grafische Darstellung von Kennzahlen. Dazu sollten die Kennzahlen in verschiedenen Diagrammtypen visualisiert werden. Das Analyse-Dashboard soll quantitative Daten anzeigen. Falls notwendig, sollen auch Funktionen oder Daten so weit bereitgestellt werden, dass zeitliche Analysen mit entsprechenden Kennzahlen später möglich sind. Zudem soll die Darstellung der Kennzahlen auf Basis von Soll- und Ist-Werten angepasst werden können, beispielsweise durch eine Farbänderung bei Erreichen bestimmter Schwellenwerte. Außerdem soll es die Möglichkeit geben, verschiedene Aggregationen mit den Kennzahlen durchzuführen.

Bei den nichtfunktionalen Anforderungen ist die Benutzerfreundlichkeit von Wichtigkeit. Das Analyse-Dashboard sollte intuitiv bedienbar sein und eine klare Struktur aufweisen, um den Benutzern eine einfache Handhabung zu ermöglichen. Obendrein soll ein schnelles Laden der Kennzahlen im Analyse-Dashboard gewährleistet sein. Das führt zu einer höheren Akzeptanz des Analyse-Dashboards. Zudem wird auf die Einhaltung der Gestaltgesetze besonders Wert gelegt, sodass eine visuelle Klarheit gewährleistet wird und somit die kognitive Last des Nutzers sinkt. Weiterhin wird Wert auf die Skalierbarkeit des Systems gelegt, um zu gewährleisten, dass große Datenmengen effizient verarbeitet werden. Dies bedeutet, dass es große Dateien, bspw. Logdaten, ohne Probleme verarbeiten kann. Dabei spielt die Entscheidung der Technologien eine große Rolle. Weiterhin soll der Code mit Kommentaren erklärt werden. Das führt dazu, dass ein fremder Entwickler diesen mit wenig Aufwand verstehen kann, um in Zukunft Erweiterungen oder Optimierungen durchzuführen.

Danach werden die gesammelten Kennzahlen nach der MoSCow Methode („Must have, Should have, Could have, Won't have“) für die prototypische Umsetzung priorisiert. Bei der Gestaltung des Analyse-Dashboard-Prototyps wurden die Gestaltgesetze berücksichtigt, wie das Gesetz der Ähnlichkeit – dies erleichterte das Verständnis der Informationen. Zum Beispiel wurden die Fallback-Rate und die Cosine-Similarity-Werte in denselben Farben dargestellt, um dem Benutzer klar zu zeigen, ob ein Zustand gut, akzeptabel oder kritisch war.

Basierend auf den Gestaltgesetzen wurden Mockups mit allen gesammelten Kennzahlen erstellt. Diese sind in die drei Kategorien „Allgemeine KPIs“, „Technologische Chatbot Performance“ und „Nutzerverhalten“ unterteilt worden. Dies dient als Orientierung und Hilfe für zukünftige Erweiterungen und Entwicklungen des Analyse-Dashboards.

Fazit

Der entwickelte Prototyp eines Analyse-Dashboards stellt ein gutes Werkzeug zur Überwachung und Optimierung des Chatbots Winfy darstellt. Er bietet den Administratoren und Entwicklern vielfältige Informationen, die zur Verbesserung des Chatbot-Systems genutzt werden können. Beispielsweise ist die implementierte Fallback-Rate eine wichtige KPI, die den prozentualen Wert anzeigt, mit der ein Chatbot keine passende Antwort findet.

Weiterhin kann das Analyse-Dashboard zukünftig durch die identifizierten Kennzahlen, die detailliert in einem Kennzahlensteckbrief festgehalten wurden, weiterentwickelt werden. Diese werden konzeptionell erklärt, um die spätere Entwicklung zu erleichtern.

Hervorzuheben ist die Fokussierung auf die Visualisierung, die durch Mockups unterstrichen wurde. In diesen wurden die Gestaltgesetze der Nähe, Ähnlichkeit, Umschließung und der Kontinuität demonstriert. Das soll als Hilfestellung und Basis dienen, die Gestaltgesetze umzusetzen und den Verantwortlichen des Analyse-Dashboards eine intuitive und einfache Übersicht zu geben.

Literatur

Few, S. (2013). Information dashboard design: Displaying data for at-a-glance monitoring (2. Aufl.). Burlingame, CA: Analytics Press. ISBN 978-1-938377-00-6.

Kohlhammer, J., Proff, D. U., & Wiener, A. (2018). Visual Business Analytics: Effektiver Zugang zu Daten und Informationen. [Online] dpunkt.verlag https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1811559&site=ehost-live.

Ludwig, S.; Ritz, H. (2023). Entwicklung einer plattformunabhängigen Chatbot-Frontend-Anwendung, in: Anwendungen und Konzepte in der Wirtschafts-informatik (AKWI), Nr. 18 (2023), S.170-171, DOI: https://doi.org/10.26034/lu.akwi.2023.n18

Ritz, Harald; Tansel, Dogus (2023). Entwicklung eines KI-basierten FAQ-Chatbots für die Hochschule im Bereich Prüfungsangelegenheiten, in: Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI), Nr. 17 (2023), S.81-92, DOI: https://doi.org/10.26034/lu.akwi.2023.n17

Schloß, D., D'Onofrio, S., & Meinhardt, S. (2023). Chatbot Analytics mittels Betriebsdaten. Robotik in der Wirtschaftsinformatik [Online] Springer Fachmedien Wiesbaden. https://link.springer.com/book/978-3-658-39621-3.

Skodowski, M. (2023). 20 Chatbot KPIs – Wie der Erfolg virtueller Assistenten gemessen wird. [Online] BOT friends GmbH. https://botfriends.de/blog/chatbot-kpi/.

Entwicklung eines KI-Screenreaders zur Verbesserung der Barrierefreiheit von Lerninhalten

Cosimo Teklenburg
HTW Berlin
Wirtschaftsinformatik
Treskowallee
10318 Berlin
E-Mail:
teklenburgcosimo@gmail.com

Prof. Dr. Birte Malzahn
HTW Berlin
Wirtschaftsinformatik
Treskowallee
10318 Berlin
E-Mail:
birte.malzahn@htw-berlin.de

Schlüsselwörter

Lernmanagementsysteme, SAP Enable Now, Künstliche Intelligenz, ChatGPT 4o, digital adoption platform

Zusammenfassung

Diese Bachelorarbeit (Teklenburg 2024) untersuchte die Integration von Künstlicher Intelligenz (KI) in den SAP Enable Now Manager, der zur Erstellung und Verwaltung von Schulungsmaterialien für SAP-Systeme genutzt wird.

Ein zentraler Teil der Arbeit war die Entwicklung eines Prototyps eines KI-Screenreader. Dieser wurde so programmiert, dass er Text und visuelle Informationen simultan erfassen und in gesprochene Sprache umwandeln kann. Dabei analysiert er die Inhalte und passt die Beschreibung an den Kontext an, um die Inhalte möglichst verständlich und umfassend wiederzugeben. Der für diese Arbeit relevante Kern der KI liegt damit bei dem Erkennen von Mustern, dem Analysieren großer Datenmengen und dem Generieren neuer Inhalte (vgl. Kreutzer, 2023, S. 30). Die Navigation wurde einfach gestaltet, sodass Nutzer*innen mit wenigen Befehlen (z.B. starten: alt+s) durch die Lerninhalte geführt werden. Ziel war es zum einen, sehbehinderten Personen die Nutzung von Lernmaterialien zu erleichtern. Aber auch sehende Menschen profitieren im Sinne des multimedialen Lernens, nach der eine Kombination von visuellen und auditiven Informationen als besonders effektiv gilt (vgl. Mayer, 2009, S. 118-122).
Der KI Screenreader wurde als Chrome-Erweiterung umgesetzt. Er besteht aus clientseitigen Komponenten sowie einer serverseitigen Komponente. Die clientseitigen Komponenten steuern die Benutzeroberfläche. Die Nutzer*innen können den Screenreader über die Benutzeroberfläche aktivieren, indem über ein Popup ein Screenshot des aktuellen Tabs erstellt wird. Nach dem Start wird ein Hintergrundsound abgespielt, um die Verarbeitung anzuzeigen. Es erfolgt die Weiterleitung des Screenshots an den Server für die Bild- und Texterkennung.

Die serverseitige Komponente, implementiert in Node.js und gehostet auf Heroku, ist für die Verarbeitung des Bildmaterials verantwortlich. Das Hosting auf Heroku gewährleistet die Skalierbarkeit und Zuverlässigkeit, die für den gleichzeitigen Zugriff vieler Nutzer*innen erforderlich sind. Der empfangene Screenshot wird in ein Base64-Format konvertiert. Die OpenAI API wird genutzt, um den Screenshot zu analysieren und eine Beschreibung zu generieren. Die erstellte Beschreibung wird anschließend in Audio umgewandelt und an die Chrome-Erweiterung zurückgesendet. Die Ergebnisse werden den Nutzer*innen über einen Audioplayer ausgegeben.

Eine durchgeführte Evaluation sowohl mit sehenden als auch blinden Personen bestätigte, dass die Nutzung des KI-basierten Screenreaders die Aufnahme von Inhalten verbessert. Die Fähigkeit des Screenreaders, Abbildungen verständlich zu erläutern, wurde von beiden Nutzergruppen als sehr hilfreich bewertet. Indem er Lerninhalte für blinde und sehende Menschen gleichermaßen erschließt, trägt der KI-Screenreader somit maßgeblich zur Barrierefreiheit von Lerninhalten und somit zur Inklusion bei.

Die Integration eines KI-Screenreaders im SAP Enable Now Manager bietet das Potenzial zur Verbesserung der Lerninhaltskonsumption und kann als Grundlage für zukünftige Anwendungen dienen, die die Inklusion und Effizienz im Bildungsbereich weiter vorantreiben.

Zukünftige Weiterentwicklungsmöglichkeiten für den Screenreader-Prototyp umfassen die Erweiterung der Anpassungsoptionen, wie etwa die Möglichkeit, verschiedene Stimmen für die Sprachausgabe auszuwählen. Technische Optimierungen, wie eine schnellere Verarbeitung der Inhalte, könnten die Benutzererfahrung weiter verbessern. Zudem könnte ein Textmodus eingeführt werden, der den gesprochenen Inhalt auch als Text anzeigt, um Nutzer*innen das Kopieren und Wiederverwenden von Informationen zu erleichtern.

Literatur

Kreutzer, Ralf T. (2023): Künstliche Intelligenz verstehen. Grundlagen - Use-Cases - unternehmenseigene KI-Journey. 2. Auflage: Springer Gabler

Mayer, Richard E. (2009): Multimedia learning. 2. Auflage. Cambridge, New York: Cambridge University Press.

Teklenburg, C. (2024): Untersuchung der Auswirkungen der Integration von Künstlicher Intelligenz in den SAP Enable Now Manager. Bachelorarbeit, Hochschule für Technik und Wirtschaft Berlin.

Hinweis: Der KI-Screenreader ist frei verfügbar, ein Zugriff kann auf Anfrage beim Erstautor gewährt werden.

Fortschritte in der KI-Entwicklung für einen proaktiveren Analyseansatz der Business Intelligence

Niklas Zährl
Technische Hochschule Mittelhessen
Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
niklas.zaehrl@mnd.thm.de

Prof. Dr. Harald Ritz
Technische Hochschule Mittelhessen
Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
harald.ritz@mni.thm.de

Prof. Dr. Frank Kammer
Technische Hochschule Mittelhessen
Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
frank.kammer@mni.thm.de

Kategorie
Abschlussarbeit

Schlüsselwörter
Artificial Intelligence as a Service, Business Analytics, Business Intelligence, Data-to-Decission, Gradient Boosting, Künstliche Intelligence, Large Language Models, Machine Learning, Natural Language Processing, Predictive Analytics, Proaktive Analysen, Prototyp.

Abstract

Der Einsatz von Künstlicher Intelligenz (KI) im Bereich der Business Intelligence (BI) eröffnet neue Möglichkeiten, den BI-Prozess effizienter und proaktiver zu gestalten. Diese Masterarbeit untersucht systematisch die theoretischen Grundlagen und technologischen Fortschritte, die erforderlich sind, um den gesamten BI-Prozess durch KI zu erweitern und effizienter zu gestalten. Die Arbeit gliedert sich in einen theoretischen Literaturreview und einen praktischen Teil, der sowohl reale Use Cases als auch die Entwicklung eines fiktiven Use Cases und die darauf aufbauende Entwicklung eines BI-Prototyps umfasst.

Im theoretischen Teil der Arbeit werden die Konzepte von BI und KI detailliert beleuchtet. Die Arbeit beginnt mit einer historischen Betrachtung der Entwicklung von BI, angefangen von der Datenanalyse über Data Warehousing bis hin zu modernen Data-Lakehouse-Architekturen, die eine flexible Verarbeitung großer, heterogener Datenmengen ermöglichen. Zentrale Begriffe wie deskriptive, diagnostische, prädiktive und präskriptive Analysen werden eingeordnet, wobei besonderes Augenmerk auf die Transformation hin zu proaktiven Analysen gelegt wird. Diese zielen darauf ab, Unternehmen in die Lage zu versetzen, zukünftige Herausforderungen und Chancen frühzeitig zu erkennen und darauf basierend strategisch zu handeln. Ein weiterer Schwerpunkt liegt auf den Grundlagen der KI, insbesondere auf maschinellem Lernen und Deep Learning. Der theoretische Rahmen beschreibt die hierarchische Struktur von KI-Technologien, beginnend mit allgemeinen Konzepten bis hin zu spezifischen Ansätzen wie Natural Language Processing und Large Language Models.

Anschließend wird durch die Zerlegung des Data-to-Decision- und Business-Analytics-Prozesses eine Struktur geschaffen, die aufzeigt, wie KI die Datenverarbeitung, Automatisierung sowie Benutzer´-freundlichkeit und Vorhersage und Prognose unterstützen und verbessern kann. Insbesondere im Bereich der Predictive Analytics können maschinelle Lernmodelle mit Ansätzen wie Gradient Boosting, aber auch mit Large Language Models und generativer KI proaktive Analysen liefern und dabei oft bessere Ergebnisse erzielen als traditionelle Analyseverfahren. Zudem spielt Natural Language Processing eine wichtige Rolle in der KI-gestützten Datenverarbeitung sowie Automatisierung und bietet eine neue Schnittstelle, um auch Nicht-Datenexperten die Datenanalyse zu ermöglichen. Dadurch werden Unternehmen in die Lage versetzt, proaktiv statt reaktiv auf künftige Unsicherheiten zu reagieren.

Die Bereitstellung dieser KI-Modelle wird im praktischen Teil der Arbeit beschrieben und erfolgt häufig über etablierte Cloud-Plattformen im Rahmen eines Artificial-Intelligence-as-a-Service-Modells. Als beispielhafte Cloud-Plattform wird in dieser Arbeit Microsoft Azure verwendet, das zentrale Komponenten für den in dieser Arbeit entwickelten Prototypen bereitstellt. Die Daten werden mithilfe einer automatisierten ETL-Pipeline, Python-Skripten und AzureML-Modulen bereinigt, transformiert und für

maschinelles Lernen vorbereitet. Im Zentrum des Prototyps steht die Implementierung eines Two-Class Boosted Decision Trees, der basierend auf historischen Kundendaten präzise Vorhersagen zur Kundenabwanderung trifft. Das fertige Modell wird in eine produktive Umgebung integriert, wobei interaktive Dashboards in Power BI die Ergebnisse visualisieren und mithilfe von Diagrammen sowie anderen Techniken konkrete Handlungsempfehlungen ableiten und Abwanderungsursachen verdeutlichen. Die Implementierung zeigt, dass KI nicht nur bestehende Prozesse beschleunigen kann, sondern auch neue Möglichkeiten zur proaktiven Analyse eröffnet, die Unternehmen strategische Vorteile verschaffen.

Der erfolgreiche Einsatz von KI in diesem Kontext hängt maßgeblich von der Qualität der Daten und der Skalierbarkeit der Infrastruktur ab. Insbesondere der Prozess der Datentransformation kann in der Praxis einen erheblichen Umfang annehmen, was sich bereits im Prototyp zeigt. Ungenaue, unvollständige oder nicht standardisierte Daten können die Effektivität der KI-Modelle erheblich beeinträchtigen. Ebenso wird die Skalierbarkeit der zugrunde liegenden Infrastruktur als kritisch betrachtet, insbesondere wenn große Datenmengen in Echtzeit verarbeitet werden sollen. Ein weiterer entscheidender Punkt ist die Transparenz und Erklärbarkeit der eingesetzten KI-Modelle. Gerade in geschäftskritischen Anwendungen ist es wichtig, dass die Ergebnisse der KI nachvollziehbar und verständlich für die Nutzer sind, um Vertrauen und Akzeptanz zu gewährleisten. Fehlende Transparenz könnte die Akzeptanz der Systeme bei Entscheidungsträgern erheblich einschränken.

Die Ergebnisse der Arbeit zeigen, dass trotz der genannten Herausforderungen die Integration von KI in BI-Systeme einen erheblichen Mehrwert für Unternehmen bietet. Der erste Teil der Arbeit zeigt, wie KI den BI-Prozess in Bezug auf Datenverarbeitung, Automatisierung und Benutzerfreundlichkeit verbessern kann, um letztendlich bessere Prognosen und Vorhersagen für proaktive Analysen zu ermöglichen. Der Prototyp und die Anwendungsfälle geben ein konkretes Beispiel, wie eine KI-Unterstützung im gesamten BI-Prozess in der Praxis aussehen könnte und untermauern damit das zuvor erarbeitete theoretische Konzept. Die durch die KI unterstützten proaktiven Analyseansätze im BI-Prozess ermöglichen es somit, zukünftige Entwicklungen genauer vorherzusagen, Risiken frühzeitig zu erkennen und Chancen effektiv zu nutzen.

Literatur

Azmi, M; Mansour, A; Azmi, C. (2023): A Context-Aware Empowering Business with AI: Case of Chatbots in Business Intelligence Systems. In: Procedia Computer Science, S. 479–484. DOI:10.1016/j.procs.2023.09.068.

Döbel, I. et al. (2018): Maschinelles Lernen– Kompetenzen, Anwendungen und Forschungsbedarf. In: Fraunhofer-Gesellschaft: Frauenhofer IAIS, Frauenhofer IMW.

Eboigbe, E. O. et al. (2023): Business Intelligence Transformation through AI and Data Analytics. In: Engineering Science & Technology Journal 4, Nr. 5, S. 285–307. DOI:10.51594/estj.v4i5.616.

Figalist, I. et al. (2022): Breaking the vicious circle: A case study on why AI for software analytics and business intelligence does not take off in practice. In: Journal of Systems and Software 184. DOI:10.1016/j.jss.2021.111135.

Gurcan, F. et al. (2023): Business Intelligence Strategies, Best Practices, and Latest Trends: Analysis of Scientometric Data from 2003 to 2023 Using Machine Learning. In: Sustainability 15, Nr. 13. DOI:10.3390/su15139854.

Haselbeck, F. et al. (2022): Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions. In: Machine Learning with Applications 7. DOI:10.1016/j.mlwa.2021.100239.

Kieninger, M., Hrsg. (2017): Digitalisierung der Unternehmenssteuerung: Prozessautomatisierung, Business Analytics, Big Data, SAP S/4HANA, Anwendungsbeispiele. Stuttgart: Schäffer-Poeschel Verlag, 2017.

Lins, S. et al. (2021): Artificial Intelligence as a Service. In: Business & Information Systems Engineering 63, Nr. 4, S. 441–456. DOI:10.1007/s12599-021-00708-w.

Michael, C. I. et al. (2024): Data-driven decision making in IT: Leveraging AI and data science for business intelligence. In: World Journal of Advanced Research and Reviews 23, Nr. 1, S. 472–480. DOI:10.30574/wjarr.2024.23.1.2010.

Seiter, M. (2023): Business Analytics: Wie Sie Daten für die Steuerung von Unternehmen nutzen. Verlag Franz Vahlen GmbH, 2023. 3. Auflage. DOI:10.15358/9783800669295-I.

Tripathi, M. A. et al. (2023): Machine learning models for evaluating the benefits of business intelligence systems. In: The Journal of High Technology Management Research 34, Nr. 2. DOI:10.1016/j.hitech.2023.100470.

Zohuri, B; Masoud Moghaddam (2020): From business intelligence to artificial intelligence. In: Journal of Material Sciences & Manufacturing Research.