**Editorial**


Christian Müller

Liebe Leserinnen und Leser,

vor Ihnen liegt nunmehr die bereits zweiundzwanzigste Ausgabe des E-Journals **Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI).** Wir hoffen, dass wir Ihnen erneut eine Reihe spannender Beiträge aus dem Umfeld der Wirtschaftsinformatik zusammenstellen konnten. Die regulären Artikel durchlaufen einen vollständig anonymisierten Review-Prozess, in dem sie von zwei Gutachtern sowie einem Redakteur beziehungsweise Herausgeber begleitet werden.

Die aktuelle Ausgabe gliedert sich in die Rubriken Praxis, Trends und Abschlussarbeiten. Die Beiträge behandeln Fragestellungen aus den Bereichen Controlling, Anforderungs- und Projektmanagement, Geschäftsprozessmanagement, Optimierung, maschinelles Lernen und Simulation.


Norbert Ketterer

Die Rubrik Praxis umfasst fünf Beiträge. Zwei Arbeiten befassen sich mit kunden- und entscheidungsorientierten Dashboards: Ein Artikel stellt ein KI-fähiges Power-BI-Dashboard auf Basis von Wertgeneratoren zur Unterstützung der strategischen Planung im Controlling vor. Eine weitere Arbeit behandelt die prototypische Neugestaltung des Controlling-Dashboards eines Gewerbekundenportals für Versicherungsmakler. Zwei Beiträge widmen sich der Gestaltung betrieblicher Softwareprojekte. Einer entwickelt auf Grundlage einer Ist- und Nutzwertanalyse ein Optimierungskonzept für das Anforderungsmanagement eines Elektronikdienstleisters. Der andere beschreibt die Konzeption und Umsetzung eines Administrationsmoduls zur Verwaltung der Mehrsprachigkeit einer Datenschutzmanagementsoftware. Schließlich werden fallbasiertes und objektzentriertes Process Mining hinsichtlich ihrer Eignung für das Supply Chain Resilience Management verglichen.


Frank Herrmann

Die sieben Beiträge der Rubrik Trends stammen aus den Bereichen Optimierung, maschinelles Lernen, Simulation und Process Mining. Ein Artikel zeigt mögliche Grenzen genetischer Algorithmen bei multimodalen Optimierungsproblemen auf. Zwei Arbeiten untersuchen die Rekonstruktion der photosynthetisch aktiven Strahlung in Gewässern: Einerseits wird die Übertragbarkeit neuronaler Netze auf unterschiedliche Meeresregionen betrachtet, andererseits werden geeignete Wellenlängenkombinationen für Regressionsmodelle mithilfe genetischer Algorithmen bestimmt. Drei weitere Beiträge behandeln die simulationsgestützte Temperaturüberwachung bei der Pasteurisierung, einen hochskalierbaren Simulator für den städtischen Straßenverkehr sowie die Anwendung des Fossen-Modells zur dynamischen Analyse von Wasserfahrzeugen. Ein weiterer Artikel stellt einen inkrementellen Ansatz zur effizienteren Konstruktion von Transitionssystemen für das regionenbasierte Process Mining vor.

Die Rubrik Abschlussarbeiten umfasst drei Beiträge zum Einsatz künstlicher Intelligenz. Die Arbeiten behandeln die automatisierte Qualitätsanalyse wissenschaftlicher Texte mit feinabgestimmten Sprachmodellen, die KI-gestützte Informationsextraktion aus Diagrammen sowie die Potenziale und Voraussetzungen einer skalierbaren Einführung von SAP-Joule im Unternehmensumfeld.

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

Regensburg, Fulda, Luzern und Wildau, im Juli 2026.

Frank Herrmann, Norbert Ketterer und Christian Müller

# Nutzung von KI durch Dashboards zur innovativen Steuerung im Controlling – Best Practises für den Strategischen Planungsprozess mit Hilfe von Wertgeneratoren


| Bettina Binder | Frank Morelli | Alisa Bay | Florian M. Rottner | Katharina M.-T. Schmidt |
|---|---|---|---|---|
| Hochschule Pforzheim | Hochschule Pforzheim | SAP SE | Hochschule Pforzheim | Hochschule Pforzheim |
| Tiefenbronnerstr. 66<br>75175 Pforzheim<br>bettina.binder@hs-pforzheim.de | Tiefenbronnerstr. 66<br>75175 Pforzheim<br>frank.morelli@hs-pforzheim.de | Dietmar-Hopp-Allee 16<br>69190 Walldorf<br>alisa.bay@sap.com | Tiefenbronnerstr. 66<br>75175 Pforzheim<br>rottnerf@hs-pforzheim.de | Tiefenbronnerstr. 66<br>75175 Pforzheim<br>schmid5t@hs-pforzheim.de |



**SCHLÜSSELWÖRTER**

Dashboard, Wertgeneratoren, Controlling, Künstliche Intelligenz, Power BI, strategische Planung, KPI, Budgetierung



**ABSTRACT**

Dashboards werden im Controlling bislang vorwiegend zur operativen Steuerung eingesetzt. Dieser Beitrag zeigt, wie ein KI-gestütztes Dashboard auf Basis von Wertgeneratoren den strategischen Planungsprozess innovativ unterstützen kann. Am Beispiel eines international tätigen Fahrrad- und Spielzeugherstellers wird die Entwicklung und Implementierung eines Werttreiberbaums in Microsoft Power BI illustriert. Dabei werden die konzeptionellen Grundlagen zur Auswahl von Spitzenkennzahlen und Wertgeneratoren ebenso erläutert wie die datentechnischen Voraussetzungen einer Business-Intelligence-Architektur. Ergänzend werden KI-Grundlagen, Kernfähigkeiten und Anwendungsmuster – von Prediction über Recommendation bis hin zu Conversational Interfaces – systematisch aufgearbeitet und auf den Controlling-Kontext übertragen. Die praktische Umsetzung zeigt, wie KPI-Visuals, Farbcodierungen und Filterfunktionen in Power BI eine effektive Soll-Ist-Analyse ermöglichen. Das Ergebnis ist ein integriertes, KI-fähiges Dashboard, das als Single Source of Truth quantitative und qualitative Steuerungsinformationen bündelt und Controller bei Analyse, Visualisierung und Entscheidungsunterstützung nachhaltig entlastet.


## EINLEITUNG

Dashboards werden heutzutage häufig in der Controlling-Abteilung genutzt, um operative Kennzahlen und deren Entwicklungen der letzten Stunden, Tage oder Wochen darzustellen. In diesem Artikel jedoch erfährt das Dashboard eine erweiterte Einsatzmöglichkeit als Werkzeug zur strategischen Berichterstattung auf Basis von Wertgeneratoren. Eingebettet in ein KI-System werden die KI-Funktionen der Anwendung sowie die Verknüpfung mit dem zugehörigen Controllingprozess näher erläutert.

## DER STRATEGISCHE PLANUNGSPROZESS – TRADITIONELL UND DIGITALISIERT

Der strategische Planungsprozess bildet einen zentralen Bestandteil eines Unternehmens und richtet dies langfristig aus. Darüber hinaus werden konkrete Maßnahmen für den operativen Alltag abgeleitet. Eine effektive interne Steuerung entsteht durch die enge Verbindung von Planung, Kontrolle und Informationsmanagement. In diesem System werden strategische Ziele Schritt für Schritt in konkrete Vorgaben umgesetzt, wobei die Budgetplanung eine Schlüsselrolle übernimmt. Sie wandelt geplante Aktivitäten in messbare Key Performance Indicators (KPIs) um und schafft damit die Grundlage für eine zielgerichtete Unternehmenssteuerung (vgl. Horváth et. al. 2024, S. 80 ff.).

Die traditionelle Budgetplanung folgt starren, zyklischen und hierarchischen Prozessen. Dabei werden die langfristigen Ziele in mittelfristige Pläne und festgelegten Budgets differenziert (vgl. Horváth et.al., 2020, S. 94ff.). Sie ist geprägt von langen Planungsphasen, intensiven Abstimmungen zwischen Fachabteilungen und einer starken Orientierung an historischen Daten. Dies führt zu einer Einschränkung der Flexibilität, schafft jedoch Stabilität (vgl. Horváth et. al. 2020, S. 80ff.). Budgets dienen hier als Richtlinie für operative Umsetzung, Zielvereinbarungen sowie ein Vergleich zwischen Soll- und Ist- Zustand zur Überprüfung des Fortschritts (vgl. Horváth et. al., 2024, S. 135ff.; vgl. Britzelmaier, 2020, S. 320ff.).

In der digitalen Welt nutzen Unternehmen bei den strategischen Planungsprozessen IT-gestützte Systeme, um eine effizientere und datengetriebene Budgetierung um-

zusetzen. Darüber hinaus ist eine automatisierte Erfassung von Daten und eine flexiblere Anpassung an neuen Entwicklungen wichtig (vgl. Horváth et.al., 2015, S. 340 ff.). Moderne Ansätze integrieren in ihre Entscheidungen Business Intelligence und Business Analytics zur Analyse großer Datenmengen sowie verbesserte Forecast-Methoden. Dies ermöglicht flexiblere Planung, die über starre Strukturen hinausgeht. (vgl. Horváth et.al. 2015, S. 344ff.)

## VORAUSSETZUNGEN UND LÖSUNGSSCHRITTE ZUR EINFÜHRUNG EINES DASHBOARDS MIT WERTGENERATOREN IN POWER BI

Eine zentrale Voraussetzung für die Einführung eines Dashboards mit Wertgeneratoren ist zunächst die Grundkonzeption eines Kennzahlensystems sowie den darin berücksichtigten Wertgeneratoren (vgl. Weber et al., 2017, S. 271). Kennzahlen übermitteln hierbei allgemein quantitative Informationen über das Unternehmen. Sie sind „Maßgrößen, die willentlich stark verdichtet werden zu absoluten oder relativen Zahlen, um mit ihnen in einer konzentrierten Form über einen zahlenmäßig erfassbaren Sachverhalt berichten zu können." (Gladen, 2014, S. 9). Folglich ermöglichen Kennzahlensysteme den Fokus des Managements auf wesentliche Informationen zu lenken und damit eine effektive Unternehmensanalyse und -steuerung zu ermöglichen (Vgl. Gladen, 2014, S.9). Die Verhinderung einer Informationsüberflutung ist damit die Grundidee von Kennzahlensystemen und muss bei der Grundkonzeption berücksichtigt werden. Im ersten Schritt erfolgt hierbei die Auswahl der Spitzenkennzahl. Als Auswahlkriterium sollte vor allem die Verständlichkeit dieser für die Entscheidungsträger herangezogen werden und im besten Fall vertraute Steuerungsgrößen wie Bilanz oder Gewinn- und Verlust (GuV) enthalten. In manchen Fällen kann es auch Sinn ergeben, eine zweite Spitzenkennzahl zu definieren, sollte eine nicht alle vom Management gewünschten Sachverhalte abbilden (z.B. zusätzliche Abbildung des Liquiditätsmanagements) (vgl. hierfür auch Reichmann, 2011, S. 36ff.). Um eine effektive Unternehmensanalyse und -steuerung zu gewährleisten, sollte das Kennzahlensystem die Organisationsstruktur des Unternehmens berücksichtigen (z.B. Produktlinien oder Geschäftsregionen). Im dritten Schritt muss ein Rahmenkonzept der Wertgeneratoren definiert werden. Dieses sollte eine funktionale oder sachlogische hierarchische Struktur ergeben und sich bis zur Spitzenkennzahl aggregieren (vgl. Weber et al. S. 271-274). Bei der Auswahl der Wertgeneratoren im Speziellen gilt der Grundsatz: „Weniger ist Mehr". Eine begrenzte Anzahl von Wertgeneratoren fördert die Verständlichkeit und ermöglicht insbesondere eine effektive Unternehmenssteuerung (vgl. Gladen, 2014, S. 96).

Neben der konzeptionellen gilt es ebenfalls die informationstechnologische Voraussetzung der Datenbereitstellung und -anbindung zu erfüllen. Für das Business Intelligence (BI) gestützte Controlling im Rahmen der Dashboardentwicklung eignen sich transaktionale Systeme, die entlang der Wertschöpfungskette eingesetzt werden, wie bspw. Enterprise-Resource-Planning (ERP)-Systeme oder Customer-Relationship-Management (CRM)-Systeme je nach Werttreiberkonzeption nur selten (vgl. Schön, 2022, S. 467ff.). Besser geeignet als Datenquelle ist ein sog. Data-Warehouse (DWH). Es ist dadurch gekennzeichnet, dass Daten aus verschiedenen transaktionalen Systemen sowie unternehmensexternen Quellen harmonisiert und integriert vorliegen, nach Themenbereichen sortiert, häufig bereits aggregiert sind und Zeiträume betrachtet sowie entgegen der häufig nicht dauerhaften Speicherung von Daten in den transaktionalen Systemen eine Historie abbildet (vgl. Inmon, 2005, S. 29ff. und Baars/Kemper, 2021, S. 20f.). Zudem ermöglichen die im Rahmen der Datentransformation durchgeführte Anreichung der Daten bereits die Berechnung der für das Kennzahlensystem benötigten Wertgeneratoren sowie der Spitzenkennzahl. Das DWH liefert somit die notwendigen fachlichen entscheidungsunterstützenden Daten, die für die Informationsbereitstellung im Dashboard direkt genutzt werden können (vgl. Baars/Kemper, 2021, S. 35f). Weitere Möglichkeiten wie z.B. Data Lakehouses werden im Kontext dieser Abhandlungen nicht vertieft.

## GRUNDLAGEN DER KÜNSTLICHEN INTELLIGENZ

Die ersten Arbeiten zu KI (engl. „Artificial Intelligence" – AI) entstanden in den 1940er Jahren durch Warren McCulloch und Walter Pitts. Aufbauend auf früheren Überlegungen zum maschinellen „Denken" entwickelten sie das Konzept künstlicher Neuronen mit binären Zuständen und logischen Verknüpfungen nach dem Vorbild organischer, synaptischen Strukturen. 1950 wurde auf Basis dieser Prinzipien der erste Computer mit einem neuronalen Netzwerk entwickelt (Russell & Norvig, 2021, S. 35). Im selben Jahr stellte Alan Turing die Frage „Can machines think?" (Russell & Norvig, 2021, S. 20) und schlug den Turing-Test als Operationalisierung dieser Frage vor. Dieser bewertet die Intelligenz einer Maschine anhand ihrer Fähigkeit, menschliches Verhalten so zu imitieren, dass es nicht von dem eines Menschen unterscheidbar ist (Bartneck et al., 2021, S. 9).

Wenige Jahre später wurde der Begriff „Artificial Intelligence" auf der Dartmouth Conference von John McCarthy et al. geprägt. In ihrem Vorschlag definierten McCarthy et al. KI als die Annahme, dass „every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy, 1955, S. 11). Diese These legte den Grundstein für das Forschungsfeld der KI. In den 80er Jahren prägten insbesondere Expertensysteme die KI-Forschung, in denen explizite Wenn-Dann-Regeln mit Hilfe von Inferenzmechanismen implementiert wurden. Ende der 80er bzw. Anfang der 90er

Jahre erfolgte ein Paradigmenwechsel, in dem maschinelles Lernen an Bedeutung gewann. Anstelle der Formulierung von Regeln steht dort algorithmisches Lernen anhand von Beispielen im Vordergrund.

KI lässt sich als interdisziplinäres Forschungsgebiet aus Informatik, Kognitionswissenschaft, Psychologie, Philosophie, Ökonomie und weiteren Disziplinen begreifen. Dabei zeigt es sich, dass der Begriff „Intelligenz“ nur schwer zu definieren ist (Kaplan, 2021, S. 22). Moderne Definitionen orientieren sich primär an agentenbasierten und rationalen Handlungsmodellen. Neben wissenschaftlichen Definitionen gewinnen regulatorische Perspektiven zunehmend an Bedeutung. Die EU definiert im Artificial Intelligence Act ein KI-System als „a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment […]“.

Teilbereiche der KI lassen sich u.a. nach Entwicklungsstufen differenzieren (Kaplan und Haenlein, 2019, S. 16 f.):

- Artificial Narrow Intelligence (ANI): Dabei handelt es sich um hochspezialisierte, auf spezifische Aufgaben optimierte KI. Hierzu zählen nahezu alle bestehenden KI-Systeme.
- Artificial General Intelligence (AGI) und
- Artificial Superintelligence (ASI): Diese repräsentieren theoretische Konzepte ohne Relevanz für heutige Unternehmensanwendungen.

Bei Machine Learning (ML) handelt es sich um ein Teilgebiet der KI, das Systeme entwickelt, die aus Daten lernen und sich verbessern (Sarferaz, 2023, S. 374). ML umfasst grundsätzlich drei Lernparadigmen:

- Überwachtes Lernen: Zugehörige Modelle lernen aus gelabelten Daten (Herrmann, 2022, S. 37). Hierunter versteht man Daten, denen ein korrektes Ergebnis (Label) zugeordnet wurde.
- Unüberwachtes Lernen: Es erfolgt eine Mustererkennung in ungelabelten Daten. Zu dieser Kategorie zählt auch das sog. Self-Supervised Learning (Herrmann, 2022, S. 37–43). Dabei lernt ein Modell ohne manuell gelabelte Daten, indem es sich eine Trainingsaufgabe selbst aus den vorhandenen Daten erstellt.
- Reinforcement Learning: Lernen erfolgt in diesem Zusammenhang durch Belohnungsmechanismen. Dieser Lerntyp eignet sich bei Systemen, deren Performance evaluiert werden kann, ohne ihnen die richtige Antwort vorzugeben.

Semi-überwachtes Lernen bezeichnet eine Mischform von überwachtem und unüberwachtem Lernen. Es werden sowohl gelabelte als auch nicht annotierte Datensets verwendet: Zunächst erfolgt ein Training mit annotierten Daten und im Anschluss mit den nicht annotierten Daten, um die Lernperformance zu verbessern.

Deep Learning ist eine ML-Kategorie, bei der sog. tiefe neuronale Netze komplexe Mustererkennungen ermöglichen (Altenfelder et al., 2025, S. 3 f.). Es handelt sich dabei um neuronale Netze, die aus vielen hintereinander geschalteten Schichten (Eingabeschicht, versteckte Schichten und Ausgabeschicht) bestehen. Wichtige Architekturen sind Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) sowie Transformer-Modelle. CNNs nutzen Faltungen, um lokale Muster zu erkennen und sind spezialisiert auf räumliche Daten, insbesondere Bilder und Dokumente. RNNs besitzen Rückkopplungsschleifen, die Daten aus vorherigen Schritten speichern. Sie sind für sequenzielle Daten konzipiert. Bei Transformer-Modellen handelt es sich um eine moderne Deep-Learning-Architektur, die 2017 eingeführt wurde. Diese nutzen einen Aufmerksamkeitsmechanismus (Self-Attention). Diese Technik erlaubt es, alle Teile einer Sequenz parallel zu betrachten, kontextabhängige Beziehungen zu erkennen und große Modelle zu trainieren. Die Transformer-Architektur basiert auf einer Encoder-Decoder-Struktur, bei der sowohl der Encoder als auch der Decoder aus gestapelten, identischen Schichten bestehen.

Transformer-Modelle bilden die Grundlage für Foundation Models und Large Language Models (LLMs) wie GPT, BERT oder LLaMA. Bei Foundation Models handelt es sich um vortrainierte KI-Modelle, die auf sehr umfangreichen, breit gefächerten Datensätzen trainiert wurden. Sie fungieren als Grundlage für eine Vielzahl spezialisierter Anwendungen. LLMs sind darauf spezialisiert, natürliche Sprache zu verstehen, zu verarbeiten und zu erzeugen. „Large“ steht in diesem Zusammenhang für die Größe des Modells (das oft aus Milliarden von Parametern besteht), der Menge an Trainingsdaten sowie die Breite der beherrschten Aufgaben.

Der Begriff „Conversational AI“ steht für integriertes Sprachverstehen, Generierung und Dialogsteuerung (Sahu et al., 2024, S. 1). Dabei kommt in modernen Systemen auch Retrieval-Augmented Generation (RAG) zum Tragen, um Antworten auf Unternehmensdaten zu stützen und Halluzinationen zu reduzieren. Die Modelle greifen hierbei auf externe Wissensquellen zu, ohne neu trainiert zu werden.

„Agentic AI“ umfasst digitale Systeme, die in der Lage sind, eigenständig zu handeln, zu planen und Ziele zu verfolgen. Sie lassen sich durch folgende Merkmale charakterisieren (Cheng et al. 2024, S. 1 ff; Huang, 2025, S. 2):

- Zielorientiertheit: Die KI ist auf spezifische Ziele ausgerichtet und erstellt mehrstufige Aktionspläne.
- Wahrnehmung: Das System interpretiert seine Umgebung, Daten und Kontext.
- Autonomie: Agenten sind in der Lage, unabhängig zu agieren bzw. selbstständig Aktionen auszuführen und damit ihre Umgebung zu beeinflussen.

- Schlussfolgerung und Selbstüberwachung: Die KI kann Entscheidungen begründen, Ergebnisse überprüfen und Fehler korrigieren.

## EIGENSCHAFTEN UND FÄHIGKEITEN KI-BASIERTER IT-LÖSUNGEN

Bei der Identifikation und Bewertung von KI-Geschäftspotenzialen gegenüber traditionellen Lösungen lassen sich im Unternehmenskontext verschiedene, untereinander komplementäre Abstraktionsebenen bilden (Bawack et al., 2021, S. 656; Jöhnk et al., 2020, S. 10 f):

(1) Kernfähigkeiten der KI: Hierbei geht es um grundlegende, kognitive Fähigkeiten von KI-Systemen. Im Einzelnen handelt es sich um Wahrnehmung, Verstehen, Handeln und Lernen. In jüngerer Zeit wird auch Generierung als fünfte Kernfähigkeit genannt (Storey et al., 2025, S. 2):

- Wahrnehmen ergibt sich aus der Datenerfassung und -erkennung, beispielsweise aus der Transaktions- und Dokumentenverarbeitung oder aufgrund von Sensordaten.
- Mit Verstehen ist die Interpretation und das semantische Verständnis von Eingaben gemeint. Hierzu gehören z.B. NLP, Kontextanalyse und regelbasierte Schlussfolgerungen.
- Handeln beinhaltet die Ausführung von Aufgaben oder Empfehlungen, beispielsweise im Sinne einer Entscheidungsautomatisierung.
- Lernen ergibt sich aus Leistungsverbesserungen durch Feedback.
- Generieren beinhaltet das Erstellen neuer Inhalte oder Ideen.

(2) KI-Funktionen: Befähigung zur Lösung abstrakter, an menschlicher Kognition angelehnte Aufgaben durch konkrete Aktivitäten. Während die Kernfähigkeiten beschreiben, was ein KI-System grundsätzlich tun kann, erklären KI-Funktionen, wie sich diese Fähigkeiten in geschäftsrelevanten kognitiven Aufgaben manifestieren. Hofmann et al. (2020, S. 10 f.) schlagen sieben KI-Funktionen vor:

- Wahrnehmen: Erfassen und Verarbeiten relevanter Daten,
- Merkmalsextraktion & Identifikation: Erkennen zentraler Merkmale aus Rohdaten,
- Schlussfolgern: Ableiten von Zusammenhängen und Beziehungen,
- Vorhersagen: Prognose zukünftiger Ereignisse oder Ergebnisse,
- Entscheiden: Auswahl optimaler Handlungsoptionen,
- Generieren: Erstellen neuer Artefakte, Inhalte oder Lösungen,
- Handeln: Ausführen von Aufgaben oder Empfehlungen.

Diese Funktionen bilden eine Brücke zwischen abstrakten Fähigkeiten und konkreten Geschäftsapplikationen.

(3) Anwendungsmuster: Standardisierte, wiederkehrende Vorgehensweisen zur Anwendung von KI zur Generierung eines Mehrwerts. Hierdurch sollen Geschäftsziele wie Prozessautomatisierung, Erkenntnisgewinn oder verbesserte Nutzerinteraktion operationalisiert werden (Hasan, 2022, S. 446; Sarferaz, 2024, S. 123 f.). Zu den zentralen Anwendungsmustern gehören:

- Matching: Identifikation von Zusammenhängen zwischen Daten, z. B. Zuordnung von Zahlungen zu offenen Forderungen (Sarferaz, 2024, S. 124).
- Recommendation: Unterstützung von Entscheidungen durch Vorschläge auf Basis historischer Daten (Sarferaz, 2024, S. 125).
- Ranking: Priorisierung, z. B. Sortierung von Forderungen nach Zahlungseingangswahrscheinlichkeit (Sarferaz, 2024, S. 126).
- Prediction: Prognose zukünftiger Ergebnisse (Sarferaz, 2024, S. 126 f.).
- Categorization: Zuordnung von Daten zu vordefinierten Klassen (Sarferaz, 2024, S. 127).
- Conversational Patterns: Natürliche Sprachinteraktion, z. B. Chatbots (Sarferaz, 2024, S. 127 f.).

Die Trennung in verschiedene Abstraktionsebenen ermöglicht eine systematische Zuordnung zu Geschäftsanforderungen, KI-Fähigkeiten und technischen Designelementen.

| **KI-Anwendungsmuster** | **KI-Funktionen** | **Kernfähigkeiten** |
|---|---|---|
| Prediction | Vorhersagen | Lernen |
| Recommendation | Schlussfolgern, Entscheiden | Verstehen, Handeln |
| Categorization | Merkmalsextraktion & Identifikation, Schlussfolgern | Verstehen |
| Matching | Merkmalsextraktion & Identifikation | Wahrnehmung |
| Ranking | Schlussfolgern, Entscheiden | Verstehen, Handeln |
| Conversational Interface | Generieren, Handeln, Wahrnehmen | Generierung, Handeln |

Abb. 1: Zuordnung von KI-Anwendungsmustern, Funktionen und Fähigkeiten
Quelle: Bay, A. (2025, S. 16)

Während KI-Fähigkeiten die grundlegenden Bausteine von KI-Systemen beschreiben, liefern KI-Funktionen eine stärker geschäftsorientierte Operationalisierung dieser Fähigkeiten. Sie eignen sich daher besonders, um Nutzerherausforderungen mit potenziellen KI-gestützten Lösungen zu verknüpfen. Im vorliegenden Artikel wird der Schwerpunkt auf KI-Funktionen gelegt:

## ANFORDERUNGEN, CHANCEN UND RISIKOPOTENZIALE FÜR DEN KI-EINSATZ

Für die Mensch-Maschine-Kooperation lassen sich je nach Einsatzszenario unterschiedliche Autonomiegrade unterscheiden: In assistierenden Systemen „Human in-the-loop"(HITL) unterstützt die KI bei Aufgaben, die Menschen grundsätzlich auch selbst ausführen könnten. In erweiterten Systemen („Human-on-the-loop") bereitet die KI komplexe Entscheidungen vor, die nicht immer vollständig nachvollziehbar sind. In diesem Fall liegt die finale Aufsicht und Kontrolle weiterhin beim Menschen (Dellermann et al., 2021; Schmalzried et al., 2023).

Die Integration von KI in eine bestehende IT-Landschaft bringt technische, organisatorische, ethische und regulatorische Herausforderungen mit sich. Obwohl KI verbesserte Automatisierung, Entscheidungsfindung und Prozessoptimierung verspricht, wird die Realisierung dieser Vorteile häufig durch organisatorische, technische und ethische Hürden behindert (Mhaskey, 2024, S. 4 f.).

Eine grundlegende Voraussetzung ist die Verfügbarkeit hochwertiger, strukturierter Daten. So erzeugen beispielsweise ERP-Systeme große Mengen an Transaktions- und Stammdaten, die die Grundlage für KI-Modelle bilden können. Allerdings beeinträchtigen Dateninkonsistenzen, fragmentierte Systemlandschaften und unzureichende historische Datensätze häufig die Wirksamkeit von KI (Duan et al., 2019, S. 64; Mhaskey, 2024). Aus architektonischer Sicht erhöht die Komplexität moderner IT-Umgebungen die Schwierigkeiten zusätzlich.

Neben diesen technischen Herausforderungen beinhaltet die organisatorische Bereitschaft einen entscheidenden Faktor (Jöhnk et al., 2020, S. 11). Viele Anwender aus den Unternehmensbereichen verfügen nicht über notwendige digitalen Kompetenzen, um KI-Werkzeuge effektiv zu verstehen, anzuwenden und zu bewerten. Dies spiegelt nicht nur einen Mangel an technischen Fähigkeiten wider, sondern auch begrenzte Erfahrung im Umgang mit KI-Systemen. Die Schließung dieser Lücke ist daher nicht nur eine Frage der Schulung, sondern erfordert die Schaffung eines lernorientierten Umfelds.

Weiterhin erweist sich die Nutzerakzeptanz als kritischer Erfolgsfaktor für die Nutzung von KI-Fähigkeiten. Diese hängt zum einen von der wahrgenommenen Nützlichkeit und Benutzerfreundlichkeit von KI-Systemen ab. Zum anderen spielt Vertrauen eine zentrale Rolle. Anwender aus den Fachbereichen benötigen transparente und nachvollziehbare Ausgaben von KI-Systemen (Zebec & Štemberger, 2024, S. 9). Explainable-AI-(XAI-)Techniken können helfen, algorithmische Entscheidungsfindung zu entmystifizieren und menschliche Aufsicht zu unterstützen, was essenziell für das Vertrauen der Nutzer ist (Kaplan & Haenlein, 2019, S. 23; Weber et al., 2023, S. 872 f.). Ferner spielt die adäquate Einbindung in Unternehmensprozesse eine wichtige Rolle, beispielsweise um bessere Entscheidungen zu unterstützen, ohne die Komplexität zu erhöhen (Sarferaz, 2025, S. 6).

Eng damit verknüpft sind ethischen Aspekte und regulatorischer Compliance-Lösungen (Jöhnk et al., 2020, S. 13). Die Sicherstellung von Fairness, Transparenz und Verantwortlichkeit ist eine Voraussetzung für den langfristigen Vertrauensaufbau in integrierte KI-Funktionalitäten. KI-Empfehlungen müssen beispielsweise frei von Verzerrungen (Bias), nachvollziehbar und korrigierbar sein. Rechtsvorschriften wie die DSGVO (Mhaskey, 2024, S. 4) und der EU-AI-Act (2020, S. 9 f.) setzen strenge Grenzen für Datennutzung und -verarbeitung, insbesondere in sensiblen Bereichen wie Personalwesen und Finanzen. Rechnungslegungsvorschriften verlangen Nachvollziehbarkeit und Prüfbarkeit (Köbler et al., 2022, S. 12). Diese rechtlichen Anforderungen verstärken die Notwendigkeit für KI-Systeme, die nicht nur präzise, sondern auch überprüfbar und erklärbar sind (OECD, 2019, S. 32).
Ferner gilt es aus Unternehmenssicht, die wirtschaftliche und operative Machbarkeit einer KI-Skalierung zu evaluieren (Mhaskey, 2024, S. 4 f.). Während Pilotprojekte häufig vielversprechende Ergebnisse zeigen, erfordert der Übergang in produktive Umgebungen Modell-Retraining und kontinuierliches Monitoring (Jöhnk et al., 2020, S. 12). Der Aufbau nachhaltiger Governance-Strukturen und die Maximierung des langfristigen Return on Investment sind daher ein wichtiger Erfolgsfaktor (Mhaskey, 2024, S. 5).

Um diese multidimensionalen Herausforderungen zu bewältigen, muss ein Unternehmen umfassende KI-Leitprinzipien etablieren, die über die reine Gesetzeskonformität hinausgehen und explizit Fairness, Transparenz, menschliche Aufsicht und Verantwortlichkeit über den gesamten KI-Lebenszyklus hinweg betonen. Gleichzeitig sollte die Komplexität der KI-Integration weniger als Barriere, denn als Katalysator für organisatorische Transformation und Lernen betrachtet werden. Die Diskrepanz zwischen verfügbaren KI-Fähigkeiten und der aktuellen Bereitschaft vieler Unternehmen schafft Raum für die Entwicklung neuer Kompetenzen, Prozesse und Denkweisen. Nach Zebec und Štemberger (2024, S. 17 f.) ist organisatorisches Lernen ein zentraler vermittelnder Faktor, der KI-Adoption in verbesserte Prozess- und Entscheidungsqualität übersetzt und die Bedeutung eines lernorientierten ERP-Umfelds hervorhebt.

## EVALUIERUNG EINES PRAXISBEISPIELS

Zur Veranschaulichung der praktischen Anwendung dient das im Folgenden beschriebene Praxisbeispiel. Das betrachtete Unternehmen ist ein international tätiger Hersteller von Fahrrädern, Mountainbikes, Rasenmähern und verschiedenartigen Kinderrädern (z.B. Dreiräder, Kettcars, Tretroller). Die Planung und Budgetierung erfolgten bislang in einem stark zentralisierten, hierarchisch geprägten Top-down-Prozess. Die Fortschreibung der Plan- und Ist- Werte basierte primär auf prozentualen

Anpassungen der Vorjahresdaten. Eine systematische Berücksichtigung von Einflussfaktoren fand dabei keine Berücksichtigung. In Anbetracht der zunehmenden Komplexität des Produktportfolios, einhergehend mit gestiegenen Anforderungen an Transparenz und Steuerungsqualität, wird die Einführung eines harmonisierten Planungs- und Budgetierungssystems verfolgt. Ausgangspunkt dieses Ansatzes ist die Entwicklung und Implementierung eines Werttreiberbaums in Microsoft BI, der die zentralen Key Performance Indicators der Unternehmensleistung strukturiert abbildet und quantitativ in einem Dashboard verknüpft.

## LÖSUNGSSCHRITTE ZUR EINFÜHRUNG EINES DASHBOARDS MIT WERTGENERATOREN IN POWER BI

Ist die Grundkonzeption des Kennzahlensystems und die dazugehörige Hierarchie der Werttreiber definiert sowie die Datenanbindung sichergestellt kann mit dem Aufbau des Dashboards begonnen werden (siehe Abb. 2).

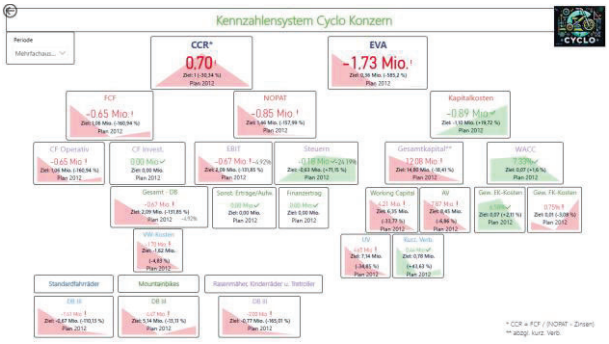



Abb. 2: Power BI Visualisierung eines Werttreiberbaums
Quelle: Eigene Darstellung

Die definierte Hierarchie der Wertgeneratoren und Spitzenkennzahl(en) sollte sich auch bei der Visualisierung in Power BI widerspiegeln. Daher sollte die Anordnung der Kennzahlen so gewählt werden, dass daraus hervorgeht, in welcher Beziehung sie zueinanderstehen und wie sie sich beeinflussen. Eine passende Anordnung erleichtert somit einerseits das Verständnis des Kennzahlensystems und anderseits die Analyse sowie Interpretation der Ergebnisse (siehe Abb. 2). In dem hier vorliegenden Fallbeispiel wurde sich für zwei Spitzenkennzahlen entschieden. So werden auf der Ertragsseite einerseits die Wertgeneratoren zur Spitzenkennzahl Economic Value Added (EVA®) aggregiert wohingegen die Liquiditätsseite durch die Cash Conversion Rate (CCR) abgebildet wird. Die Anordnung der darunterliegenden Wertgeneratoren zeigt hierbei, dass bspw. der EVA® aus den Wertgeneratoren Net Operating Profit after Tax (NOPAT) und den Kapitalkosten besteht. Die Kapitalkosten werden wiederum aus dem Gesamtkapital (Hier wird das Gesamtkapital abzgl. der kurzfristige Verbindlichkeiten interpretiert, da diese mit einem kurzfristigen Abfluss an liquiden Mitteln führen und somit den Wertgenerator Working Capital und in der Folge das Gesamtkapital mindern) sowie dem gewichteten Kapitalkostensatz (Weight Average Costs of Capital (WACC)) ermittelt, wohingegen das Gesamtkapital aus dem Working Capital und Anlagevermögen (AV) abgeleitet wird. Die unterste Wertgeneratoren-Ebene der Kapitalkosten wird durch das Umlaufvermögen und den kurzfristigen Verbindlichkeiten gebildet. Der zweite Ast der Ertragsspitzenkennzahl EVA® mündet im NOPAT. Dieses ermittelt sich aus den Earnings before Interest and Taxes (EBIT) abzgl. der Steuern. Die Wertgeneratoren des EBIT sind aus dem Gesamt-Deckungsbeitrag (DB), den Sonstigen Erträgen und Aufwendungen sowie dem Finanzertrag definiert. Der Gesamt-DB wird aus der Summe der Produktlinien-DB III abzgl. der Verwaltungskosten (VW-Kosten) bestimmt. Das Kennzahlenmodell spiegelt durch die Trennung des DB III in die Produktlinien die Organisationsstruktur des Unternehmens wider und ermöglicht eine erste Einschätzung der Ertragssituation dieser. Eine weitere Aufgliederung wäre grundsätzlich möglich, führt jedoch zu einer zunehmenden Unübersichtlichkeit des Dashboards. Es empfiehlt sich somit weitergehende Thrill-down Visualisierungen bspw. in Form der Deckungsbeitragsanalyse in einer separaten Visualisierung aufzubereiten. Die Liquiditätsspitzenkennzahl CCR wird als Quotient des Free Cashflow (FCF) und das NOPAT abzgl. Zins berechnet. Der Free Cashflow wird hierbei als Summe der Cashflows aus operativer und Investitionsbedingter Tätigkeit bestimmt.

Für eine effektive Unternehmenssteuerung ist es zudem wichtig, dass das Dashboard einen direkten Vergleich der IST-Werte mit den Plan-Werten ermöglicht bzw. die Plan-Werte mit Zielwerten aus dem strategischen Planungsprozess verglichen werden können. Für eine schnelle Erfassung dieser Abweichungen spielt eine entsprechende Farbgebung (Farbcodierung) eine wichtige Rolle. So sollten Abweichungen, die einen negativen Einfluss auf den entsprechenden Wertgenerator haben rot geschrieben sein, wohingegen Abweichungen mit positivem Einfluss grün dargestellt werden sollten. Die Farbcodierung mithilfe dieser Ampelfarben ermöglicht es den Anwendern Abweichungen schnell zu erkennen und zu interpretieren. Sie lenkt damit den Fokus auf positive und problematische Ausprägungen der Wertgeneratoren, die einer negativen Abweichung einer genaueren Analyse bedürfen. Dies erhöht die Effizienz im Rahmen der Unternehmenssteuerung. Die Farbcodierung vermittelt jedoch lediglich eine nominale Information (Abweichung positiv oder negativ) und nicht die Ausprägung der Abweichungshöhe. Für Anwender des Dashboards kann jedoch diese Information ebenso wichtig sein, um zu erkennen, ob die Zielsetzung nur knapp oder deutlich über- bzw. unterschritten wurde. Die Angabe einer prozentualen Abweichung zum Zielwert kann hierbei zielführend sein und ermöglicht Anwendern eine schnelle und einfache Interpretation der Abweichungshöhe (siehe Abb. 3).

Darüber hinaus können sich Anwender ebenfalls für die Entwicklung der Wertgeneratoren und Spitzenkennzahl(en) im Zeitverlauf interessieren, z.B. wenn die Entwicklung im Rahmen einer mehrjährigen Budgetplanung analysiert werden soll. Ein Liniendiagramm als Trendlinie kann hierbei auf einen Blick darstellen, wie sich die Kennzahlen über die betrachteten Perioden entwickelt hat. Auch hierbei ist die Verwendung einer entsprechenden Farbcodierung mithilfe der Ampelfarben sinnvoll, um die Informationsinterpretation zu vereinfachen (siehe Abb. 2).


Abb. 3: KPI-Visualisierungselement (sog. Visual) in Power BI
Quelle: Eigene Darstellung

Power BI bietet für die obigen Anforderungen das „KPI"-Visualisierungselement (sog. „KPI"-Visual) an. Diese Visualisierung ermöglicht es im ersten Schritt die Kennzahl (Key Performance Indicator (KPI)) festzulegen, die dargestellt werden soll. Darüber hinaus wird die Trendlinie über die betrachteten Perioden im Hintergrund des Visuals visualisiert und entsprechend der Entwicklung mit einer Farbcodierung versehen. Ebenfalls ist die Hinterlegung des Zielwertes möglich (siehe Abb. 4). Das Visual vergleicht dessen Ausprägung automatisch mit dem tatsächlichen Wert und passt daraufhin die Farbcodierung des Visuals an. Die Abweichung mit einem negativen Einfluss auf den Wertgenerator wird zusätzlich zu der roten Ampelfarbe ein rotes Ausrufezeichen angegeben sowie die prozentuale Abweichung berechnet (siehe Abb. 2). Wird ein Zeitraum visualisiert, zeigt das Visual die Werte der letzten Periode an (in Abb. 2 bspw. „Plan 2012"). Über die Formatierung des Visuals kann definiert werden, welche Entwicklung der Trendlinie bzw. Abweichung zum Zielwert als positiv bzw. negativ interpretiert werden soll (siehe Abb. 4 – rechter roter Kasten).

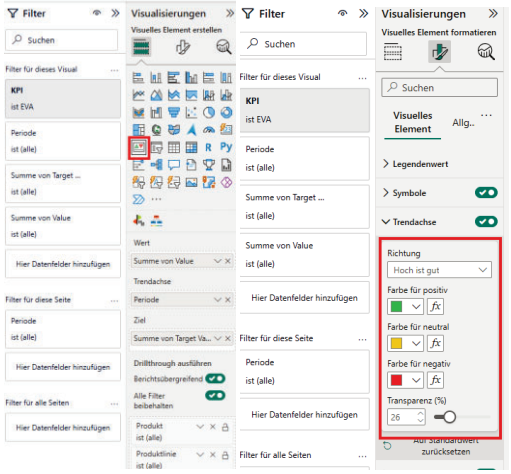



Abb. 4: Spezifikation des KPI-Visuals in Power BI
Quelle: Eigene Darstellung

Für manchen Wertgeneratoren kann es zudem sinnvoll sein auch ihren relativen Wert als Zusatzinformation anzugeben (bspw. EBIT in Prozent vom Umsatz oder Steuerquote). Möchte man dies mit dem „KPI"-Visual umsetzen, müsste zusätzlich zur Visualisierung des absoluten Wertes dieses Wertgeneratoren ein weiteres Visual kreiert werden. Dies kann jedoch eine Unübersichtlichkeit generieren. Eine andere Möglichkeit ist es diese Information in das bestehende „KPI"-Visual der absoluten Darstellung des Wertgenerators zu integrieren. Hierfür kann ebenfalls das „KPI"-Visual genutzt werden, jedoch wird dieses ohne Zielwert und ohne Trendlinie angegeben, da diese Information bereits mit der Darstellung des absoluten Wertes abgebildet ist. Dies kann erreicht werden, indem das Visual mit transparentem Hintergrund formatiert wird, sodass eine Einpassung in das bestehende Visual des Wertgenerators möglich ist (siehe Abb. 5).

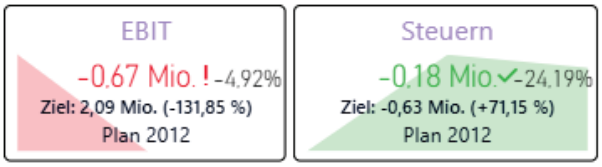



Abb. 5: „KPI"-Visual inklusiver der Darstellung der relativen Ausprägung der Wertgeneratoren
Quelle: Eigene Darstellung

Zuletzt sollte ebenfalls berücksichtigt werden, dass das Dashboard im Rahmen einer „Self-Service" Nutzung der Anwender auch eine Dynamisierung durch unterschiedliche Auswertmöglichkeiten ermöglicht. Für Kennzahlensysteme ist häufig die Entwicklung über mehrere Perioden für die Anwender von Interesse. Es empfiehlt sich daher ein Filtermenü einzufügen, das den Anwendern eine nutzerspezifische Auswahl der Perioden ermöglicht. So sind die Anwender selbst in der Lage eigene Analysen vorzunehmen und unterschiedliche Perioden und die

dazu gehörigen Entwicklungen zu vergleichen (bspw. Budget- und IST-Werte wie in Abb. 6 dargestellt).

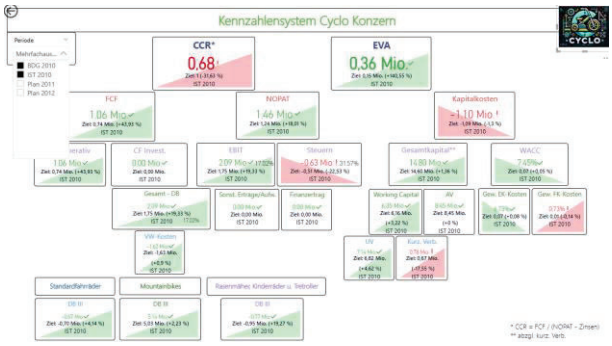



Abb. 6: Filterfunktion zur Analyse der Periode Budget 2010 und IST-Werte 2010
Quelle: Eigene Darstellung

Wird Power BI im geschäftlichen Umfeld und mit einer Pro- oder Premium-Lizenz genutzt ist zudem das Hinzufügen von Kommentaren zu einem Dashboard möglich (Vgl. Microsoft, URL). Dies kann besonders hilfreich sein, um Abweichungen direkt im Dashboard zu erläutern und dem Management qualitative Hintergründe über die Ausprägung und Entwicklung der Wertgeneratoren und Spitzenkennzahl(en) zu geben. Somit werden alle für die Unternehmenssteuerung wichtigen quantitativen und qualitativen Informationen an einem Ort verfügbar gemacht. Das Dashboard fungiert damit als sog. „Single Source of Truth“.
Eine sinnvolle Weiterentwicklung des Werttreiberbaums stellt die erweiterte Deckungsbeitragsrechnung dar, da sie eine präzisere Analyse der Kosten- und Ergebnisstruktur ermöglicht. Durch die Aufteilung der Kosten in weitere Deckungsbeitragsstufen wird eine systematische Untersuchung von Kostensensitivitäten und eine verbesserte Bewertung der Auswirkungen von Veränderungen in Bezug auf Mengen, Preise oder Kosten erzielt. Dementsprechend können ökonomische Risiken frühzeitig lokalisiert werden und fundierte Entscheidungen hinsichtlich des Produktportfolios, der Marktstrategien sowie der Internationalisierung abgeleitet werden.

## HANDLUNGSEMPFEHLUNGEN UND FAZIT

Traditionelle Kenntnisse des Controllers bleiben unerlässlich zur Kennzahlenbestimmung und Analyse. Neue Kenntnisse der Einbindung dieser Steuerungskennzahlen operativ und strategisch (vgl. Binder, Morelli 2025, S.52 ff.) mit Hilfe von Wertgeneratoren in ein Power-BI erfor-dern weitere vertiefte IT-Kenntnisse der Controller. Dar-über hinaus ist der Einsatz von Dashboards mit Unterstützung von KI die Entwicklung der Zukunft bei der die Rohdaten (z.B. in Excel) per Texteingabe (Prompt) hochgeladen werden können, so dass automatisiert Diagramme und Key Performance Indicators (KPIs) generiert werden können. Damit übernimmt die KI die Datenanalyse, Layoutgestaltung und Visualisierung, so dass der Controller bei der Erarbeitung eines Dashboards damit hilfreich unterstützt werden kann.

## LITERATUR


Altenfelder, K., Kieffer-Radwan, S., & Schönfeld, D. (2025). Einleitung: Wie die Algorithmen Einzug in das Services Management gehalten haben, Die historische Entwicklung der Künstlichen Intelligenz. In K. Altenfelder, S. Kieffer-Radwan, & D. Schönfeld, Services Management und Künstliche Intelligenz, Grundlagen und Anwendungsfelder für den Einsatz von KI-unterstützten Service (pp. 1-18). Wiesbaden: Springer Nature. doi:https://doi.org/10.1007/978-3-658-46665-7

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). An Introduction to Ethics in Robotics and AI. Cham, Switzerland: Springer Nature. doi:https://doi.org/10.1007/978-3-030-51110-4

Bawack, R. E., Wamba, S. F., & Carillo, K. D. (2021). A framework for understanding artificial intelligence research: insights from practice. Journal of Enterprise Information Management, 34(2), 645-678

Bay, Alisa (2025). Artificial Intelligence in Financial Closing: Identifying, Designing, And Demonstrating AI-Driven Use Cases for an Enterprise Software Company, Master´s Thesis, University of Ljubljana. School of Economics and Business

Binder B.C.K., Morelli F. (2025), Digitale Zwillinge zur Optimierung des Budgetierungsprozesses – Dashboards und Process Mining unterstützen ein prozessorientiertes Performance Measurement, in: Industry 4.0 Science, Ausg. 2/2025, S. 52 - 58

Britzelmaier, B. (2020): Controlling. Grundlagen, Praxis, Handlungsfelder, 3. Auflage, Pearson Studium, München

Cheng, Yuheng; Zhang, Ceyao; Zhang, Zhengwen; Meng, Xiangrui; Hong, Sirui; Li, Wenhao; Wang, Zihao; Wang, Zekai; Yin, Feng; Zhao, Junhua und He, Xiuqiang (2024): Exploring Large Language Model based Intelligent Agents: Def-initions, Methods, and Prospects. arXiv Preprint arXiv:2401.03428, URL: https://arxiv.org/pdf/2401.03428.pdf, abgerufen am 08.11.2025

Dellermann, D., Lipusch, N., Ebel, P., & Leimeister, J. M. (2021). CrowdServ –Konzept für ein hybrides Entscheidungsunterstützungssystem zur Validierung von Geschäftsmodellen. In D. Beverungen, J. H. Schumann, V. Stich, & G. Strina (Eds.), Dienstleistungsinnovationen durch Digitalisierung: Band 2: Prozesse – Transformation – Wertschöpfungsnetzwerke (pp. 299–331). Springer. https://doi.org/10.1007/978-3-662-63099-0_8

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management, Vol. 48, 63-71.

doi:https://doi.org/10.1016/j.ijinfomgt.2019.01.021

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 168, 1-159.

Gladen, W. (2014). Performance Measurement- Controlling mit Kennzahlen. Wiesbaden: Springer Gabler.doi:10.1007/978-3-658-05138-9

Gleich, R., Gänßlen, S., Kappes, M., Kraus, U., Leyk, J., Tschandl, M. (2015). Moderne Instrumente der Planung und Budgetierung: Innovative Ansätze und Best Practive für die Unternehmenssteuerung. 2. Auflage. Haufe Verlag. München

Gleich, T., Kappes, M., Leyk, J. (2019). Planung, Budgetierung und Forecasting: Innovative und digitale Instrumente für die Unternehmenssteuerung. 1. Auflage. Haufe Verlag. München

Hasan, A. R. (2022). Artificial Intelligence (AI) in Accounting & Auditing: A Literature Review. Open Journal of Business and Management. 10, 440-465. doi:https://doi.org/10.4236/ojbm.2022.101026

Herrmann, J. (2022). Künstliche Intelligenz mit den Themenschwerpunkten maschinelles Lernen und künstlichen neuronalen Netzen, dargestellt anhand des Beispiels von Blended-Learning-Übungen. In C. Aichele, & J. Herrmann, Betriebswirtschaftliche KI-Anwendungen, Digitale Geschäftsmodelle auf Basis Künstlicher Intelligenz (2. ed., pp. 17-76). Wiesbaden: Springer. doi:https://doi.org/10.1007/978-3-658-40099-6

Hofmann, P., Jöhnk, J., Protschky, D., & Urbach, N. (2020). Developing purposeful AI use cases - A structured method and its application in project management. 15. International Conference on Wirtschaftsinformatik (WI) 2020. Proceedings. 1, pp. 1-16. Potsdam: GITO Verlag. doi:10.30844/wi_2020_a3-hofmann

Horváth, P., Gleich, R., Seiter, M. (2024). Controlling. 15. überarbeitete Auflage. München. Vahlen.

Huang, K. (2025). The Genesis and Evolution of AI Agents. In K. Huang, Agentic AI, Theories and Practices (pp. 1-22). Cham, Switzerland: Springer Nature. doi:https://doi.org/10.1007/978-3-031-90026-6

Inmon, W. H. (2005). Building the Data Warehouse (4. Ausg.). New York: Wiley Computer. Publishing.

Jöhnk, J., Weißert, M., & Wyrtki, K. (2020). Ready or Not, AI Comes - An Interview Study of Organizational AI Readiness Factors. Business & Information Systems Engineering, 63, 5-20

Kaplan, A. (2021). Artificial Intelligence (AI): When Humans and Machines Might Have. In P. Verdegem, AI for Everyone? Critical Perspectives (pp. 21-32). London: University of Westminster Press. doi:https://doi.org/10.16997/book55.b

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25. doi:https://doi.org/10.1016/j.bushor.2018.08.004

Köbler, M., La Kier, H. A., Prabakar, S., & Six, M. (2022). Introducing SAP S/4HANA Cloud for Advanced Financial Closing. Bonn: Rheinwerk Verlag GmbH

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Hanover, New Hampshire: Dartmouth College.

Mhaskey, S. (2024). Integration of Artificial Intelligence (AI) in Enterprise Resource Planning (ERP) Systems: Opportunities, Challenges, and Implications. International Journal of Computer Engineering in Research Trends, 11(12), 1-9

Microsoft. (URL). Hinzufügen von Kommentaren zu einem Dashboard oder Bericht. Abgerufen am 15.03.2016 von https://learn.microsoft.com/de-de/power-bi/explore-reports/end-user-comment

OECD. (2019). Artificial Intelligence in Society. Paris: OECD Publishing. doi:https://doi.org/10.1787/eedfee77-en

Reichmann, T. (2011). Controlling mit Kennzahlen - Die systemgestützte Controlling-Konzeption mit Analyse- und Reportingsinstrumenten (8. Ausg.). München: Vahlen.

Russell, S. J., & Norvig, P. (2021). Artificial Intelligence, A Modern Approach (4. ed.). Harlow: Pearson.

Sarferaz, S. (2023). ERP-Software: Funktionalität und Konzepte, Basierend auf SAP S/4HANA. Wiesbaden: Springer Nature.

Sarferaz, S. (2024). Embedding Artificial Intelligence into ERP Software, A Conceptual View on Business AI with Examples from SAP S/4HANA. Cham, Switzerland: Springer Nature. doi:https://doi.org/10.1007/978-3-031-54249-7

Sarferaz, S. (2025). Implementing Generative AI into ERP Software. IEEEAccess, 1-13. doi:10.1109/ACCESS.2025.3564133

Schmalzried, D., Hurst, M., Wentzien, M., & Gräser, M. (2023). Analyse der Rolle künstlicher Intelligenz für eine menschenzentrierte Industrie 5.0. HMD Praxis der Wirtschaftsinformatik, 60(6), 1143–1155. https://doi.org/10.1365/s40702-023-01001-y

Schön, D. (2022). Planung und Reporting im BI-gestützten Controlling - Grundlagen, Business Intelligence, Mobile BI, Big-Data-Analytics und KI (4 Ausg.). Wiesbaden: Springer Gabler. doi:https://doi.org/10.1007/978-3-658-35475-6

Storey, V., Yue, W., Zhao, J., & Lukyanenko, R. (2025). Generative Artifcial Intelligence: Evolving

Technology, Growing Societal Impact, and Opportunities for Information Systems Research. Information Systems Frontiers,, 1-22. doi:https://doi.org/10.1007/s10796-025-10581-7

Weber, J., Bramsemann, U., Heineke, C., & Hirsch, B. (2017). Wertorientierte Unternehmensführung - Konzepte-Implementierung-Praxis-Statement (2. Ausg.). Wiesbaden: Springer Gabler. doi:10.1007/978-3-658-15216-1

Zebec, A., & Štemberger, M. (2024). Creating AI business value through BPM capabilities. Business Process Management Journal, 30(8), 1-26. doi:10.1108/BPMJ-07-2023-0566

KONTAKT

Bettina C.K. Binder ist Professorin für Controlling, Prozessmanagement und Strategische Unternehmensführung an der Hochschule Pforzheim. Zuvor war sie 14 Jahre in der Managementberatung und als Finance Director tätig.

Frank Morelli ist Professor für Wirtschaftsinformatik an der Hochschule Pforzheim. Er forscht im Bereich Business Intelligence und hat die Hochschule Pforzheim als Celonis Academic Center of Excellence etabliert.

Alisa Bay ist Business Processes Consultant bei der SAP Deutschland SE & Co. KG mit dem Schwerpunkt Finance. Sie absolvierte ein Double-Degree-Masterstudium in Information Systems und Business Informatics an der Hochschule Pforzheim und University of Ljubljana.

Florian Rottner ist Controlling Professional bei der Siemens Energy Global GmbH & Co. KG im Geschäftsbereichcontrolling. Er absolvierte ein Duales Studium mit dem Schwerpunkt BWL-Industrie an der Dualen Hochschule Baden-Württemberg in Karlsruhe. Zurzeit studiert Florian Rottner im Master Controlling, Finance und Accounting an der Hochschule Pforzheim.

Katharina Marie-Therese Schmidt ist Studentin im Masterstudiengang Controlling, Finance & Accounting an der Business School Pforzheim. Ihren Bachelorabschluss erwarb sie 2025 am Campus Künzelsau der Hochschule Heilbronn mit einer Abschlussarbeit zur Implementierung einer Self-Service BI Anwendung im Controlling eines mittelständischen Unternehmens.

# Analyse und Optimierung des Anforderungsmanagements am Beispiel des Bereichs Global Engineering der Zollner Elektronik AG

Professor Dr. Frank Herrmann
*Innovationszentrum für Produktionslogistik und Fabrikplanung*
*Ostbayerische Technische Hochschule Regensburg*
Galgenbergstraße 32, 93053 Regensburg
Germany
Frank.Herrmann@OTH-Regensburg.de

Denise Beil
*Zollner Elektronik AG*
Postgasse 4 93462 Lam
Germany
denise_beil@t-online.de

***Abstract*—Das vorliegende Projekt beschäftigt sich mit dem Fachgebiet Requirements-Engineering. Im Mittelpunkt der Arbeit steht die Betrachtung des Requirements-Engineerings eines international agierenden Elektronikdienstleisters. Die Erstellung eines Konzepts zu dessen Durchführung bildet hierbei das Ziel der Arbeit. Zu Beginn werden die gängigen Vorgehensweisen im Bereich Requirements-Engineering erläutert. Anhand einer IST-Analyse wird die aktuelle Situation aufgenommen und mithilfe einer Nutzwertanalyse bewertet. Im Anschluss daran werden Kriterien zur Optimierung des Requirements-Engineerings ausgewählt. Zu jedem Kriterium werden mögliche Lösungen gesammelt, um so eine höhere Bewertung zu erzielen. Aus der Menge der Lösungen ergeben sich verschiedene Lösungsvarianten zur Optimierung. Hieraus wird die Lösung ausgewählt, welche die größte Nutzwertsteigerung liefert. Abschließend werden die notwendigen Schritte zur Umsetzung der Lösungsvariante in einer Handlungsempfehlung beschrieben.**

***Keywords*—Anforderungsmanagement, Anforderung, Requirements-Engineering, Requirements-Management.**

## I. Einleitung

Wie aktuelle Studien belegen, stellen fehlerhafte, unvollständige oder unklare Anforderungen einen der Hauptgründe für das Scheitern von Projekten dar. Aus diesem Grund gewinnt das Requirements-Engineering (RE) in den verschiedensten Branchen an immer größerer Bedeutung. In der Abteilung Global Engineering (GE) der Zollner Elektronik AG, gibt es zum gegenwärtigen Zeitpunkt keine einheitliche Vorgehensweise, für die Bearbeitung eingehender Anforderungen. Vor diesem Hintergrund sollen die notwendigen Rahmenbedingungen für das RE im Bereich GE erarbeitet werden, wodurch eine Beurteilung der aktuellen Situation auf Basis des Stands der Technik ermöglicht wird. Dabei ist aufzuzeigen, welche Aufgaben des REs bereits gut umgesetzt sind und bei welchen Themen Handlungsbedarf besteht.

Das Ziel dieses Projekts besteht darin, ein Konzept für die Durchführung des REs zu erstellen. Durch die spätere Umsetzung der Ausarbeitungen soll der Nutzwert für das RE im Bereich GE nachweislich gesteigert werden. Zusätzlich ist die Verwendbarkeit von Checklisten bzw. Templates innerhalb des REs zu klären. Darüber hinaus soll eine geeignete Software zur Dokumentation und Verwaltung der Anforderungen vorgeschlagen werden.

## II. Requirements-Engineering

### A. Notwendigkeit

Der Hauptgrund für das Scheitern von Projekten bzw. für das Nichterreichen der Marktziele von Produkten ist ein mangelhaftes RE. Dies wird von verschiedenen Studien belegt. Die jährlichen Studien von Gartner liefern beispielsweise folgendes Resultat: Obwohl das RE nur 2-5% des Projektaufwands darstellt, haben die Fehler in Anforderungen die größten Auswirkungen. [9, vgl. S. 3] Hinzu kommt, dass ein Anforderungsfehler umso höhere Kosten verursacht, je später er im Projektverlauf entdeckt wird. [6, vgl. S. 10]

### B. Begriffsdefinitionen

**Anforderung**

Um den Begriff der Anforderung zu definieren, wird in der Literatur beispielsweise bei Rupp, Ebert, Pohl und Hruschka auf die Begriffsdefinition des Institute of Electrical and Electronics Engineers (IEEE) verwiesen. Das IEEE, ein weltweiter Berufsverband von Ingenieuren [1, vgl. S. 13], hat den Anforderungsbegriff im Standard 610.12-1990 wie folgt festgelegt.

Zum einen bezeichnet eine Anforderung eine Eigenschaft oder Fähigkeit, die von einem Benutzer zur Lösung eines Problems oder zur Erreichung eines Ziels benötigt wird. Der Benutzer kann hierbei ein System oder auch eine Person sein. Außerdem stellt eine Anforderung eine Eigenschaft oder Fähigkeit dar, die ein System oder Teilsystem erfüllen oder besitzen muss, um einen Vertrag, eine Norm, eine Spezifikation oder

weitere formale Vorgaben zu erfüllen. Eine Anforderung ist zudem, eine dokumentierte Repräsentation einer Eigenschaft oder Fähigkeit, wie in den ersten beiden Punkten beschrieben. [2, vgl. S. 65]

Der oben genannte Standard des IEEE wurde im Jahre 2011 von dem neuen Standard ISO/IEC/IEEE 29148 abgelöst. Im neuen Standard des IEEE wird der Begriff der Anforderung nur sehr knapp behandelt, dennoch wird zur Vollständigkeit darauf eingegangen. Eine Anforderung ist demnach als eine Aussage, welche ein Bedürfnis und die damit verbundenen Einschränkungen und Bedingungen ausdrückt bzw. übersetzt, definiert. [3, vgl. S. 9]

Mit der Anforderungsspezifikation gibt es ein eigenes Dokument, in dem alle beschriebenen Anforderungen zusammengefasst werden. [1, vgl. S. 12]

**Requirements-Engineering**
Laut dem International Requirements-Engineering Board (IREB), einem Expertengremium welches sich mit der Zertifizierung von Fachkräften im Bereich RE beschäftigt, handelt es sich beim RE um eine Disziplin zur Spezifikation und zum Management von Anforderungen unter Erreichung bestimmter Ziele. Ein Ziel ist es, die erforderlichen Anforderungen zu kennen, Übereinstimmung der Stakeholder im Bezug auf die Anforderungen zu erreichen, die Anforderungen nach bestimmten Standards zu dokumentieren und sie systematisch zu managen. Außerdem sollen die Wünsche und Bedürfnisse der Stakeholder verstanden und dokumentiert werden. Des Weiteren soll ein System ausgeliefert werden, das auch mit den Wünschen und Bedürfnissen der Stakeholder übereinstimmt, deshalb sind die Anforderungen dementsprechend zu spezifizieren und zu managen. [5] [1, vgl. S. 13]

Der Begriff des RE wird auch als (System-)analyse [1, vgl. S. 15] oder Anforderungsanalyse [1, vgl. S. 10] bezeichnet.

In mehreren Werken zum RE, wie z.B. bei Hruschka, Ebert, Pohl oder auch Rupp, erfolgt eine Gliederung des REs in wenige Kernaktivitäten. Laut Rupp und den SOPHISTen [1, vgl. S.14] ist das RE in vier Haupttätigkeiten gegliedert. Dazu gehören die Ermittlung, das Dokumentieren, das Prüfen und Abstimmen und die Verwaltung von Anforderungen. Die einzelnen Aktivitäten werden in den nachfolgenden Kapiteln näher erläutert.

**Requirements-Management**
Das Requirements-Management (RM) gehört zu den vier Haupttätigkeiten des REs und bildet mit der Verwaltung der Anforderungen einen Teilbereich hiervon. Gleichbedeutende Begrifflichkeiten sind Anforderungsverwaltung oder Anforderungsmanagement (AM). [1, vgl. S. 368] Laut der Definition von Pohl und Rupp umfasst das RM alle für die Dokumentation, Änderung und Nachverfolgung der Anforderungen, erforderlichen Maßnahmen. [4, vgl. S. 14 f. + S. 125] [1, vgl. S. 368]

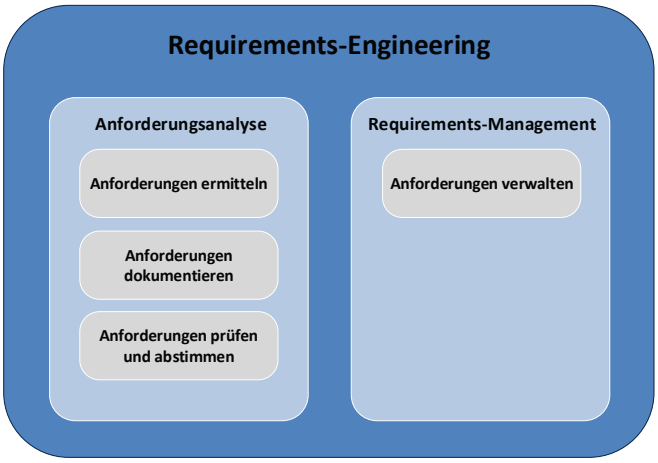



Figure 1. Zusammenhang von RE und RM.

### C. Vorgehensweisen

Betrachtet man die zeitliche Einordnung des REs, ergeben sich aus der Literatur zwei grundsätzliche Ansätze für die Vorgehensweise. Da sich dieses Projekt im Gegensatz zur Literatur nicht auf Entwicklungsprojekte bezieht, wurden die folgenden Ansätze der einzelnen Vorgehensweisen an Projekte allgemein angepasst.

*1) RE als Phase:*
Einerseits wird das RE in Modellen wie z.B. dem Wasserfallmodell oder dem V-Modell, als einzelne Phase betrachtet. Der RE-Prozess wird für jedes Projekt zu Beginn des eigentlichen Projektes durchlaufen. Am Ende der Phase geht die Anforderungsspezifikation hervor, die alle Anforderungen für dieses Projekt beinhaltet und Informationsgeber für die weiteren Phasen ist. Die Anforderungsermittlung wird immer nur für ein einziges Projekt durchgeführt und ist somit völlig autark im Bezug auf sonstige Projekte. Hierbei kann weiter zwischen dem „Sequenziellen RE“ und dem „Parallelen RE“ unterschieden werden. Die zeitliche Einordnung der einzelnen Projekte grenzt diese beiden Unterarten voneinander ab. [6, vgl. S. 30 f.] Bei der sequentiellen Vorgehensweise werden die kompletten Anforderungen für das beispielhafte Projekt A ermittelt und spezifiziert. Im Anschluss daran beginnt Projekt A. Erst wenn Projekt A komplett abgeschlossen ist, kann mit der RE-Phase für ein neues Projekt, in diesem Fall für Projekt B, begonnen werden. [6, vgl. S. 31] Beim parallelen Vorgehen werden die Anforderungen weiterhin getrennt für die Projekte A, B und C ermittelt und dokumentiert. Der Unterschied zum sequentiellen RE besteht darin, dass beim parallelen RE die RE-Phase und die Projektarbeit simultan ablaufen können. Während der Durchführung des Projektes B kann beispielsweise schon mit der RE-Phase für Projekt C begonnen werden. [6, vgl. S. 30 f.]

*2) Kontinuierliches RE:*
Das kontinuierliche RE zeichnet sich durch phasen- und projektübergreifende Eigenschaften aus. Dieser Ansatz ist bereits in vielen Vorgehensmodellen verwirklicht, wobei sich

die Stärke der Umsetzung je nach Projekt unterscheidet. Beispiele für solche Modelle sind der Rational Unified Process oder Scrum. Demnach fallen die in der Literatur oftmals separat behandelten agilen Vorgehensweisen bei diesem Ansatz mit unter das kontinuierliche RE. Pohl unterscheidet zwei Qualitätsstufen innerhalb des kontinuierlichen REs: Das phasenübergreifende und das projektübergreifende RE. Diese beiden Aktivitäten werden im Folgenden näher erläutert. [6, vgl. S. 34] Beim phasenübergreifenden RE erstreckt sich die RE-Phase über den gesamten Projektzeitraum, wodurch die Anforderungen beständig und nachvollziehbar erfasst und verwaltet werden. Die Projekte A, B und C stehen beispielsweise mit dem zugehörigen RE über ihren gesamten Lebenszyklus in Kontakt. Dadurch erhält das RE beispielsweise Änderungsanträge aus dem laufenden Projekt. Veränderungen werden somit während des kompletten Entwicklungsprozesses erkannt und ausgewertet, wodurch eine frühzeitige Berücksichtigung ermöglicht wird. [6, vgl. S. 34 f.] Das projektübergreifende Vorgehen zeichnet sich durch eine andauernde Erfassung und Verwaltung der Anforderungen aus. Diese ist nicht zwingend an einzelne Projekte gebunden und stellt somit eine Erweiterung zum phasenübergreifenden RE dar. Dadurch wird die Zusammenstellung von Anforderungen für ein neues System aus der Menge der bereits bekannten Anforderungen ermöglicht. Anschließend wird das Projekt auf Grundlage der Anforderungen durchgeführt. Veränderungen von Anforderungen oder Kontext werden fortlaufend aktualisiert. Anforderungen können sowohl für Projekt B als auch für Projekt C gelten. [6, vgl. S. 35]

### *D. Tätigkeiten des Requirements-Engineerings*

*1) Anforderungen ermitteln:*
Der erste Schritt in der Anforderungsanalyse, die Ermittlung der Anforderungen, beginnt mit der Identifizierung der notwendigen Anforderungsquellen. Die drei typischen Anforderungsquellen werden im folgenden beschrieben. [1, vgl. S. 77]

Wichtige Informationen zur Generierung von Anforderungen können in Dokumenten enthalten sein. Beispiele hierfür sind Gesetzestexte, Normen oder andere Standards. Die darin enthaltenen Vorgaben können erhebliche Auswirkungen auf Projekte haben, indem sie z.B. bestimmte Standards für die Entwicklung vorgeben. Außerdem dienen branchen- oder organisationsspezifische Dokumente als weitere Anforderungsquelle. Hierzu zählen etwa Anforderungsdokumente, Schulungsunterlagen oder Fehlerberichte des Altsystems und Dokumentationen zu Prozessen und Arbeitsanweisungen. [1, vgl. S. 77]

Eine weitere Quelle für Anforderungen bilden Systeme im Betrieb. Darunter fallen Alt- bzw. Vorgängersysteme oder auch Konkurrenz- und Nachbarsysteme. Die aktuellen Systeme können von den Stakeholdern getestet werden. So ist es möglich, Entscheidungen bezüglich Erweiterungen oder Änderungen leichter zu treffen. [1, vgl. S. 77]

Unter einem Stakeholder versteht man eine Person oder Organisation mit potenziellem Interesse am zukünftigen System. Aus diesem Grund gehen von den Stakeholdern auch Anforderungen an dieses System ein, wodurch sie eine weitere wesentliche Anforderungsquelle bilden. [6, vgl. S. 65] [1, vgl. S. 77]

Für die unterschiedlichen Arten von Anforderungen gibt es verschiedene Techniken, um diese zu ermitteln.

Durch den Einsatz von Kreativitätstechniken ist es möglich, neuartige Ideen zu erarbeiten und so die unterbewussten Begeisterungsfaktoren zu ermitteln. Zu diesen kreativen Techniken gehören das Brainstorming, die Methode 6-3-5 und der Wechsel der Perspektive. [1, vgl. S. 99]

Um die Basisfaktoren zu erhalten, die die meisten Stakeholdern voraussetzen oder ihnen unterbewusst bekannt sind, eignen sich Beobachtungstechniken. Dazu gehören beispielsweise die Feldbeobachtung, wobei die Stakeholder bei der Ausführung ihrer Tätigkeiten beobachtet werden. Des Weiteren können Anforderungen durch Apprenticing, also durch Erlernen der Tätigkeiten der Stakeholder, ermittelt werden. [1, vgl. S. 103 f.]

Die Befragungstechniken zählen zu den klassischen Methoden zur Ermittlung von Anforderungen mit unterschiedlichster Detailtiefe. Wichtig ist, dass sich der Stakeholder den Anforderungen bewusst sein muss und sie sprachlich beschreiben kann. Durch geschicktes Einsetzen dieser Technik können so Leistungsfaktoren, Basisfaktoren und auch nicht-funktionale Anforderungen ermittelt werden. Die Durchführung von Befragungstechniken erfolgt durch den Einsatz von Fragebögen oder Interviews. [1, vgl. S. 105 f.]

Bei den artefaktbasierten Techniken geht es um die Wiederverwendung von Lösungen oder Erfahrungen aus bestehenden Systemen. Die Technik ermittelt die gesamte Funktionalität und das detaillierte Verhalten bestehender Systeme. Somit werden die Basisfaktoren des Systems aufgedeckt und zusätzlich enthaltene Leistungsfaktoren erkannt. Rupp empfiehlt eine kombinierte Anwendung der artefaktbasierten Techniken mit anderen Ermittlungstechniken, um gültige, alte und neue Anforderungen zu erhalten. Die Technik wird weiter in die Systemarchäologie und den Reuse untergliedert. [1, vgl. S. 110]

Die Systemarchäologie ermittelt Anforderungen anhand des vorhandenen Systems oder Dokumenten zum jeweiligen System. Man kann sich hierbei immer tiefer vorarbeiten wie z.B. bei Softwareprojekten bis hin zur Code-Recherche. [1, vgl. S. 110 f.]

Eine weitere Möglichkeit um Anforderungen artefaktbasiert zu ermitteln ist der Reuse, also die Wiederverwendung. Wurde bereits ein Projekt durchgeführt, welches Ähnlichkeiten zum aktuellen Projekt aufweist, können Dokumente oder sogar Anforderungen aus dem früheren Projekt für das aktuelle wiederverwendet werden. Deshalb ist auch eine gute

Dokumentation der Anforderungen sinnvoll, um diese auch wiederverwenden zu können. In der Praxis werden Probleme oft einfach neu gelöst, anstatt nach einer bereits bestehenden Lösung zu suchen. Dies kann mit dem Einsatz des Reuse verhindert werden. [1, vgl. S. 111 f.]

Zur Steigerung der Effizienz und Qualität der oben genannten Ermittlungstechniken gibt es die Kategorie der unterstützenden Techniken. Diese können ergänzend zu verschiedenen Ermittlungstechniken angewandt werden. Beispiele hierfür sind Audio- und Videoaufzeichnungen, Workshops oder Mind Mapping. [1, vgl. S. 112]

*2) Anforderungen formulieren und dokumentieren:*

Die Verwaltung aller Anforderungen eines Projekts wird in den Anforderungsdokumenten realisiert. [6, vgl. S. 231] Die gebräuchlichsten Bezeichnungen für die Dokumente sind Anforderungsdokument, Anforderungsspezifikation, Requirements Specification oder Requirements-Spezifikation. [7, vgl. S. 235] Oft ist eine Aufteilung der Anforderungsdokumente in z.B. Lasten- und Pflichtenheft notwendig. Die Erstellung der Anforderungsdokumente unterscheidet sich je nach Vorgehensmodell. Das kontinuierliche RE generiert die Anforderungsdokumente mithilfe der vorliegenden aktuellen Menge von Anforderungen. Beim phasenorientierten RE stellen die Anforderungsdokumente das Ergebnis einer Projektphase dar. Das Lastenheft und das Pflichtenheft sind die zwei Typen von Anforderungsdokumenten, die sich laut Pohl im deutschen Sprachraum durchgesetzt haben. [6, vgl. S. 231]

Die Dokumentation der einzelnen Anforderungen kann entweder natürlichsprachlich oder mithilfe von unterschiedlichsten Diagrammen erfolgen. [1, vgl. S. 186] Für den Bereich GE bei der Zollner Elektronik AG, wird der Fokus auf die natürlichsprachliche Dokumentation gelegt.

Anforderungen die in natürlicher Sprache dokumentiert werden, nennt man auch Prosaanforderungen. Diese Art der Anforderungsdokumentation ist die am meisten genutzte Technik zur Dokumentation. Ein großer Vorteil dieser Technik besteht darin, dass die Sprache allen Beteiligten ausreichend bekannt ist. Zugleich kann die Verwendung der natürlichen Sprache auch einige Probleme mit sich bringen. [1, vgl. S. 187]

Dazu gehören die Tilgung, die Verzerrung und die Generalisierung, die auch sprachliche Effekte genannt werden. Bei der Tilgung werden Teile der Information, die als selbstverständlich oder dem Empfänger der Information bekannt angesehen werden, weggelassen. Geht es dabei um Anforderungen können so relevante Informationen übersehen werden. Eine Umgestaltung oder schlimmsten Falls Verfälschung der Wirklichkeit tritt bei dem Effekt der Verzerrung auf. Hierbei wird eine reale Tatsache derart abgeändert, dass es beim Betrachter zu einem Zerrbild kommt. Neue Informationen werden so besser mit bestehenden Vorstellungen des Menschen zusammengefasst. Dadurch können wiederum wichtige Informationen verloren gehen. Der Vorgang der Generalisierung überträgt einmalige Erfahrungen auf andere ähnliche Situationen oder verwandte Sachverhalte und macht sie somit allgemeingültig. Dies kann beim RE nützlich sein. Eine zu starke Generalisierung sollte vermieden werden, denn dadurch können globale Anforderungen entstehen, die nur für einen Teil des Systems angemessen sind, wodurch Fehler- und Sonderfälle übersehen werden. [1, vgl. S. 129 f.]

Um die eben beschriebenen sprachlichen Effekte bei der Anforderungsdokumentation zu minimieren, schlägt Rupp den Einsatz des SOPHIST-REgelwerks und die Verwendung von Anforderungsschablonen vor. [1, vgl. S. 216 f.]

Das so genannte SOPHIST-REgelwerk beschäftigt sich mit der Identifizierung sprachlicher Effekte in Anforderungen. Hierbei werden verlorene oder verfälschte Informationen analysiert und die sprachlichen oder inhaltlichen Mängel durch Umformen eliminiert. [1, vgl. S. 133 ff.] Für diesen Vorgang wurden von Rupp 18 Regeln aufgestellt, die es zu beachten gilt. Diese sind wiederum nach ihrer Priorität in die Klassen „hoch“, „mittel“ und „niedrig“ aufgeteilt. [1, vgl. S. 163] Da der Aufwand dafür, jede Anforderung nach diesen 18 Regeln zu untersuchen, sehr hoch ist und in der Realität meist nicht umgesetzt werden kann, ist es möglich nur einige der Regeln gemäß der Priorität anzuwenden. [1, vgl. S. 216]

Der schablonenbasierte Ansatz nach Rupp hilft dabei, den Aufwand zur Dokumentation der Anforderungen signifikant zu verringern. Die Technik ist leicht anwendbar und liefert Anforderungen hoher Qualität unter optimalen Zeit- und Kostenfaktoren. Außerdem eignen sich die Anforderungsschablonen neben dem Aufbau auch zur späteren Qualitätssicherung der Anforderungen. Das Prinzip hierfür ist denkbar einfach: Es geht darum wie die Anforderung syntaktisch korrekt aussieht. Damit wird bereits zu Beginn eine Vielzahl der klassischen Fehler bei der Formulierung verhindert. Ein Beispiel hierfür sind Passivsätze, welche nichts darüber aussagen für wen die jeweilige Funktion gedacht ist. Der große Vorteil dieser Schablonen ist die Tatsache, dass alle Anforderungen einen gleichartigen Aufbau besitzen, deshalb werden sie auch syntaktische Anforderungsschablonen genannt. Rupp stellt für verschiedene Arten der Anforderungen eine eigene Schablone vor. [1, vgl. S. 217 ff.] Der Ansatz der Anforderungsschablonen wird auch von Ebert [9, vgl. S. 107], Hruschka [7, vgl. S. 129], Pohl [6, vgl. S. 246] und Niebisch [8, vgl. S. 106] zur Dokumentation vorgeschlagen bzw. in Anlehnung an Rupp übernommen.

Figure 2 zeigt das Beispiel einer Anforderungsschablone für funktionale Anforderungen in Anlehnung an Rupp [1, vgl. S. 220] und Ebert [9, vgl. S. 107]. Die Anforderungen werden mithilfe der Schablone schrittweise aufgebaut. Im ersten Schritt werden das System, das Produkt oder die Komponenten benannt. Als Nächstes wird die Wichtigkeit unter Verwendung von „muss“, „soll“ oder „wird“ festgelegt. Danach wird die Funktion des Systems durch ein Vollverb, auch Prozesswort genannt, ausgedrückt. Anschließend wird die Art der Funktionalität des Systems festgelegt. Es kann einerseits „fähig sein“ etwas zu tun, also etwas selbsttätig

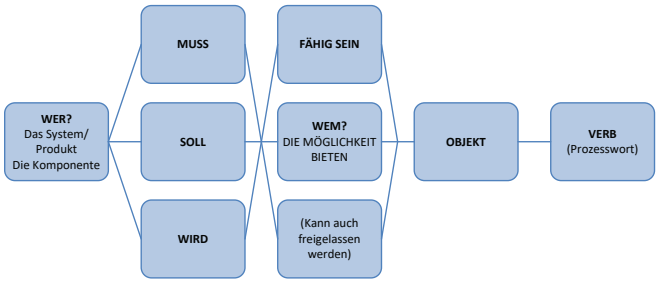



Figure 2. Beispiel einer Schablone für funktionale Anforderungen.

durchführen, andererseits kann es eine Interaktionsmöglichkeit anbieten oder etwas in Abhängigkeit eines Dritten ausführen. Gibt es ein Objekt, auf das sich das festgelegte Prozesswort bezieht, wird dieses als Nächstes definiert. Zuletzt können Bedingungen angegeben werden, die eine Durchführung der Funktionalität auslösen. [1, vgl. S. 220 ff.]

Durch die Verwendung der Anforderungsschablone wurde bereits der Beitrag zur syntaktischen Korrektheit der einzelnen Anforderungen gemacht. Damit ist die Anforderung aber noch nicht auf dem gewünschten Qualitätsniveau. Um das zu erreichen, ist es notwendig, die semantischen Inhalte in der Anforderungsschablone richtig zu füllen. Eine gleiche Verwendung der Wortwahl ist durch mehrere Personen nur dann gewährleistet, wenn der verfügbare Wortschatz eingeschränkt ist. Zudem müssen die zu verwendenden Wörter eindeutig festgelegt werden. [1, vgl. S. 225]

Die Schlüsselwörter zur Festlegung der rechtlichen Verbindlichkeit können beispielsweise in einer Tabelle dargestellt werden. Zu den einzelnen Schlüsselwörtern wird dessen Bedeutung für alle Beteiligten verständlich formuliert. [1, vgl. S. 226]

Auch die verwendeten Prozesswörter, also Vollverben, sollten klar definiert werden. Somit soll sichergestellt werden, dass die Beteiligten alle genau das Gleiche unter einem Prozesswort verstehen. Umgesetzt werden kann dies anhand einer so genannten Prozesswortliste. Diese Liste enthält alle Vollverben die innerhalb des Projekts verwendet werden. Rupp schlägt den Aufbau der Liste weiterhin so vor, dass eine Spalte für die Prozesswörter, eine Spalte für deren Bedeutung und eine weitere Spalte für Synonyme zum jeweiligen Prozesswort existieren. Gibt es für ein Wort verschiedene Interpretationsmöglichkeiten, sollte die gewünschte Bedeutung ebenfalls angegeben werden. Für die Definition des betreffenden Objekts werden in der Regel Substantive verwendet. Auch diese sollten, wie die Prozesswörter, in einer entsprechenden Liste, dem Glossar, geführt werden. Eventuelle Abkürzungen für verwendete Substantive sollten in dieser Liste ergänzt werden. Enthalten Anforderungen zudem eine Bedingung, sollte diese genau und einheitlich formuliert werden. Hilfreich ist es hierzu beispielsweise die Bedingungen immer mit den Konjunktionen zu beginnen. [1, vgl. S. 227 ff.]

*3) Anforderungen prüfen und abstimmen:*
Mit der Prüfung und Abstimmung von Anforderungen wird gewährleistet, dass die Anforderungen unter Einhaltung ihrer Qualitätskriterien dokumentiert wurden. [4, vgl. S. 101] Laut Hruschka spielt der Begriff des Quality Gateways, abgekürzt Quality Gate, eine zentrale Rolle bei der Prüfung von Anforderungen. Die dokumentierten Anforderungen werden zur Überprüfung durch solche Quality Gates getrieben, um die Einhaltung der Qualitätskriterien zu einem bestimmten Zeitpunkt sicherzustellen. [7, vgl. S. 276] Betrachtet man die zeitliche Lage der Quality Gates gibt es drei verschiedene Möglichkeiten: Es wird „einmal“, „öfter“ oder „andauernd“ durchlaufen. Beim phasenorientierten RE gibt es beispielsweise am Ende der Analysephase ein zentrales und einmaliges Quality Gate. Hruschka empfiehlt jedoch, mehrere Quality Gates in sein eigenes Vorgehensmodell einzubauen und rät zudem zum regelmäßigen Prüfen der Anforderungen durch solche Gateways. Fest steht, dass die Quality Gates benötigt werden, um die genauen Zeitpunkte zur Prüfung von Anforderungen zu definieren. [7, vgl. S. 277]

Pohl und Rupp definieren die folgenden drei Hauptziele bzw. Qualitätsaspekte der Prüfung von Anforderungen: Inhalt, Dokumentation und Abgestimmtheit. [4, vgl. S. 103 ff.]

Werden Anforderungen auf inhaltliche Fehler untersucht, fällt dies unter den Qualitätsaspekt „Inhalt“. Enthält eine Anforderungen inhaltliche Fehler ziehen sich diese auch durch alle folgenden Aktivitäten im Projekt. Beispiele für inhaltliche Fehler sind die Verletzung der Qualitätskriterien für Anforderungen (und das Anforderungsdokument). [4, vgl. S.104]

Beim Prüfen von Anforderungen nach dem Qualitätsaspekt „Dokumentation“ geht es um Fehler in der Dokumentation oder um Verstöße bei der Einhaltung festgelegter Vorschriften zur Dokumentation. Solche Fehler können nachfolgende Entwicklungsaktivitäten aufhalten. Anforderungen können von den Empfängern dadurch nicht mehr verstanden werden. Es kann ebenfalls zu nicht dokumentierten Informationen kommen. Außerdem können Anforderungen übersehen werden, da sie sich nicht am richtigen Platz innerhalb der Spezifikation befinden. [4, vgl. S. 105 f.]

Mängel in der Abstimmung von Anforderungen zwischen relevanten Stakeholdern betreffen den Qualitätsaspekt „Abgestimmtheit“. Da die Stakeholder während des REs neues Wissen über die Planung des Systems erhalten, ist es möglich, dass sich ihre Meinung zu bereits abgestimmten Anforderungen verändert. Beim Prüfen können die Stakeholder Änderungswünsche benennen ohne nachfolgende Tätigkeiten zu beeinflussen. [4, vgl. S. 106]

Um die Einhaltung der oben genannten Punkte zu überprüfen, gibt es verschiedene Techniken. Für dieses Projekt sind die Reviews und der Einsatz von Checklisten besonders interessant.

Der Großteil der verwendeten Techniken zum Prüfen der Anforderungen sind manuelle Techniken, diese werden unter

dem Oberbegriff Review zusammengefasst. Sie lassen sich weiter in Stellungnahme, Inspektion und Walkthrough unterteilen. Nachfolgend wird die Stellungnahme beschrieben, da diese Technik auch für dieses Projekt relevant ist. Die Anforderungen werden dabei vom Autor an eine dritte Person (beispielsweise einen Kollegen) weitergegeben. Dieser soll die Anforderungen anschließend auf definierte Qualitätskriterien untersuchen. Die gefundenen Mängel werden dokumentiert und kurz beschrieben. [4, vgl. S. 109 f.]

Eine weitere Technik zur Überprüfung von Anforderungen ist der Einsatz von Checklisten. Für das Prüfen von Anforderungen enthält eine Checkliste Fragen, welche die Prüfer von Anforderungen beim Ermitteln von Fehlern unterstützt. In der Praxis ist dies laut Rupp und Pohl eine gängige Technik. Um Checklisten benutzen zu können, müssen vorher die jeweiligen Fragen oder Aussagen dafür festgelegt werden. Quellen für diese Fragen können die Qualitätskriterien für Anforderungen (und Anforderungsdokumente), die drei Hauptziele des Prüfens von Anforderungen oder die potentielle Erfahrung von Prüfern aus abgeschlossenen Projekten sein. Eine Checkliste kann als Leitfaden für die Prüfer dienen. Des Weiteren kann sie als Strukturierungshilfe eingesetzt werden und Fragen enthalten, die beantwortet werden müssen. Somit wird die Arbeit verschiedener Prüfer einheitlich und vergleichbar. Um eine erfolgreiche Anwendung von Checklisten zu erzielen, sollten folgende Punkte eingehalten werden. Bei Verwendung einer Checkliste sollte darauf geachtet werden, diese regelmäßig zu aktualisieren und auf dem neusten Stand zu halten. Um eine gute Handhabbarkeit zu unterstützen, sollte der Umfang der Liste eine Seite nicht überschreiten. Außerdem sind die Fragen möglichst präzise zu formulieren, um viele verschiedene Antworten auszuschließen. [4, vgl. S. 116 f.]

Abschließend wird auf ein wichtiges Prinzip bei der Prüfung von Anforderungen eingegangen. Das Prinzip schreibt die Trennung von Fehlersuche und Fehlerkorrektur vor. Die gefundenen Mängel werden beim Prüfen der Anforderungen nur dokumentiert. Erst anschließend wird für jeden Mangel untersucht, ob dieser tatsächlich einen Fehler darstellt. Diese Trennung fördert die Konzentration auf die Fehlersuche. Die eigentliche Korrektur erfolgt erst nach dem Prüfen, wodurch keine zusätzlichen Fehler produziert werden. [4, vgl. S. 106 f.]

*4) Anforderungen verwalten:*

Das RM beschäftigt sich mit der Verwaltung von Anforderungen. Zu den Tätigkeiten in diesem Bereich gehört die Vergabe von Attributen, die weiterführende Informationen über eine Anforderung angeben. Eine weitere Aufgabe ist die Generierung von Sichten für eine bestimmte Absicht, also die Auswahl von speziellen Daten aus der gesamten Informationsmenge. Außerdem wird die Priorität von Anforderungen festgelegt, wenn aufgrund finanzieller Kapazitäten eine Auswahl getroffen werden muss. Um eine Basis für zukünftige Tätigkeiten zu schaffen, sollten die Anforderungen auch versioniert werden. Überdies ist die Verfolgbarkeit von Anforderungen (Traceability) zu gewährleisten. Für die Änderung von bestehenden Anforderungen ist zudem ein Änderungsmanagement notwendig. [7, vgl. S. 304]

**Attributierung**

Unter der Attributierung von Anforderungen versteht man die Vergabe von Zusatzinformationen für die Anforderungen. Für die Frage, welche zusätzlichen Informationen eine Anforderung besitzen sollte, verweist Hruschka auf das Template von Volere. Demnach ist die Formulierung der Anforderung, also deren Beschreibung, zwingend notwendig. Außerdem sollte als weiteres Attribut eine eindeutige Bezeichnung bzw. Nummer vergeben werden. Diese ermöglicht eine einfache Bestimmung und Referenzierung der Anforderungen. Auch die Angabe einer Quelle sollte unter den Attributen einer Anforderung zu finden sein. Damit sind die Personen gemeint, die an dieser Anforderung beteiligt sind, wie beispielsweise Ersteller, Prüfer, Verantwortlicher der Anforderung. Da jedes Zusatzattribut einen Aufwand bedeutet und auch Zeit kostet, sollte man sich genau überlegen, wie viele Quellen hier angegeben werden. Diese sind so gering wie möglich zu halten. Die Vorlage von Volere enthält zudem Felder zur Angabe der Kundeneinschätzung. Somit ist es möglich die Anforderungen nach Basis-, Leistungs- und Begeisterungsfaktoren darzustellen. Steht eine Anforderung im Konflikt zu einer anderen Anforderung, sollte dies ebenfalls in einem Attribut vermerkt werden. Als weitere Information soll auch die Priorität als Attribut enthalten sein. Zusätzlich kann die Abhängigkeit von Anforderungen zu anderen Anforderungen als Attribut gepflegt werden. Die Angabe von Versionsnummer oder verantwortlichen Personen, sollte im Idealfall für das gesamte Anforderungsdokument angegeben werden, was einen enormen Arbeitsaufwand erspart. In der Praxis kann dies bei vielen Beteiligten oder bei sich unterschiedlich ändernden Anforderungen nicht umgesetzt werden, dann müssen die Attribute auf einer tieferen Ebene vergeben werden. [7, vgl. S.305 ff.]

**Sichtenbildung**

Die Möglichkeit zur Bildung von bestimmten Sichten, ergibt sich aus der vorherigen Attributierung von Anforderungen. Bei der Sichtenbildung wird nach gewünschten Attributen gefiltert. Ausgewählt werden kann z.B. nach Stakeholder, Status oder auch Anforderungsart. [7, vgl. S. 310]

**Priorisierung**

Die Priorisierung der Anforderungen ist eine weitere Aufgabe des RMs. Meist werden hierfür dreistufige Bewertungsmöglichkeiten beginnend mit „A: hohe Priorität“ der zweiten Abstufung „B: nice-to-have“ und der letzten Stufe „C: nicht zu schaffen“ verwendet. [7, vgl. S. 311]

Die Priorisierung kann außerdem aufgrund verschiedener Kriterien vorgenommen werden. Ein mögliches Merkmal ist die Kundenzufriedenheit, daneben kann auch die Größe bzw. das bisherige Auftragsvolumen des Kunden miteinbezogen werden. Des Weiteren könnte der Aufwand bezüglich Zeit und Budget berücksichtigt werden. Ebenso ist es möglich, eine Priorisierung anhand der Risiken einzelner Anforderun-

gen durchzuführen. Die Dringlichkeit einzelner Anforderungen eignet sich ebenfalls zur Priorisierung. Im Bereich der Finanzen können auch die Auswirkungen der Anforderung auf Unternehmensumsätze eine Rolle zur Festlegung der Rangfolge spielen. [7, vgl. S. 311 f.]

Aus den Vorschlägen werden einige Kriterien zur Priorisierung ausgewählt. Diese werden von verschiedenen Personen nach Punkten bewertet, woraus sich am Ende eine gewichtete Priorität ergibt. Laut Hruschka werden die Prioritäten in der Praxis meist durch eine Bewertung mit A, B oder C nach dem „Bauchgefühl“ aufgestellt. [7, vgl. S. 313]

**Versionierung**
Anforderungen können sich über die Dauer des Lebenszyklus eines Systems verändern. Deshalb empfehlen Pohl und Rupp die Anforderungen angemessen zu versionieren. Dieses Vorgehen erlaubt den Zugriff auf spezifische Änderungsstände einzelner Anforderungen während des gesamten Lebenszyklus eines Systems. Eine solche Version einer Anforderung kennzeichnet sich durch eine eindeutige Versionsnummer und einen inhaltlichen Stand der Änderung. Eine Verwaltung der Versionen kann sowohl für Sätze, einzelne textuelle Anforderungen, Teile eines Anforderungsdokuments als auch Anforderungsmodelle durchgeführt werden. Eine Versionsnummer besteht aus den Teilen Version und Inkrement. Das Inkrement wird bei kleinen inhaltlichen Änderungen erhöht. Wurden größere inhaltliche Änderungen vorgenommen, wird die Version erhöht und das Inkrement auf den Wert „0“ zurückgesetzt. Zur Verdeutlichung der Verwendung als Versionsnummer kann diese auch mit einem „V“ begonnen werden. [4, vgl. 142 f.]

**Traceability**
Zwischen allen Informationen die im Laufe des REs zusammengetragen werden, bestehen verschiedene Verbindungen. [1, vgl. S. 421] Mit dem Bereich Traceability, auch Nachverfolgbarkeit oder Verfolgbarkeit genannt, wird eine Nachverfolgung der Anforderung über den kompletten Lebenszyklus des Systems ermöglicht. [4, vgl. S. 136] Rupp erweitert diese Definition der Traceability insofern, dass alle „Verbindungen und Abhängigkeiten zwischen Informationen, welche während der Analyse, Entwicklung, Wartung und Weiterentwicklung eines Systems anfallen, jederzeit nachvoll[zogen werden] können“. [1, S. 421]

Dabei verfolgt die Traceability folgende Ziele:

- Die Umsetzung der Anforderung kann nachgewiesen werden
- Bestimmte Dokumente können zur Wiederverwendung in anderen Projekten genutzt werden
- Auswirkungen auf Dokumente aufgrund von Änderungen einzelner Anforderungen können ermittelt werden
- Der Aufwand für die Umsetzung einer Anforderung kann ermittelt werden
- Es können Abschätzungen bezüglich des Aufwands für Fehlerbehebungen oder Ursachenforschungen durchgeführt werden

[1, vgl. S. 422]

Es gibt drei verschiedene Arten dieser Verbindungen. Die Eltern-Kind-Verbindung existiert zwischen Anforderungen mit unterschiedlicher Detailebene. Diese Verbindung stellt die Verfeinerung dar. Bei der Verfeinerung wird eine Anforderung in eine oder mehrere detailliertere Anforderungen unterteilt. Die ursprüngliche Anforderung und ihre Verfeinerungen besitzen gleichzeitig Gültigkeit. Zum Beispiel kann eine der vorgestellten natürlichsprachlichen Anforderungen anhand von einer oder mehreren detaillierteren natürlichsprachlichen Anforderungen ausführlicher beschrieben werden. Mithilfe der Traceability können solche Anforderungen exakt dokumentiert und die Verbindungen nachvollzogen werden. [1, vgl. S. 422 f.]

Des Weiteren gibt es eine Verbindung zwischen Anforderungen auf der gleichen Ebene. Darin ist auch die Verbindung einer Anforderung mit ihrer Vorgänger- oder Nachfolgerversion eingeschlossen. Diese Verbindungen können eine fachliche Abhängigkeit, funktionale Abhängigkeit, Ausnahme, Randbedingung oder Realisierungsvorgabe sein. [1, vgl. S. 425]

Die dritte Möglichkeit ist die Verbindung zwischen verschiedenen Informationsarten. Hierzu zählen z.B. außer Anforderungen auch Interviewprotokolle. So wird im Falle einer Anforderungsänderung ermittelt, welche dazugehörigen Dokumente ebenfalls geändert werden müssen. [1, vgl. S. 425 f.]

Die oben genannten Verbindungen, auch „Traces“ genannt, können technisch auf unterschiedliche Weise dargestellt werden. In den Anforderungstext lassen sie sich in textueller Form als Verweis einbauen. Die Referenzierung der Verbindung erfolgt über die ID sowie über die Art der entsprechenden Information. Eine Sortierung ist bei dieser Methode jedoch nicht möglich. Wird mit professionellen Tools für das RE gearbeitet, legen diese Programme Hyperlinks zwischen bestimmten Informationen an. Die Verbindungen können auch durch textuelle Verweise in den Attributen der Anforderungen selbst gekennzeichnet werden. [1, vgl. S. 426 ff.]

**Änderungsmanagement**
Betrachtet man den gesamten Entwicklungs- und Lebenszyklus eines Systems, wird ersichtlich, dass sich Anforderungen stetig ändern. Entweder werden sie verändert, entfernt oder es kommen neue Anforderungen hinzu. Ursachen für die Änderung der Anforderungen können Fehler, unvollständige Anforderungen oder Änderungen der Umgebung des Systems sein. Hierzu zählen beispielsweise Gesetzesänderungen, neue Technologien, Änderungswünsche von Seiten der Stakeholder oder Fehlverhalten von Systemen. [4, vgl. S. 148]

Wenn Anforderungen bereits übergeben oder festgelegt wurden, können sie nicht einfach geändert werden, es muss ein Änderungsantrag gestellt werden. Die Änderungsanträge werden vom sogenannten Change Control Board bzw. Change

Management Board überprüft. Es kann sich dabei um eine Gruppe von Personen oder eine Einzelperson handeln. Dieses Gremium nimmt die Anträge von den Auftraggebern entgegen, bearbeitet sie, trifft eine Entscheidung bezüglich der Änderung und teilt sie den Betroffenen mit. Die gesamten Abläufe hierzu müssen unternehmensspezifisch definiert sein. [7, vgl. S. 315 ff.]

Der Änderungsantrag, der auch Change Request genannt wird, dokumentiert die gewünschten Änderungen. Folgende Bestandteile sollte ein solcher Antrag aufweisen:

- Eindeutiger Identifikator
- Titel
- Genaue Beschreibung der Änderung
- Grund der Änderung
- Datum der Antragstellung
- Name des Antragstellers
- Priorität aus Sicht des Antragstellers

[4, vgl. S. 148]

*5) Werkzeugeinsatz für das Requirements-Engineering:*

Zur Erfassung, Dokumentation und Verwaltung von Anforderungen im RE existieren verschiedene Werkzeuge, die diese Tätigkeiten unterstützen. Grande unterscheidet hierbei zwischen Einfach-Werkzeugen und spezialisierten Anforderungswerkzeugen. [10, vgl. S. 26 ff.]

Unter dem Begriff der Einfach-Werkzeuge versteht man Standardanwendungen, dazu gehören beispielsweise die Textverarbeitung oder Tabellenkalkulation. Diese Werkzeuge haben den Vorteil, dass sie bereits im Unternehmen vorhanden sind. Die Anwender besitzen in der Regel die nötigen Kenntnisse um damit zu arbeiten. Dadurch entfallen aufwändige Schulungen der Mitarbeiter und diese erhalten einen schnellen Zugang zur Thematik. Es gibt jedoch auch Nachteile. Handelt es sich um größere Projekte mit einer großen Anzahl von Anforderungen, stoßen die Einfach-Werkzeuge an ihre Grenzen. Gerade die Verwaltung der Anforderungen ist hierbei nicht optimal zu realisieren. Die Dokumentation der Anforderungen kann alternativ auch mit Datenbankanwendungen umgesetzt werden. Diese ermöglichen die Sichtenbildung mit den dazugehörigen Möglichkeiten zum Filtern der Daten. [10, vgl. S. 26 f.]

Neben den Einfach-Werkzeugen gibt es die spezialisierten Anforderungswerkzeuge. Dabei handelt es sich um Anwendungen, die speziell für das RE konzipiert sind. Der Vorteil dieser Programme ist die automatische Vergabe von eindeutigen Identifikationsnummern für die einzelnen Anforderungen, die automatische Versionierung und die Abbildung der Nachverfolgbarkeitsbeziehungen. [10, vgl. S. 28 ff.]

Ebert geht in diesem Zusammenhang näher auf bestimmte Werkzeuggruppen ein. Seine Ausführungen zu Office-Werkzeugen, Wikis, Workflow-Tools und speziellen RE"=Werkzeugen ist für dieses Projekt von besonderem Interesse.

Die Office-Werkzeuge zählen zu den in diesem Kapitel beschriebenen Einfach-Werkzeugen für das RE. In dieser Kategorie findet die Verwendung von eigens erstellten Vorlagen in Tabellenkalkulations oder Textverarbeitungsprogrammen Anwendung. In kleineren Projekten können die Anforderungen in Excel so effizient überwacht werden. Ebert beschreibt den ergänzenden Einsatz von Excel in agilen Projekten mit der anschließenden Möglichkeit zur Übertragung in spezielle RE-Werkzeuge. Jedoch nennt er auch dessen Grenzen bezüglich komplexer Anforderungen wie Bildern, Wiederverwendung, Änderungsmanagement und auch beim verteilten Arbeiten. Diese Probleme versucht man häufig mithilfe eines Textverarbeitungsprogramms zu lösen. Da in diesen Programmen die Verwendung von Prosaanforderungen leicht überhand nimmt, sollte Word auch nur unterstützend seinen Einsatz im RE finden. [9, vgl. S. 279]

Ein sehr verbreitetes Mittel zur Kommunikation ist in kleinen Unternehmen und bei kontinuierlichen Vorgehensweisen die Verwendung von Wikis. Mit diesen Umgebungen können verschiedene Anwender Anforderungen organisieren. Die sinnvolle Nutzung ist jedoch nur möglich, wenn vorher bestimmte Vorschriften zur Form vereinbart wurden. Laut Ebert herrscht bei Wikis in der Praxis häufig Chaos wodurch sie dann für den Einsatz im RE zwecklos sind. [9, vgl. S. 280]

Ist es notwendig Aufgaben zwischen verschiedenen Benutzern zu verteilen oder handelt es sich um ein verteiltes Projekt, ist der Einsatz von Workflow-Lösungen interessant. Die einzelnen Tätigkeiten im RE können somit von der Ermittlung bis hin zur Freigabe automatisch an die betreffenden Personen adressiert werden. Diese erhalten über das Tool die Aufforderung bestimmte Aktivitäten durchzuführen. Ebert stellt in dieser Kategorie einige Tools vor, darunter auch Open-Source Anwendungen. Jedes der Tools hat bestimmte Fähigkeiten zur Lösung von Teilaufgaben des REs. Beispiele hierfür sind Redmine oder Jira. Diese bieten primär die Unterstützung durch Workflows, zeigen jedoch ihre Schwächen bei Dokumentenverwaltung und Versionierung. Jira geht zwar in die Richtung der RE-Werkzeuge, setzt den Schwerpunkt aber auf Workflow und Verwaltung, hierbei fehlt jedoch der Dokumentationsaspekt. [9, vgl. S. 280 f.]

Die Verwaltung, Nachverfolgung und Verknüpfung verschiedenster Dokumente im RE wird von den speziellen RE-Werkzeugen angeboten. Diese Programme liefern gute grafische Umgebungen zur Darstellung von Anforderungen, Abhängigkeiten, Workflows oder Modellen. Komplexe Anforderungen können anhand vielseitiger Filter verknüpft, gruppiert, weiterbearbeitet und verwaltet werden. Die Oberfläche für die Benutzer ist ähnlich wie bei gängigen Office-Werkzeugen. Des Weiteren ist der Import bzw. Export der Informationen mit Word, auf Basis von XML oder speziellen Formaten für Anforderungen wie z.B. Requirements Interchange Format (ReqIF) möglich. [9, vgl. S. 300] Modellierungsmethoden werden je nach Werkzeug unterschiedlich unterstützt, was bei deren Auswahl beachtet werden sollte.

Beispiele für spezielle RE-Werkzeuge sind DOORS, Integrity oder Quality Center. Obwohl der Preis für diese speziellen Programme laut Ebert hoch ist, sind sie im RE weit verbreitet. Ein Grund hierfür sind ihre Stärken in den Bereichen Rechteverwaltung und verteiltes Arbeiten. [9, vgl. S. 281 f.]

## III. IST-ANALYSE

In diesem Kapitel wird die aktuelle Umsetzung des REs im Bereich GE der Zollner Elektronik AG genauer betrachtet. Dazu wird die aktuelle Situation bezüglich des REs im Bereich GE als Untersuchungsobjekt für die Ist-Analyse definiert. Um die aktuelle Umsetzung des REs bewertbar zu machen und später mit weiteren Umsetzungsmöglichkeiten vergleichen zu können wird als Entscheidungsinstrument eine Nutzwertanalyse durchgeführt. Deren Durchführung wird im folgenden näher erläutert.

### *A. Ableitung der Kriterien*

Zur Durchführung der Nutzwertanalyse werden als erstes die zu bewertenden Kriterien definiert. Die Ableitung der Kriterien findet hierbei in mehreren Schritten statt. Im ersten Schritt erfolgt die Sammlung möglicher Kriterien aus den Grundlagen und Tätigkeiten des REs, siehe section II. Zudem werden Interviews mit den Stakeholdern des Bereichs GE zum Thema RE durchgeführt, welche ebenfalls als Quelle zur Bildung der Kriterien dienen. Die Vorauswahl der Kriterien wird gesammelt in Form eines Fragebogens an Mitarbeiter aus dem Bereich GE und den Prozesseigner des REs bei Zollner zur Bewertung weitergegeben. Auf dem Fragebogen ist zu vermerken, welche Kriterien sinnvoll für die Nutzwertanalyse übernommen werden können. Außerdem wird mithilfe des Fragebogens nach weiteren möglichen Beurteilungskriterien gesucht, indem die Befragten zusätzliche Vorschläge angeben können. Auf diese Weise ergeben sich die in Tabelle I enthaltenen Kriteriengruppen und Kriterien zur Beurteilung des REs im Bereich GE bei der Zollner Elektronik AG.

Table I
ABGELEITETE KRITERIEN.

| Kriteriengruppe | Kriterium |
|---|---|
| Allgemein | Vorgehensweise |
| | Verantwortlichkeit |
| Ermitteln | Anforderungsquellen |
| | Einsatz von Techniken |
| Dokumentieren | Form |
| | Inhalt |
| Prüfen | Zeitpunkt |
| | Ablauf |
| Verwalten | Attributierung |
| | Priorisierung |
| | Versionierung |
| | Traceability |
| | Änderungsmanagement |
| | Werkzeugeinsatz |

Die Kriterien sind in Kriteriengruppen mit zugehörigen Unterkriterien eingeteilt. Im Bereich *Allgemein* werden die Kriterien *Vorgehensweise* und *Verantwortlichkeit* bewertet. Unter den Punkt Vorgehensweise fällt die Überprüfung, ob das RE nach einer der vorgestellten Vorgehensweisen durchgeführt wird. Beim Kriterium Verantwortlichkeit erfolgt zum einen die Bewertung der Definition und Abgrenzung der Begriffe Anfrage und Anforderung im Bereich GE. Zum anderen muss die Zuständigkeit des REs festgelegt und eine verantwortliche Person für das RE im Bereich GE benannt und bekannt sein.

Die Kategorie *Ermitteln* enthält die Kriterien *Anforderungsquellen* und *Einsatz von Techniken*. Im Bereich der Anforderungsquellen wird bewertet, ob bei der Ermittlung der Anforderungen alle zur Anforderungsermittlung notwendigen Anforderungsquellen berücksichtigt werden. Des Weiteren fließt die richtige und umfassende Auswahl der internen Stakeholder mit in die Bewertung dieses Kriteriums ein. Das Kriterium Einsatz von Techniken untersucht die Verwendung von gängigen Techniken zur Ermittlung der Anforderungen.

In der Gruppe *Dokumentieren* sind die Kriterien *Form* und *Inhalt* vertreten. Unter dem Punkt Form wird überprüft, ob zur Dokumentation der einzelnen Anforderungen eine einheitliche Struktur festgelegt ist und verwendet wird. Daneben erfolgt eine Bewertung bezüglich der Existenz einheitlicher Vorgaben zur Ablage aller Daten im RE und zu Struktur und Format des Lastenhefts. Durch das Kriterium Inhalt wird die Verwendung einer identischen Wortwahl bei der Anforderungsdokumentation begutachtet.

Der Bereich *Prüfen* enthält die Kriterien *Zeitpunkt* und *Ablauf*. Hierbei bewertet der Zeitpunkt die Existenz von regelmäßigen Quality Gates zum Prüfen der Anforderungen. Der Punkt Ablauf beleuchtet das Vorgehen beim Prüfen der Anforderungen nach der Abarbeitung der drei Qualitätskriterien, der Verwendung von geeigneten Techniken und einer festgelegten Verantwortlichkeit für das Prüfen.

Die letzte Kategorie enthält alle Anforderungen zum Bereich *Verwalten*. Im Einzelnen sind dies die Kriterien *Attributierung*, *Priorisierung*, *Versionierung*, *Traceability*, *Änderungsmanagement* und *Werkzeugeinsatz*. Bei der Attributierung wird überprüft, ob eine Vergabe von Zusatzinformationen zu den einzelnen Anforderungen vorgenommen wird. Der Aspekt Priorisierung untersucht die Bildung einer Rangfolge für die Menge an Anforderungen. Das Kriterium Versionierung bewertet, ob bei Anforderungsänderungen auf die verschiedenen Stände zugegriffen werden kann. Bei der Bewertung der Traceability geht es um die Umsetzung der Nachverfolgbarkeit zwischen den verschiedenen Informationsarten. Unter dem Kriterium Änderungsmanagement wird betrachtet, ob für Änderungen von Anforderungen ein entsprechendes Änderungsmanagement vorhanden ist. Das Kriterium Werkzeugeinsatz beschäftigt sich mit dem Einsatz und der

Ausprägung von computergestützten Hilfsmitteln für das RE.

### *B. Gewichtung der abgeleiteten Kriterien*

Um die Bedeutung der soeben definierten Kriterien festzulegen, werden diese als nächstes priorisiert. Hierbei werden zuerst die Kriteriengruppen, gefolgt von den einzelnen Unterkriterien gewichtet. Um eine möglichst eindeutige Gewichtung zu erreichen wird unter anderem die Methode des paarweisen Vergleichs eingesetzt. Hierbei werden die Kriterien anhand einer Matrix dargestellt. Wenn das Zeilenkriterium „wichtiger“ als das Spaltenkriterium ist, wird die Zahl eins in das Schnittfeld der beiden Kriterien eingetragen. Ist das Zeilenkriterium „weniger wichtig“, wird dem Schnittfeld die Zahl null zugewiesen. Eine Bewertung mit „gleich wichtig“ wird bewusst nicht angeboten, um eine exaktere Gewichtung zu erhalten. Es macht keinen Sinn, ein Kriterium mit sich selbst zu vergleichen, diese Felder müssen nicht gefüllt werden. In einer Summenspalte wird angegeben, wie oft ein Kriterium als wichtig bewertet wurde. Mithilfe dieser Spalte und deren Gesamtsumme erfolgt die Ermittlung der Gewichtung in der letzten Spalte unter Verwendung des Dreisatzes. [12, vgl. S. 14]

Bei den Kriteriengruppen, die jeweils nur zwei Kriterien enthalten wird die stufenweise Gewichtung zur Priorisierung angewendet. Dabei werden jeweils 100 % auf die Kriterien aufgeteilt. Das genaue Vorgehen bei der Kriteriengewichtung wird im Folgenden erläutert.

Die Gewichtung der Kriteriengruppen erfolgt mithilfe des paarweisen Vergleichs, welcher von vier Teilnehmern mit entsprechender Fachkompetenz durchgeführt wird. Diese setzen sich aus Mitarbeitern des Bereichs GE sowie dem Prozesseigner für das RE der Zollner Elektronik AG zusammen. Aus den Punktwerten der einzelnen Vergleiche wird der Mittelwert gebildet und in Prozentwerte umgewandelt.

Um zu den Gewichtungen der einzelnen Unterkriterien zu kommen wird für die Kategorien *Allgemein*, *Ermitteln*, *Dokumentieren* und *Prüfen* die stufenweise Gewichtung vorgenommen. Hierzu sind 100 Punkte pro Kategorie auf die Unterkriterien zu verteilen. Dieser prozentuale Anteil der Kriterien wird in Bezug auf das Gewicht der jeweiligen Kategoriengruppe in einen Punktwert umgerechnet. Bei der Kategoriengruppe Verwalten ergibt sich der prozentuale Anteil der einzelnen Kriterien durch den paarweisen Vergleich, welcher analog zum paarweisen Vergleich für die Kriteriengruppen angewendet wird. Die erhaltenen Ergebnisse werden anschließend auf Grundlage der Gewichtung der Kriteriengruppe *Verwalten* in Punktwerte umgerechnet.

### *C. Bewertung der abgeleiteten Kriterien*

In diesem Abschnitt werden die einzelnen Kriterien nach ihrem Erfüllungsgrad bewertet und bestimmte Punktwerte vergeben. Für die Bewertung der Kriterien wird die in Table II dargestellte Skala festgelegt. Diese enthält die möglichen Punktwerte, die pro Kriterium zu vergeben sind. Außerdem ist angegeben, was ein bestimmter Punktwert über den Erfüllungsgrad eines Kriteriums aussagt. Demnach kann ein Kriterium nicht erfüllt, unzureichend erfüllt, mit Mängeln erfüllt, gut erfüllt oder sehr gut erfüllt sein.

Table II
FESTGELEGTE BEWERTUNGSSKALA.

| Punkte | Kriterium |
|---|---|
| 1 | nicht erfüllt |
| 2 | unzureichend erfüllt |
| 3 | mit Mängeln erfüllt |
| 4 | gut erfüllt |
| 5 | sehr gut erfüllt |

Anhand der Bewertungsskala aus Tabelle II werden die Kriterien in einer Diskussion zwischen Teilnehmern aus dem Bereich GE und dem Prozesseigner des REs der Zollner Elektronik AG bewertet.

### *D. Ergebnisse der IST-Analyse*

Aus den Ergebnissen der Kriteriengewichtung und Kriterienbewertung lässt sich der Nutzwert zur aktuellen Situation des REs im Bereich GE errechnen. Hierzu wird zuerst der Punktwert pro Kriterium ermittelt, indem das Kriteriengewicht mit der zugehörigen Kriterienbewertung multipliziert wird. Den Gesamtnutzwert für das aktuelle RE bildet die Summe aus den Punktwerten der einzelnen Kriterien.

Aus der Nutzwertanalyse ist zu erkennen, dass das RE im GE mit weniger als der Hälfte der erreichbaren Punkte, Potenzial für Verbesserungen bietet. Im nächsten Abschnitt werden die Kriterien ausgewählt, bei denen eine Optimierung die höchste Steigerung des Nutzwerts ermöglicht.

## IV. OPTIMIERUNG

Bei der Auswahl der zu optimierenden Kriterien spielen zwei Faktoren eine Rolle. Es handelt sich hierbei um die Gewichtung und die Bewertung der Kriterien. Ein großes Potenzial zu einer hohen Nutzwertsteigerung bieten Kriterien mit hoher Gewichtung und niedrigem Erfüllungsgrad. Das Filtern der gesamten Kriterien nach diesen beiden Eigenschaften liefert die in Table III enthaltenen Kriterien mit dem höchsten Optimierungspotenzial.

Aus Table III werden die Kriterien *Anforderungsquellen*, *Inhalt*, *Form* und *Werkzeugeinsatz* zur Bearbeitung in der Optimierungsphase ausgewählt. Das Kriterium *Einsatz von Techniken* wird auf Wunsch des Bereichs GE aus dieser Betrachtung ausgeschlossen.

Table III
KRITERIEN MIT HÖCHSTEM OPTIMIERUNGSPOTENZIAL.

| Kriterien |
|---|
| Anforderungsquellen |
| Inhalt |
| Form |
| Einsatz von Techniken |
| Werkzeugeinsatz |

### *A. Beschreibung der Lösungsausprägungen*

Dieser Abschnitt beschreibt die Lösungsfindung zur Nutzwertsteigerung. Diese wird für die im obigen Abschnitt ausgewählten Kriterien, unter Verwendung des morphologischen Kastens durchgeführt. Die Anwendung des morphologischen Kastens gliedert sich in drei Schritte. An erster Stelle stehen die Definition, Analyse und Beschreibung des Problems. Hierbei wird das Problem in seine elementaren Bestandteile aufgeteilt. Im zweiten Schritt werden alle möglichen Lösungsausprägungen pro Element erläutert. Abschließend werden verschiedene Ausprägungen kombiniert und als mögliche Lösungsvarianten aufgezeigt. Hieraus wird eine Lösungsausprägung als bevorzugte Variante bestimmt. [11, vgl. S. 172 f.]

Auf dieses Projekt übertragen, stellt der aus der Nutzwertanalyse hervorgehende niedrige Erfüllungsgrad des REs, das Problem dar. Die einzelnen Bestandteile des Problems stellen die zur Optimierung festgelegten Kriterien dar. Abbildung 3 zeigt die Visualisierung des morphologischen Kastens. Die erste Spalte enthält das in die grundlegenden Elemente zerlegte Gesamtproblem. Die Spalten zwei bis vier umfassen die Lösungsausprägungen. Jede Zeile beinhaltet mögliche Lösungen für ein bestimmtes Element.

| | Ausprägungen | | |
|---|---|---|---|
| **Anforderungsquellen** | Checkliste | Formular | |
| **Inhalt** | Wortwahl-Liste | Listenfeld | Textanalyse-Werkzeug |
| **Format** | Leitfaden | Template | |
| **Werkzeugeinsatz** | Einfach-Werkzeuge | JIRA | DOORS |

Figure 3. Morphologischer Kasten zur Lösungsfindung.

**Checkliste**
Die Tätigkeit der Anforderungsermittlung kann durch den Einsatz einer Checkliste transparenter gestaltet werden. Den Hauptinhalt dieser Checkliste stellen die zu berücksichtigenden Anforderungsquellen dar. Die Checkliste dient somit als Wegweiser für den Anforderungsermittler. Es wird sichergestellt, dass die wichtigen Anforderungsquellen *Dokumente*, *Systeme im Betrieb* und *Stakeholder* bedacht werden. Zusätzlich hierzu sollte in der Checkliste angegeben werden, welche internen Stakeholder bei den verschiedenen Projektarten zu konsultieren sind. Analog hierzu werden die bekannten Standards oder vergleichbaren Projekte zu einem aktuellen Projekt dokumentiert. Voraussetzung für eine erfolgreiche Anwendung ist dabei, eine beständige Aktualisierung der Checkliste durch einen ausgewählten Verantwortlichen.

**Formular**
Ein weiterer Lösungsansatz zur Optimierung des Kriteriums Anforderungsquellen, ist die Anfertigung eines Formulars zur Erfassung von Anforderungen. Diese Idee wird durch das Ergebnis der Kundenbefragungen mit den internen Stakeholdern bestärkt. Demnach benötigen die verschiedenen Stakeholder je nach Projektart gewisse Informationen zur Bearbeitung des Projekts. Diese Informationen sollen vom RE ermittelt, als Anforderung erfasst und an die Stakeholder weitergeleitet werden. Deshalb wird vorgeschlagen ein Formular zu entwerfen, das mit allen notwendigen Informationen gefüllt werden kann. Da es im Bereich GE viele verschiedene Projektarten und mehrere interne Stakeholder gibt, ist es erforderlich, mehrere Formulare zu erstellen. Somit ist es möglich, den Stakeholdern die Anforderungen in der gewünschten Form zu übergeben.

**Wortwahl-Liste**
Zum Kriterium Inhalt ist bisher noch keinerlei Erfüllung vorhanden. Aus diesem Grund besteht die erste Lösungsvariante darin, die Erkenntnisse aus den Tätigkeiten des RE zunächst grundsätzlich auszuarbeiten und zum Einsatz verfügbar zu machen. Dazu wird die Anfertigung von Tabellen vorgeschlagen. Auf diese Weise können die Prozesswortliste, das Glossar und die Definition der rechtlichen Verbindlichkeit festgelegt und dokumentiert werden. Diese Vorgaben sind anschließend allen beteiligten Mitarbeitern zu kommunizieren und müssen diesen auch verfügbar gemacht werden.

Die Prozesswortliste enthält alle Vollverben, die in den Anforderungen verwendet werden. Durch das Führen einer Prozesswortliste wird sichergestellt, dass eine Anforderung von jedem Betrachter genau gleich verstanden wird. Um die Prozesswortliste anzufertigen, müssen alle vorhandenen Unterlagen des jeweiligen Projekts nach den Prozesswörtern durchsucht werden. Mit diesen Informationen kann anschließend eine Prozesswortliste erstellt werden. Diese Liste wird im nächsten Schritt zu einer Tabelle vervollständigt. Es gibt eine Spalte für das Prozesswort selbst, eine Spalte für die Bedeutung des Prozessworts und eine Spalte für gleichbedeutende Wörter. [1, vgl. S. 227 f.] Analog zur Vorgehensweise bei der Erstellung der Prozesswortliste, muss ein Glossar angelegt werden. Auch die Schlüsselwörter zur rechtlichen Verbindlichkeit und deren Bedeutung werden in einer Tabelle festgehalten.

**Listenfeld**
Ein weiterer Lösungsvorschlag, um das Kriterium Inhalt zu erfüllen, ist der Einsatz von Listenfeldern. Dazu müssen die Prozesswörter, Substantive und rechtliche Verbindlichkeiten erarbeitet werden. In einer Standardanwendung wie z.B. Mi-

crosoft Word kann die Struktur der Anforderung vorgegeben werden. Im jeweiligen Feld werden die erarbeiteten Begriffe hinterlegt. Durch Verwendung eines Listenfelds können ausschließlich die gepflegten Einträge ausgewählt werden. Somit ist die Verwendung einer einheitlichen Wortwahl vorgegeben. Der so entstandene Anforderungssatz kann dann per Copy-and-Paste schnell und einfach an die gewünschte Stelle übertragen werden.

**Textanalyse-Werkzeug**
Zur Verbesserung der inhaltlichen Bestandteile einer Anforderung können, als weitere Steigerung, auch unterstützende Programme eingesetzt werden. Diese Anwendungen tragen dazu bei, die Qualität bei der Formulierung von natürlichsprachlichen Anforderungen zu gewährleisten.

Ein Beispiel hierfür ist das Produkt DESIRe® der HOOD GmbH. Dieses Werkzeug analysiert die in einer Anforderung verwendeten Wörter. Werden auffällige Inhalte entdeckt, stellt das Programm vordefinierte Fragen und weist auf die vorhandenen Schwächen in der Anforderung hin. Durch die Beantwortung dieser Fragen werden die Anforderungen inhaltlich überarbeitet. Außerdem ist eine Anpassung des Programms an die Bedürfnisse des Unternehmens möglich. So können beispielsweise eigene Wörter als zu erkennende Schwachstellen definiert werden. [13] Eine Verwendung dieses Werkzeugs ist für Einfach-Werkzeuge und spezielle RE-Tools möglich. Es kann sowohl mit Microsoft Word als auch mit DOORS genutzt werden. [14] Das Tool steht nach Angaben des Anbieters auf der eigenen Homepage zum kostenlosen Download zur Verfügung. [13]

**Leitfaden**
Eine mögliche Lösung besteht darin, die Anforderungen unter Berücksichtigung der vorgestellten Anforderungsschablonen zu formulieren. Dazu sollten speziell für den Bereich GE Anforderungsschablonen für funktionale Anforderungen, Qualitätsanforderungen und Umgebungsanforderungen eingesetzt werden. Rupp [1, vgl. S. 219] stellt hierzu bereits entsprechende Vorlagen zur Struktur der Anforderungssätze für verschiedene Anforderungsarten zur Verfügung.

Im Bereich GE muss die Benutzung dieser drei Anforderungsschablonen in einer Anweisung schrittweise erklärt und dokumentiert werden. Diese Dokumentation dient zukünftig jedem Mitarbeiter als Leitfaden zur Formulierung seiner Anforderungen. Dadurch wird ein einheitlicher Aufbau der Anforderungen erreicht, der zudem unabhängig vom Verfasser ist. Außerdem werden durch die Vorgabe dieser Strukturen einige sprachliche Effekte von Vornherein ausgeschlossen. Passivsätze beispielsweise können, aufgrund des vorgegebenen Satzbaus, nicht entstehen. [1, vgl. S. 217]

**Template**
Eine weitere Variante zur Verbesserung der Form ist das Bereitstellen von Templates zur Dokumentation der Anforderungen. Grundlage für diesen Ansatz bilden die Anforderungsschablonen nach Rupp. [1, vgl. S. 219] Die Struktur der Anforderungsschablonen kann mithilfe von Standardanwendungen wie z.B. Microsoft Word vorgegeben werden. Hierbei wird jeder Teil der Schablone durch ein auszufüllendes Feld repräsentiert. Die Felder sind fest vorgegeben und können in ihrer Reihenfolge nicht verändert werden. Somit ist die Grundlage für eine einheitliche Struktur aller Anforderungen geschaffen.

**Einfach-Werkzeug**
Die Standardanwendungen in heutigen Unternehmen eignen sich bis zu einem gewissen Grad auch für den Einsatz im RE. Für den Bereich GE wird vorgeschlagen, die vorhandenen Programme Microsoft Word und Microsoft Excel für die Verwaltung der Anforderungen zu nutzen. Aktuell wird Microsoft PowerPoint zusammen mit dem unternehmenseigenen Wiki verwendet. Diese Programme unterstützen das RE gar nicht oder nur sehr wenig. Durch den Einsatz von Microsoft Word können Verbesserungen im Bereich der Anforderungsstruktur und des Anforderungsinhalts erzielt werden. Dazu werden das Template, sowie das Listenfeld zusammen implementiert. Des Weiteren sollte mit Microsoft Word eine Formatvorlage für das Lastenheft aufgebaut werden, welche für alle Projekte wiederverwendet wird. Die Verwaltung der Anforderungen wird anhand von Microsoft Excel realisiert, wodurch die Attributierung und Priorisierung angewendet werden kann. Eine automatische Versionierung oder Traceability wird hierbei nicht angeboten. Für das Änderungsmanagement besteht ebenfalls keine weitere Unterstützung.

**JIRA**
Die Software JIRA ist ein Produkt, das von der Firma Atlassian angeboten wird. Nach eigenen Angaben handelt es sich dabei primär um eine Anwendung für die Bereiche Problembehandlung und Projektmanagement. Aufgrund der verfügbaren Funktionen, ist die Software auch für das RE verwendbar, da diese an den Umgang mit Anforderungen anpassbar sind. Für die Anforderungen können benutzerdefinierte Felder angelegt und ein Berichtswesen genutzt werden. Ein weiterer Vorteil von JIRA liegt in der vorhandenen Workflow-Funktionalität. Die Anforderungen können durch Teilaufgaben verwaltet und verschiedenen Benutzern zugewiesen werden. Sowohl in Beziehung stehende Anforderungen, als auch Anforderungen und die zugehörige Funktionsanfrage, können verbunden werden. Zusätzlich ist es möglich JIRA mit dem Unternehmenswiki Confluence zu integrieren, da dieses ebenfalls von Atlassian angeboten wird. Dadurch werden Verlinkungen zwischen den beiden Anwendungen ermöglicht. [15]

Auf der Website von Atlassian wird weiterhin auf die Möglichkeit hingewiesen, zusätzliche RM-Funktionen durch die Integration anderer Tools zu erhalten. Diese werden als Plug-ins kostenpflichtig von Drittanbietern angeboten. [15] Dadurch können Aufgaben wie Traceability, Änderungsmanagement und Wiederverwendung zwischen Projekten sowie Sichtenbildung von JIRA umgesetzt werden. [16] Diese Informationen dienen der Vollständigkeit, bei der späteren Be-

wertung wird die Standard JIRA-Version betrachtet.

**Rational DOORS**
Das Programm Rational DOORS von IBM ist eine spezielle Anwendung für den Bereich RE. [17, vgl. S. 3] Die Anforderungen werden mithilfe von Sichten, Links und Rückverfolgbarkeitsanalysen über den gesamten Projektlebenszyklus überwacht und verwaltet. [17, vgl. S. 4]

Zur Erfassung der Anforderungen ist eine Schnittstelle zu Textverarbeitungsprogrammen vorhanden. Es werden verschiedene Dateiformate für den Import unterstützt, darunter befinden sich auch Microsoft-Anwendungen. [17, vgl. S. 3] Die Vergabe von Attributen kann in DOORS benutzerdefiniert umgesetzt werden. [17, vgl. S. 5] Auch die Sichtenbildung wird von DOORS unterstützt. [17, vgl. S. 7 f.] Des Weiteren werden alle Projekte durch Anlegen einer geeigneten Ordnerstruktur verwaltet. [17, vgl. S. 8] Ein weiterer Vorteil ist die Rückverfolgbarkeit. So können Informationen miteinander verknüpft werden. Dies betrifft zum einen die Anforderungen untereinander, aber auch die Anforderungen und zugehörige Dokumente. Durch generierte Links stehen dem Benutzer alle in Beziehung stehenden Informationen zur Verfügung. [17, vgl. S. 6] Anhand von Baselines können zudem schreibgeschützte Versionen eines bestimmten Zustands festgehalten werden. [17, vgl. S. 10] Eine integrierte Änderungsüberwachung macht den Verlauf von Änderungen im System nachvollziehbar. Durch sichtbare Änderungsmarkierungen ist zu erkennen, welcher Benutzer eine Information zu einem bestimmten Zeitpunkt verändert hat. [17, vgl. S. 9] Hervorzuheben ist, das vorhandene Änderungsvorschlagsystem, welches die Abwicklung und Dokumentation des Änderungsmanagements durch DOORS ermöglicht. [17, vgl. S. 12] Für das Prüfen und Genehmigen der Anforderungen bietet DOORS ein hilfreiches Feature. Die Freigabe oder Prüfung von Anforderungen erfolgt mithilfe von elektronischen Signaturen. Dadurch wird bei einer Freigabe der Prüfer, eine Zeitmarke und die geprüfte Information festgehalten. [18, vgl. S. 99 f.]

### *B. Aufstellung und Bewertung von Lösungsvarianten*

Indem die einzelnen Lösungsausprägungen aus Abbildung 3 miteinander kombiniert werden, entstehen verschiedene Lösungsvarianten zur Steigerung des Nutzwertes. Auf diese Weise werden verschiedene Lösungsvarianten aufgestellt, diese sind im morphologischen Kasten in Abbildung 4 eingezeichnet.

Um die verschiedenen Varianten vergleichen zu können, werden sie anhand einer Nutzwertanalyse bewertet. Die Grundlage hierzu bildet die Beschreibung der einzelnen Varianten. Die Bewertung erfolgt analog zur Bewertung der Ist-Situation. Für jede Variante werden alle Kriterien erneut betrachtet und mit Punktwerten belegt. Daraus ergeben sich gesteigerte Nutzwerte. Hierbei ergibt sich für Lösung eins eine Nutzwertsteigerung gegenüber der Ist-Analyse von 65 %. Bei Lösungsvariante zwei handelt es sich um einen Anstieg

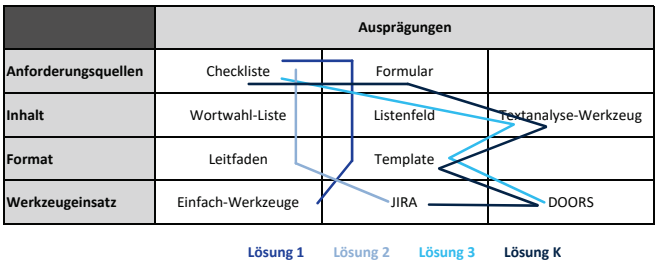



Figure 4. Mögliche Lösungsvarianten.

des Nutzwertes um 56 %. Mit der dritten Lösung kann ein um 80 % höherer Nutzwert erreicht werden. Durch erneute Kombination der Lösungen eins und drei zu Lösung K wird die größte Nutzwertsteigerung über 95 % im Vergleich zur Ausgangssituation erzielt.

### *C. Auswahl einer Lösungsvariante*

Um das Ziel der Nutzwertsteigerung zu erreichen, gilt es eine der vorgestellten Lösungsvarianten zur Optimierung des RE auszuwählen. Die Auswahl erfolgt hierbei unter Berücksichtigung der einzelnen Nutzwerte. Zudem wird der Arbeitsaufwand zur Umsetzung der verschiedenen Lösungen abgeschätzt. Außerdem sind die Lizenzkosten pro Lösungsvariante bei der Auswahl miteinzubeziehen.

Betrachtet man die Nutzwerte, liefern Lösung drei und Lösungsvariante K mit einem 80 und 95 % höheren Nutzwert als die Ausgangssituation, das höchste Potenzial zur angestrebten Nutzwertsteigerung. Der Arbeitsaufwand zur Umsetzung der verschiedenen Lösungsvarianten wurde von den Verantwortlichen des Bereiches GE für alle Lösungen als gleichwertig eingestuft. Somit kann anhand des Aufwands zur Umsetzung keine Entscheidung für eine Lösung getroffen werden. Bezüglich der Lizenzkosten können Lösung eins und zwei besonders niedrige Werte vorweisen. Dies muss jedoch vor dem Hintergrund betrachtet werden, dass die in diesen Lösungen eingesetzte Software nicht die volle Funktionalität zum Verwalten von Anforderungen bereitstellt. Für Lösungsvariante K, welche eine Kombination aus JIRA und DOORS verwendet, liegen die Lizenzkosten im Mittelfeld. Hierbei werden eine Lizenz für mehrere Benutzer bei JIRA und zwei Lizenzen für DOORS benötigt. Die Lizenzkosten für Lösungsvariante drei, welche ausschließlich die Software DOORS beinhaltet, sind für fünf Lizenzen unter allen Lösungsvarianten deutlich am höchsten.

Hinsichtlich all dieser genannten Faktoren wird Lösung K zur Optimierung ausgewählt, da es für die Bedürfnisse des Bereiches GE die beste Lösung darstellt. Zudem wird mit Lösung K die höchste Nutzwertsteigerung erreicht. Außerdem sind die Lizenzkosten, die sich im Mittelfeld befinden, vertretbar, da alle gewünschten Funktionen zur Verwaltung von Anforderungen genutzt werden können.

### D. Steigerung der Erfüllungsgrade durch Lösungsvariante K

Bei der Umsetzung der ausgewählten Lösungsvariante K ergeben sich deutliche Verbesserungen im Vergleich zur Ausgangssituation. Von den insgesamt vierzehn Kriterien kann bei 10 Kriterien eine Erhöhung des Erfüllungsgrads erzielt werden. Die Bewertung der einzelnen Kriterien kann jeweils um einen bis vier Punkte erhöht werden. Im Durchschnitt kann somit jedes der Kriterien eine um 2 Punkte verbesserte Erfüllung vorweisen. In Table IV ist detailliert dargestellt, wie der Erfüllungsgrad bei welcher Anzahl an Kriterien positiv verändert wird. Bei drei Kriterien kann beispielsweise eine Steigerung des Erfüllungsgrades um 3 Punkte erzielt werden.

Table IV
ERHÖHUNG DER ERFÜLLUNGSGRADE DURCH LÖSUNGSVARIANTE K.

| Steigerung | Anzahl der Kriterien |
|---|---|
| + 4 | 1 |
| + 3 | 3 |
| + 2 | 2 |
| + 1 | 4 |

Zudem erreicht in der Ausgangssituation keines der Kriterien den vollen Erfüllungsgrad. Wird Lösungsvariante K eingeführt, können vier der 14 Kriterien eine sehr gute Erfüllung vorweisen.

## V. FAZIT

Die Zielsetzung dieses Projekts besteht darin, ein Konzept für das RE im Bereich GE auszuarbeiten. Zusätzlich muss die Umsetzung der vorgeschlagenen Handlungsschritte den Nutzwert des REs steigern. Aufgrund der vorgestellten Vorgehensweisen im Bereich Requirements-Engineering wird die aktuelle Situation des REs im Bereich GE analysiert und bewertet. Dies führt zur Feststellung, dass bei der Mehrheit der Tätigkeiten dringender Handlungsbedarf besteht. Im Zuge einer Optimierungsphase werden deshalb die wichtigsten Kriterien zur Verbesserung ausgewählt. Um die aktuelle Umsetzung des REs zu optimieren, werden verschiedenen Lösungsvarianten vorgeschlagen und bewertet. Durch Kombination der aufgestellten Lösungen ist es möglich, eine zusätzliche Lösungsvariante zu bilden. Da diese das bestmögliche Kosten-Nutzen-Verhältnis vorweist, wird diese Variante zur Optimierung des REs im Bereich GE ausgewählt. Durch diese Lösung, kann der aktuelle Nutzwert um 95 Prozent erhöht werden. Das Ziel der Nutzwertsteigerung ist somit erreicht.

Des Weiteren gilt es, im Rahmen dieses Projekts die Anwendung von Checklisten bzw. Templates zu untersuchen. Diese Möglichkeit wird bei den Aufgaben des REs für die Tätigkeit des Prüfens von Anforderungen aufgezeigt. Zudem enthält die ausgewählte Lösungsvariante den Einsatz von Checklisten und Templates zur Ermittlung sowie zur Dokumentation von Anforderungen.

Eine weitere Vorgabe ist die Auswahl einer geeigneten Software zur Dokumentation und Verwaltung der Anforderungen. Hierbei befinden sich drei Anwendungen unter den möglichen Lösungen. Daraus werden die zwei Werkzeuge für die ausgewählte Lösungsvariante vorgeschlagen.

Abschließend ist festzustellen, dass durch diese Arbeit ein erster Schritt in Richtung eines optimalen REs gemacht wird. Da nicht alle Kriterien einer Optimierung unterzogen werden und sich jederzeit Änderungen ergeben können, ist eine stetige Verbesserung des REs anzuraten. Hierbei empfiehlt es sich, das Thema RE aktiv weiterzuverfolgen und kontinuierlich an den neuesten Erkenntnissen aus der Forschung zu messen.

## REFERENCES


[1] Chris Rupp. 2014. *Requirements-Engineering und -Management: Aus der Praxis von klassisch bis agil*. 6. Auflage. München: Hanser.

[2] IEEE Standards Board. 1990. *IEEE standard glossary of software engineering terminology*. New York: Institute of Electrical and Electronics Engineers.

[3] IEEE Standards Board. 2011. *Software & Systems Engineering Standards Committee of the IEEE Computer Society*.

[4] Klaus Pohl und Chris Rupp. 2011. *Basiswissen Requirements Engineering*. 3. Auflage. Heidelberg: dpunkt-Verlag.

[5] International Requirements Engineering Board. 2017. *IREB in Kurzform - IREB - IREB*. https://www.ireb.org/de/about/basics/.

[6] Klaus Pohl. 2008. *Requirements engineering: Grundlagen, Prinzipien, Techniken*. 2. Auflage. Heidelberg: dpunkt-Verlag.

[7] Peter Hruschka. 2014. *Business Analysis und Requirements Engineering: Produkte und Prozesse nachhaltig verbessern*. München: Hanser.

[8] Thomas Niebisch. 2013. *Anforderungsmanagement in sieben Tagen: Der Weg vom Wunsch zur Konzeption*. Berlin: Springer Gabler.

[9] Christof Ebert. 2014. *Systematisches Requirements Engineering: Anforderungen ermitteln, spezifizieren, analysieren und verwalten*. Heidelberg: dpunkt-Verlag.

[10] Marcus Grande. 2014. *100 Minuten für Anforderungsmanagement: Kompaktes Wissen nicht nur für Projektleiter und Entwickler*. 2. Auflage. Wiesbaden: Springer Vieweg.

[11] Christian Schawel und Fabian Billing. 2014. *Top 100 Management Tools: Das wichtigste Buch eines Managers*. 5. Auflage. Wiesbaden: Gabler Verlag.

[12] Jörg B. Kühnapfel. 2014. *Nutzwertanalysen in Marketing und Vertrieb*. Wiesbaden: Springer Fachmedien Wiesbaden.

[13] Hood GmbH. 2016. *DESIRe®*. https://www.hood-group.com/requirements/beratung/vorgehensentwicklung/desire/.

[14] Hood GmbH. 2016. *DESIRe®*. https://www.hood-group.com/requirements/downloadbereich/#c1204.

[15] Atlassian. 2017. *Using JIRA for Requirements Management - Atlassian Documentation*. https://confluence.atlassian.com/jirakb/using-jira-for-requirements-management-193300521.html.

[16] Atlassian. 2017. *Atlassian Marketplace*. https://marketplace.atlassian.com/plugins/com.easesolutions.jira.plugins.requirements/server/overview.

[17] IBM. *Erste Schritte mit Rational DOORS*.

[18] IBM. *Rational DOORS verwalten*.

# Konzeptionierung und Entwicklung einer administrativen Applikation für die Editierung und Erweiterung von Softwareversprachlichung.

Professor Dr. Frank Herrmann
Innovationszentrum für Produktionslogistik und Fabrikplanung
Ostbayerische Technische Hochschule Regensburg
Galgenbergstraße 32
93053 Regensburg
Frank.Herrmann@OTH-Regensburg.de

Marcel Roth
Projekt29 GmbH & Co. KG
Ostengasse 14
93047 Regensburg
marcel.roth@projekt29.de

## ABSTRACT
Dieser Artikel behandelt die Bearbeitung und Umsetzung eines Projekts mit dem obigen Titel. Dabei liegt das Augenmerk weniger auf der technischen Implementierung als viel mehr auf dem Umfang und dem Management des Projekts selbst.

## Keywords
Softwareversprachlichung, Projektmanagement, Anforderungsmanagement, Softwaredesign, Dokumentation

## 1. KERNZIELE DES PROJEKTS
Die Datenschutzmanagementsoftware, kurz DSM-Software, der *Projekt29 GmbH & Co. KG*, muss aufgrund eines wachsenden Kundenstammes immer mehr Sprachen anbieten. Um das bestehende System um weitere Sprachen zu erweitern, gibt es bereits einen Weg, welcher allerdings für jede Änderung manuell von einem Administrator ausgeführt werden muss und nicht komplett Nutzergebunden ist. Außerdem ist es schwer einen Überblick über die bereits vorhandenen Sprachen und deren Umsetzung zu bekommen. Deshalb soll eine Anwendung (genauer ein Modul für den Adminbereich der DSM-Software) entwickelt werden, die die folgenden Kernziele umsetzt.

- einspielen von Sprachen durch CSV- oder JSON-Dateien
- erweitern von Sprachen durch CSV- oder JSON-Dateien
- darstellen einer detaillierten Übersicht über alle im System verfügbaren Sprachen
- editieren einzelner Sprachwerte der Software (z.B. kundenspezifische Bezeichnungen oder Berichtigung / Erneuerung einzelner Werte)
- exportieren von Sprachen oder Sprachwerten als CSV- oder JSON-Dateien

## 2. UMFANG DES PROJEKTS
Dieses Projekt dient der Entwicklung des im Vorangegangenen beschriebenen Moduls des Administrationsbereichs der DSM-Software. Dabei geht es um ein komplettes IT-Projekt. Dazu gehört die Wahl eines geeigneten Vorgehensmodells, die Erfassung genauer Anforderungen, die Planung der Schnittstellen, die Erstellung eines Designkonzepts, die Implementierung, sowie eine Dokumentation und Analyse des Endergebnisses. Zudem ist ein neuer Administrationsbereich geplant, der im Zuge dieses Projekts initial erstellt werden soll. Dieser wird dann mit seinem ersten Modul (für Versprachlichung) ausgestattet. Im Folgenden wird das Vorgehen für jeden dieser Tätigkeitsbereiche kurz beschrieben. Außerdem wird das Resultat jedes Abschnitts aufgezeigt und abschließend zu einem Endergebnis zusammengeführt.

## 3. AUSWAHL DES VORGEHENSMODELLS
Ein geeignetes Vorgehensmodell muss zur Projektstruktur passen, darf nicht unnötig kompliziert sein und soll einen klaren Projektablauf definieren. Um ein passendes Vorgehensmodell zu finden, müssen zunächst verschiedene Vorgehensmodelle angesehen und im Bezug auf das Projekt evaluiert werden. Dazu wurden in diesem Projekt die nachfolgend aufgelisteten Vorgehensmodelle genauer begutachtet.

- Basismodell
- Wasserfallmodell
- V-Modell
- Spiralmodell
- Extreme Programming
- Scrum

Vor allem die agilen Methoden (Extreme Programming und Scrum), sowie das Spiralmodell sind zu umfangreich für dies Projekt. Außerdem besteht keine große Notwendigkeit für regelmäßige Meetings, da es sich um ein Ein-Mann-Projekt handelt. Somit bleiben nur noch die drei erstgenannten Modelle. Das V-Modell ist ein gutes Vorgehensmodell, welches das Testen in den Vordergrund stellt. Allerdings würden automatisierte Tests den zeitlichen Rahmen für das Projekt sprengen, weswegen die Implementierung im Fokus steht. Letztendlich wurde das Basismodell verwendet und nach den Wünschen des Projektbetreuers etwas abgeändert.

Abbildung 1 zeigt das angewendete Vorgehensmodell, welches vier Phasen hat, die nacheinander durchgearbeitet werden müssen. Dabei ist die Reihenfolge entscheidend, da die Ergebnisse einer Phase als Grundlage für die nächste Phase dienen. Der Ablauf und die Ergebnisse der einzelnen Phasen werden in den nächsten Abschnitten beschrieben.

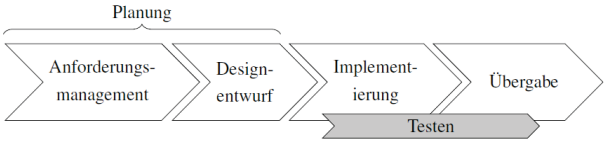



**Abbildung 1: Vorgehensmodell dieses Projekts**

# 4. ERFASSUNG DER ANFORDERUNGEN

Die erste Phase beschäftigt sich mit den Anforderungen an die Software (also das Endergebnis des Projekts). Hier geht es darum, das Ziel für das Projekt möglichst genau zu definieren. Dafür gab es einen iterativen Prozess (siehe Abbildung 2), bei dem ich als Anforderungsmanager fungierte. Dabei ging es darum möglichst detailliert die Funktionalitäten und Schnittstellen der Software festzulegen und regelmäßig mit dem Projektbetreuer abzustimmen und zu validieren. Um eine allgemein gültige Fassung zu schaffen und

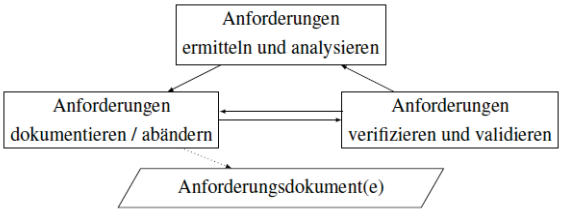



**Abbildung 2: Zyklus des Anforderungsmanagements**

spätere Streitigkeiten zu vermeiden ist es wichtig die Anforderungen korrekt zu dokumentieren. Auch dazu wurde Recherche betrieben. Aber anstatt eines speziellen Anforderungsmanagementtools wurde sich auf eine einfache Tabellendarstellung geeinigt, welche z.B. in Excel gepflegt werden kann. Der Aufbau dieses Anforderungsdokuments wird in Tabelle 1 beschrieben.

| **Attribut** | **Beschreibung** |
|---|---|
| ID | Eindeutiger Wert zur Identifikation der Anforderung |
| Bezeichnung | Sprechender Name für die Anforderung |
| Beschreibung | Kurze aber exakte Beschreibung der Anforderung |
| Quelle | Ursprung der Anforderung (Personen oder Interessensgruppen) |
| Typ | Typ der Anforderung |
| Status | Umsetzungsfortschritt der Anforderung |
| Version | Versionierung der Anforderung |
| Letzte Änderung | Datum der letzten Änderung |

**Table 1: Vorlage für den Informationsumfang einer einzelnen Anforderung**

# 5. ERSTELLUNG DES DESIGNKONZEPTS

Die zweite Phase dient dazu, die vorher festgelegten Anforderungen in ein Designkonzept für die Software umzuwandeln. Dabei geht es nicht darum das Frontend vollständig zu designen. Im Mittelpunkt steht ein Konzept für die Benutzeroberfläche der Anwendung. Dazu zählen das Farbschema, die Formatierung von bestimmten Elementen, sowie die Navigation. Beispielsweise geht es darum welche Elemente in den Kopfbereich kommen und ob Daten besser als Checkboxen oder Dropdown-Menüs dargestellt werden. Für die Umsetzung des Designkonzepts wurde der Webservice *balsamiq* verwendet. Dieser ermöglicht das Erstellen von UI-Scribbles. Außerdem können diese Zeichnungen gleich online gespeichert und anderen für einen Review bereit gestellt werden. Dieses Tool ist also perfekt geeignet für einen iterativen Ansatz, bei dem regelmäßig Rücksprache gehalten wird. Für dieses Projekt wurde wie in Abbildung 3 gezeigt vorgegangen. Die wichtigsten Ergebnisse aus dem Designkonzept

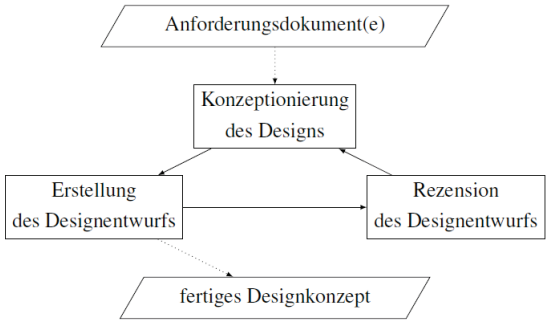



**Abbildung 3: Zyklus des iterativen Designprozesses**

waren die Aufteilung der Anwendung in Kernbereiche, sowie einzelne Komponenten (z.B. Warnungs-Popups und Navigationselemente). Die vier Kernbereiche, die nachfolgend aufgelistet sind, beinhalten zusammen alle nötigen Funktionalitäten und sollen visuell voneinander getrennt sein.

- Sprachen (Übersicht)
- Sprachwerte (Editor und Übersicht)
- alle Imports
- alle Exports

# 6. IMPLEMENTIERUNG DES SERVERS

Bevor der Server entwickelt werden kann muss zunächst die Datenbank designed und erstellt werden. Das entfällt bei diesem Projekt jedoch, da das Datenbankschema dafür bereits existiert und in die Datenbank der DSM-Software integriert ist. Abbildung 4 zeigt die Tabellen dieses sogenannten LValue-Schemas (steht für "*Language Values*"). Die Unternehmen, die bei *Projekt 29* Kunde sind, werden in der Datenbank als "*Tenants*" also Mandanten verwaltet. Es gibt somit zwei verschieden Sprachumgebungen. Zum einen die **GEN**-Umgebung, die die Standardsprachen der Software abbildet, die für alle Benutzer gleichermaßen zur Verfügung stehen. Zum anderen die **TID**-Umgebung, welche kundenspezifische Anpassungen oder Erweiterungen der Standardsprachen, sowie komplett andere Sprachen bereitstellen kann. Um die Software zu versprachlichen geht das LValue-Konzept wie folgt vor. Die **LValue**-Tabellen beinhalten die textuellen Werte für alle Sprachen und verbinden diese via IDs mit der zugehörigen Sprache (**Lang**-Tabellen) und dem repräsentativen Symbol (**Symbol**-Tabelle). In der DSM-Software an sich werden somit keine tatsächlichen Sprachwerte mehr ver-

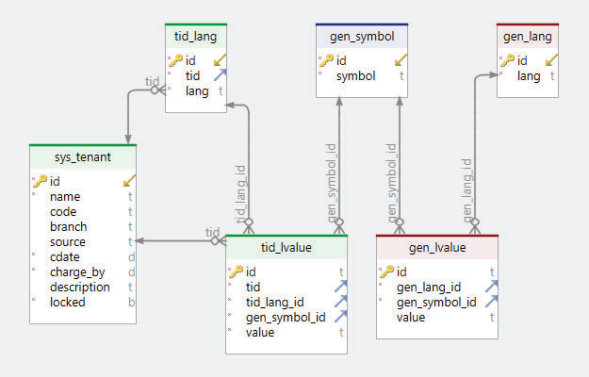



**Abbildung 4: Datenbankschema des LValue-Konzepts**

wendet sondern lediglich Symbole. Bevor dann ein Template oder eine Webseite an einen Nutzer ausgeliefert wird, werden die darin angegeben Symbole mit der passenden Sprache und Tenant-ID aufgelöst und durch den angeforderten Sprachwert substituiert.

## Severstruktur und Frameworks

Der letzte Schritt bevor die Implementierung starten kann, ist das Aufsetzen des Projekts. Dabei ging es darum zu entscheiden welcher Aufbau und welche Frameworks verwendet werden sollen. Auf Details zur Entscheidungsfindung und den einzelnen Frameworks wird in der Zusammenfassung verzichtet. Zum programmieren wird serverseitig Java in der Version 17.0.5 benutzt. Als Build-Management-Tool wird ein *Gradle* der Version 7.4.1 angewendet. Die größte Änderung ist, dass für den Administrationsbereich nicht mehr *Spring* (wird in der DSM-Software genutzt), sondern *Quarkus* als Fullstack-Java-Framework eingesetzt wird. Ein weiteres wichtiges Framework, welches aus der DSM-Software übernommen wurde, ist *jOOQ*. Dieses Framework ist für die Generierung von Java-Klassen in Bezug auf die Datenbank. Es ermöglicht einen einfachen und übersichtlichen Zugriff vom Server auf die Datenbank unter Verwendung einer eigenen *Domain Specific Language* in Java.

## Serverschnittstellen

Tabelle 2 zeigt eine Übersicht über alle, in den Anforderungen dokumentierten, Schnittstellen und ihre technische Spezifizierung. Hier kann auch die Trennung der vier Kernbereiche (siehe Punkt 5) erkannt werden. Allerdings bedienen die Schnittstellen für Symbole und LValues zusammen den Bereich für Sprachwerte.

Zuerst zu den Sprachfunktionalitäten. Der Endpunkt **/lang/cards** liefert eine Liste von Datenobjekten zur Darstellung der Übersicht einer Sprache. Mit **/lang/ls/copy** werden alle Sprachen aufgelistet, die für eine Kopieroperation zwischen den Sprachumgebungen bereitstehen. Im Anschluss kann mit **/lang/do/copy** eine gewünschte Sprache inklusive aller Sprachwerte auf eine andere Sprachumgebung kopiert werden. Der Endpunkt **/lang/tenants** gibt eine Liste von Mandanten zurück, welche danach gefiltert werden kann, ob dieser Mandant über eigene Sprachen verfügt oder nicht. Mit **/lang/bindings** bekommt man eine List von Mandanten, die über eine bestimmte Sprache verfügen (für Suchoperationen). Eine einzelne neue Sprache ohne Sprachwerte kann über **/lang/create** erstellt werden. Eine Sprache zu löschen (inklusive aller Sprachwerte) ermöglicht die Schnittstelle **/lang/delete**.

Weiter mit den Schnittstellen für Sprachwerte. Der Endpunkte **/sym/pagination** liefert seitenweise alle Symbole. Nachdem sehr viele Symbole in der Datenbank liegen wird die Ladedauer dadurch verringert, dass immer nur eine bestimmte Anzahl geladen wird. So werden nur bei Bedarf weitere Symbole geliefert ("*Lazy Loading*"). Durch **/sym/analyze** kann ein einzelnes Symbol analysiert werden. Als Rückgabe kommt eine Auflistung aller Orte an denen das analysierte Symbol verwendet wird. Nur ein Symbol das nirgends Anwendung findet kann mit **/sym/delete** entfernt werden. Der Endpunkt **/lval/pagination** funktioniert genau wie die *pagination* von Symbolen nur für textuelle Sprachwerte. Die Schnittstelle **/lval/create** ermöglicht die Erstellung eines einzelnen Sprachwertes für ein bereits existierendes Symbol. Sprachwerte können mit **/lval/update** auch einzeln bearbeitet und mit **/lval/delete** einzeln gelöscht werden.

Zuletzt geht es noch um den Import und Export. Über **/import/json** können JSON-Dateien ausgelesen und Sprachen sowie Symbole (inklusive von Sprachwerten) daraus importiert werden. Der Unterschied zum CSV-Import (**/import/csv**) ist, dass CSV-Dateien auf ein Symbol oder eine Sprache pro Datei limitiert sind. Exports werden auch für Sprachen und Symbole angeboten. Dafür gibt es die Endpunkte **/export/json** und **/export/csv**. Jedoch gilt für CSV-Dateien die gleiche Limitierung wie bei den Imports.

| Zuordnung | HTTP | Pfad |
|---|---|---|
| Sprache | GET | /lang/cards |
| | GET | /lang/ls/copy |
| | GET | /lang/tenants |
| | GET | /lang/bindings |
| | POST | /lang/create |
| | POST | /lang/do/copy |
| | DELETE | /lang/delete |
| Symbole | GET | /sym/pagination |
| | GET | /sym/analyze |
| | DELETE | /sym/delete |
| LValues | GET | /lval/pagination |
| | POST | /lval/create |
| | PUT | /lval/update |
| | DELETE | /lval/delete |
| Imports | POST | /import/json |
| | POST | /import/csv |
| Exports | GET | /export/json |
| | GET | /export/csv |

**Table 2: Übersicht der Serverschnittstellen**

# 7. IMPLEMENTIERUNG DES CLIENTS

## Clientstruktur und Frameworks

Auch beim Client muss die Projektumgebung eingerichtet werden, bevor die Implementierung des Clients beginnen kann. Der Client wurde mit Typescript anstatt von Javascript geschrieben, um zumindest etwas besser an das vom Java-Server vorgegebene Typsystem anzuknüpfen. Zudem wurde für das Projekt die Node-Version 16.20.0 verwendet. Als Build-Management-Tool kommt *Vite* (Version 4.3.9) zum Einsatz. Als Client-Development-Framework wird *SvelteKit* eingesetzt. *SvelteKit* basiert auf Svelte und bietet eine einfache Komponentenbauweise, ein eigenes Routingkonzept, sowie weitere Extras (z.B. Icons und Animationen). Außerdem wird ein CSS-Framework names *Tailwind* verwendet, welches auch in allen anderen Frontends bei *Projekt 29* Anwendung findet. Es bietet einheitliche und standardisierte Klassen für die Formatierung von HTML-Elementen und entfert zuvor die Standardformatierung von allen Elementen.

## Überblick des Clients

Wie bereits unter Punkt 5 beschrieben wurde der Client visuell in vier Teile mit unterschiedlichen Schwerpunkten aufgeteilt. Für die Erstellung des Frontends, welches nachfolgend gezeigt wird, wurden nur die in Tabelle 2 aufgelisteten Schnittstellen verwendet. In Abbildung 5 wird zunächst der Kopfbereich gezeigt, welcher in jedem Teil der Anwendung zu sehen ist. Er ist die Hauptnavigationskomponente und ermöglicht die Navigation durch die vier Kernbereiche der Software. Außerdem soll in der weiteren Entwicklung noch ein **Home**-Bereich eingeführt werden, auf dem Statistiken zu den Standardsprachen dargestellt werden. Nun folgt der


**Abbildung 5: Kopfbereich des LValue-Moduls**

Sprachteil des Moduls. Dieser besitzt eine eigene Navigationsleiste (siehe Abbildung 6 links) und folgende Funktionalitäten:

- Übersicht über die Standardsprachen
- Suche von Mandanten nach Tenant-ID
- Suche von Mandanten nach verfügbaren Sprachen

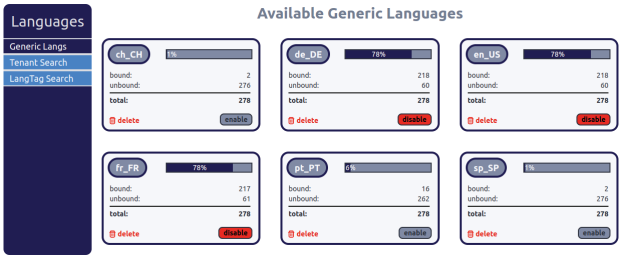



**Abbildung 6: Übersicht über die Standardsprachen**

Aus Gründen des Umfangs wird nicht jede Ansicht in dieser Zusammenfassung gezeigt. Das Herzstück des LValue-Moduls ist der sogenannte LValue-Editor, welcher in Abbildung 7 zu sehen ist. Er kann sowohl alle Symbole als auch alle LValues des Systems auflisten. Außerdem ist es möglich einzelne Symbole zu löschen. Das gleiche gilt auch für LValues, mit dem Unterschied, dass diese auch erstellt und geändert werden können. Da für jedes Symbol beliebig viele LValues existieren können, ist es auch möglich nur alle LValues für ein bestimmtes Symbol zu laden. Dabei kann oben links im Editor die gewünschte Zielgruppe (z.B. alle Symbole, alle LValues, nur Template-Symbole, usw.), ausgewählt werden. Daraufhin erscheinen je nach Zielgruppe in der linken Leiste die gewünschten Symbole oder in der Mitte alle LValues. Klickt man auf ein gelistetes Symbol erscheinen in mittleren Teil des Editors alle zugehörigen LValues. Die Leiste rechts ist für alle Operationen (erstellen, bearbeiten, löschen) auf ausgewählten Objekten zuständig. Das Info-Icon mittig oben kann benutzt werden, um ein ausgewähltes Symbol zu analysieren. Sollte dieses Symbol keine Anwendungsfälle haben, so kann es im erschienenen Popup optional gelöscht werden. Zum Schluss bleiben noch


**Abbildung 7: LValue-Editor von *pso2admin***

das Import- und das Export-Segment übrig. Für beide gibt es einen Konfigurations-Wizard, welcher durchlaufen werden muss, bevor ein Import oder Export getätigt werden kann. Dabei geht es darum die folgenden Punkte festzulegen:

- Dateiformat: JSON / CSV
- Exporttyp: Sprache / Sprachwerte
- Umgebung: TID / GEN
- welche(r) Sprache(n) oder Sprachwert(e)

Erst wenn diese Konfigurationen eingestellt wurden ist es möglich eine Datei hochzuladen (Import) oder einen Export anzufordern. Allerdings hat der Import-Teil der Anwendung noch eine weitere Funktionalität. Dabei handelt es sich um das Kopieren von Sprachen. So können Sprachen von einer Umgebung auf eine andere kopiert werden. Zunächst muss dafür eine Zielumgebung (ein spezifischer Mandant oder die **GEN**-Umgebung) ausgewählt werden. Dann werden alle kopierbaren Sprachen aufgelistet. Aus diesen können dann Sprachen ausgewählt werden, die inklusive all ihrer Sprachwerte auf die Zielumgebung kopiert werden. Das ist beispielsweise nützlich wenn eine kundenspezifische Sprache neuerdings auch als Standardsprache bereitgestellt werden soll. Oder wenn ein Kunde bestimmte Sprachwerte ändern möchte, können diese aus der **GEN**-Umgebung auf diesen Mandanten kopiert und dort angepasst werden.

## Endergebnis des Projekts

Das Kernergebniss in diesem Projekt ist die fertige erste Version des Sprachmoduls für den neuen Administrationsbereich (*pso2admin*) der DSM-Software. Also genau die Software, die im Vorangegangenen erklärt und beschrieben wurde. Insgesamt ist das Unternehmen mit der übergebenen Software sehr zufrieden. Sowohl mit dem Code als auch mit der Dokumentation dazu. Um einen groben Überblick über den Umfang des Projektes zu geben, zeigt Tabelle 3 die Menge an Code und Dokumentation. Zudem ist festzuhalten, dass alle unter Punkt 1 aufgezählten Ziele der Software erreicht wurden. Das LValue-Modul ist im Administrationsbereich integriert und ist bereits in Benutzung. Es ist in Planung das Modul um eine SQL-Skript-Generierung für die Datenbankskripte im DEV-Bereich zu erweitern und mit KI-Übersetzung auszustatten.

| **Scope** | **Lines of Code** | **Lines of Comments** |
|---|---|---|
| Serverside | 3.217 | 1.496 |
| Clientside | 2.919 | 98 |
| Total | 6.136 | 1.598 |

**Table 3: Überblick über die Codebasis des LValue-Moduls**

# Verbesserung des Gewerbekundenportals der mb Support GmbH für Real Estate-Versicherungsmakler durch kundenorientierte Daten-Visualisierungen im Versicherungsbereich

Maximilian Fischer
mb Support GmbH
Friedenstraße 18
93053 Regensburg
E-Mail: maximilian2.fischer@mbsupport.de

Professor Dr. Frank Herrmann
Innovationszentrum für Produktionslogistik und Fabrikplanung
OTH Regensburg
Galgenbergstraße 32, 93053 Regensburg
E-Mail: frank.herrmann@oth-regensburg.de

**Abstract: Dieses Projekt konzentriert sich auf die Neugestaltung des Controlling-Dashboards, um die Attraktivität des Gewerbekundenportals zu steigern. Dabei liegt der Fokus auf der Problemstellung, der Relevanz und der Definition der Anforderungen für das neue Dashboard. Die Arbeit beschreibt, wie Prototypen entwickelt werden, die den festgelegten Anforderungen entsprechen. Zudem wird der Evaluationsprozess erläutert, um die Wirksamkeit der Neugestaltung zu bewerten und sicherzustellen, dass die Ziele erreicht werden.**

## 1 Problemstellung

Die mb Support GmbH ist ein mittelständisches Unternehmen mit Sitz in Regensburg, das sich auf die Entwicklung von Softwarelösungen für Versicherungsmakler spezialisiert hat. Ihr Hauptprodukt, open VIVA c2, ist eine webbasierte und modulare Software, die es Versicherungsmaklern ermöglicht, ihre Vertriebs- und Geschäftsprozesse effizient zu steuern, Kommunikation zu optimieren und Abrechnungsprozesse zu automatisieren. Das Unternehmen verfolgt das Ziel, die gesamte Wertschöpfungskette für Versicherungsmakler abzubilden und ihnen Werkzeuge zur Verfügung zu stellen, um ihre Aufgaben bestmöglich zu erfüllen [Supp18].

Ein zentrales Element des Angebots der mb Support GmbH ist das Gewerbekundenportal. Dieses Portal wird von den Kunden der Versicherungsmakler genutzt, um ihnen einen besseren Service zu bieten. Es bietet eine Vielzahl nützlicher Funktionen, wie die Verwaltung versicherter Objekte, Vertragsprüfung und die direkte Meldung von Schäden. Durch die Benutzerfreundlichkeit des Portals und praktische Tools, wie den Prämienrechner, wird die Zusammenarbeit zwischen Versicherungsmaklern und ihren gewerblichen Kunden erleichtert.


*Abb. 1: Stakeholder-Beziehung zueinander*

Die Motivation hinter diesem Projekt liegt darin, die Attraktivität des Gewerbekundenportals weiter zu steigern. Dieses Ziel verfolgt das Unternehmen nicht nur, um die Nutzererfahrung für bestehende Kunden zu verbessern, sondern auch, um neue Kunden aus dem Immobilienbereich zu gewinnen. Das Gewerbekundenportal spielt eine entscheidende Rolle bei der Überzeugung von Geschäftsführern und Managern potenzieller Kunden. Das Controlling-Dashboard, das prominent auf der Startseite platziert ist, beeinflusst maßgeblich den ersten Eindruck des Portals. Daher hat die mb Support GmbH beschlossen, das bestehende Controlling-Dashboard genauer zu untersuchen.

In Gesprächen mit Versicherungsmaklern aus dem Immobilienbereich wurde festgestellt, dass die bisherigen Anforderungen und Bedürfnisse der Nutzer nicht vollständig berücksichtigt wurden, was zu einer unzureichenden Nutzung des Controlling-Dashboards führte.
Es besteht somit eine Forschungslücke in Bezug auf die Benutzerfreundlichkeit und Effektivität des aktuellen Controlling-Dashboards sowie die Identifizierung von Verbesserungsmöglichkeiten, um es für die Nutzer ansprechender und nützlicher zu gestalten.

Diese Projekt hat das Ziel, diese Frage zu klären und Lösungsansätze zur Steigerung der Attraktivität des Gewerbekundenportals zu entwickeln.

## 2 Relevanz

Die Analyse des aktuellen Controlling-Dashboards offenbart eine Vielzahl von Herausforderungen und Problemen. Hierzu zählen unter anderem unübersichtliche Kreisdiagramme, die aufgrund der Vielzahl von Werten die prozentualen Verteilungen nicht klar darstellen können [Poll20a]. Ebenso fällt ein Widget ins Auge, das Informationen zu "Verträgen nach Vertragsstatus" bereitstellt, jedoch durch das Fehlen eines Linienverlaufs die Daten schwerer verständlich macht. Die Einhaltung von Designrichtlinien bei Balken- und Säulendiagrammen wurde ebenfalls vernachlässigt, da die Anordnung der Werte nach Größe nicht entsprechend berücksichtigt wurde [Poll20a]. Die Ästhetik des Dashboards entspricht nicht den Nutzererwartungen, da es an visuellen Reizen und einem zeitgemäßen Design mangelt. Dies führt zu dem zentralen Problem, dass die präsentierten Daten nicht diejenigen Informationen repräsentieren, die für die Nutzer von Relevanz sind, wie sich aus den Experteninterviews deutlich ergibt.

*Abb. 2: Screenshot Gewerbekundenportal Controlling-Dashboard Quelle: Gewerbekundenportal der mb Support GmbH*

Um diese Herausforderungen zu bewältigen und das Controlling-Dashboard gezielt zu verbessern, sind Experteninterviews mit Versicherungsmaklern von entscheidender Bedeutung [PlätoJ]. Diese Interviews bieten weitere Einblicke in die Schwächen des aktuellen Dashboards und liefern wertvolle Empfehlungen zur Lösung dieser Probleme. Die Experten äußern Bedenken hinsichtlich der Sinnhaftigkeit einiger dargestellter Daten und betonen die Notwendigkeit einer sinnvolleren und informativeren Darstellung, um den Benutzern einen klaren Mehrwert zu bieten.

Außerdem wird die visuelle Gestaltung des Dashboards diskutiert, und die Experten empfehlen die Integration von Eyecatchern und ein zeitgemäßes Design, um die Attraktivität des Dashboards zu steigern. Die Bedeutung der Schadenhäufigkeit als entscheidende Kennzahl für das Immobilienwesen wird ebenfalls betont, wobei auf eine verstärkte Ausrichtung der Daten auf Versicherungsobjekte hingewiesen wird. Dies ermöglicht eine bessere Analyse von Trends und Entwicklungen im Bereich der Schadenhäufigkeit. Zusätzlich wird angemerkt, dass eine Vereinfachung des Schaden-Bearbeitungsprozesses durchaus nützlich wäre. Schließlich schlagen die Experten vor, die Interaktivität mit den Datenvisualisierungen zu erhöhen und die Berichterstattung zu verbessern, wobei sie erfolgreiche Programme wie SAP, Salesforce und Power BI als Inspirationsquelle nennen.

## 3 Anforderungen an das neue Controlling-Dashboard

Aufgrund der Erkenntnisse aus der Experteninterviews konnten Kriterien für eine erfolgreiche Lösung des Problems definiert werden. Diese Kriterien lassen sich in funktionale und nicht funktionale Anforderungen unterteilen. Funktionale Anforderungen beschreiben, welche Aufgaben das System erfüllen soll, während nicht funktionale Anforderungen sich auf die Qualität beziehen, in der diese Funktionalitäten erbracht werden sollen [HeinoJ]. Hierzu zählen Aspekte wie Leistung, Benutzerfreundlichkeit und Zuverlässigkeit, die für die Umsetzung eines effektiven Controlling-Dashboards von entscheidender Bedeutung sind.

**Funktionale Anforderungen**

- Schadenhäufigkeit aufzeigen
- Schaden-Bearbeitungsstatus visualisieren
- Hohe Interaktivität mit Datenvisualisierungen
- Größerer Fokus auf Versicherungsobjekte

**Nicht funktionale Anforderungen**

- Optisch ansprechende Gestaltung
- Hohe Benutzerfreundlichkeit
- Schnell an relevante Informationen kommen

## 4 Ansätze der Lösung

In den folgenden Abschnitten werden die Ansätze zur Lösung des Problems der Verbesserung des Controlling-Dashboards präsentiert. Hier wird insbesondere der User-Centered Designprozess erläutert, der in diesem Projekt eine zentrale Rolle spielt.

Dieser Prozess ermöglicht es, das Dashboard so zu gestalten, dass es die Bedürfnisse und Erwartungen der Benutzer optimal erfüllt [InteoJ]. Zudem werden die Entscheidungen zur Diagrammauswahl und zum Design erläutert, da die Wahl geeigneter Diagrammtypen entscheidend ist, um Daten und Kennzahlen

effektiv zu visualisieren und aussagekräftige Einblicke zu gewinnen.

### 4.1 Durchführung des Projekts anhand des User-Centered Designprozesses

Im Rahmen der Lösungsansätze zur Verbesserung des Controlling-Dashboards wird ein strukturierter Ansatz verfolgt, der den User-Centered Designprozess einbezieht. Dieser Designprozess zeichnet sich durch seine iterative Natur aus und stellt den Benutzer und seine Bedürfnisse in den Mittelpunkt des Gestaltungsprozesses. Ziel dieses Prozesses ist es, ein herausragendes Benutzererlebnis zu schaffen, indem die Anforderungen, Ziele und Perspektiven der Benutzer von Anfang an berücksichtigt werden [SeoboJ].

Der User-Centered Designprozess gliedert sich in mehrere aufeinanderfolgende Phasen:

1. Anforderungsanalyse: In dieser ersten Phase werden die Bedürfnisse und Anforderungen der Benutzer sorgfältig identifiziert. Hierbei kommen verschiedene Methoden wie Interviews, Umfragen und Beobachtungen zum Einsatz, um ein umfassendes Verständnis für die Benutzer und ihre Anforderungen zu entwickeln [InteoJ].
2. Benutzerforschung: In dieser Phase werden die verschiedenen Benutzergruppen genauer untersucht. Informationen über ihre Eigenschaften, Vorlieben, Fähigkeiten und Ziele werden gesammelt. Diese Erkenntnisse dienen als Grundlage für das spätere Design des Produkts oder Dienstes [InteoJ].
3. Ideenfindung und Konzeptentwicklung: Basierend auf den gesammelten Informationen werden in dieser Phase verschiedene Designideen entwickelt. Diese Ideen werden zu Konzepten weiterentwickelt, die darauf abzielen, die Bedürfnisse und Anforderungen der Benutzer bestmöglich zu erfüllen [InteoJ].
4. Prototypenentwicklung: Durch die Erstellung von Prototypen des Designs wird eine frühe Bewertung durch die Benutzer ermöglicht. Diese Prototypen können von einfachen Papiermodellen bis hin zu interaktiven digitalen Prototypen reichen [InteoJ].
5. Evaluation und Testing: In dieser letzten Phase werden die erstellten Prototypen von den Benutzern getestet und bewertet. Feedback und Verbesserungsvorschläge der Benutzer werden gesammelt und in den Designprozess integriert, um sicherzustellen, dass das endgültige Produkt den Bedürfnissen und Erwartungen der Benutzer entspricht [InteoJ].

Dieser strukturierte Prozess ermöglicht es, das Controlling-Dashboard kontinuierlich zu optimieren und sicherzustellen, dass es die Anforderungen und Erwartungen der Benutzer erfüllt. Indem die Benutzer frühzeitig in den Gestaltungsprozess einbezogen werden und ihre Perspektiven berücksichtigt werden, ist es möglich, ein Dashboard zu entwickeln, das effektiv, effizient und zufriedenstellend in der Anwendung ist [SeoboJ].

### 4.2 Auswahl von Diagrammtypen

Eine der herausragenden Herausforderungen in der Konzeption eines effektiven Controlling-Dashboards besteht in der Auswahl geeigneter Diagrammtypen, die die Daten und Kennzahlen anschaulich visualisieren. Die gezielte Auswahl der Diagrammtypen spielt eine entscheidende Rolle, um die Effektivität der Berichterstattung zu steigern und fundierte Entscheidungen im Controlling-Prozess zu unterstützen [Poll20b].

Die verschiedenen Diagrammtypen, darunter Liniendiagramme, Säulendiagramme, gestapelte Säulendiagramme, Kreisdiagramme, Treemaps und kartografische Diagramme, bieten vielfältige Möglichkeiten zur Darstellung von Daten und ermöglichen es, verschiedene Aspekte und Muster aufzuzeigen. Beispielsweise eignen sich Liniendiagramme besonders zur Darstellung von zeitlichen Verläufen [Poll20c], während Säulendiagramme die Vergleichbarkeit von Datenpunkten betonen [Poll20d]. Moderne Ansätze wie Treemaps ermöglichen die Darstellung hierarchischer Daten [LauboJ], während kartografische Diagramme geografische Zusammenhänge visualisieren [TibcoJ].

Die Wahl des richtigen Diagrammtyps erfordert eine sorgfältige Prüfung und Berücksichtigung verschiedener Faktoren, einschließlich der Art der Daten, die visualisiert werden sollen, und der Ziele der Berichterstattung [Poll20c].

### 4.3 Designentscheidungen

In der Gestaltung des Projekts werden vielfältige Designentscheidungen getroffen, um eine benutzerfreundliche und ansprechende Erfahrung zu schaffen. Diese Entscheidungen basieren auf bewährten Prinzipien der Webgestaltung und werden von verschiedenen Quellen inspiriert. Im Folgenden werden die wesentlichen Designaspekte und Prinzipien zusammengefasst, die zur Gestaltung des Projekts beitragen.

Diese umfassen die Berücksichtigung von asymmetrischer Balance, Einheitlichkeit, Konsistenz, Kontrasten, Symmetrie, Farbdesign und Inspiration aus etablierten Designs sowie die Anordnung von Elementen. Die Integration dieser Designprinzipien soll eine ansprechende und leicht verständliche Benutzererfahrung gewährleisten.

1. Asymmetrische Balance: Einige Gestaltungsbereiche im Projekt werden bewusst asymmetrisch gestaltet, wobei dennoch eine symmetrische Gesamtwirkung erzielt wird. Dieser Ansatz basiert auf den Prinzipien der Webgestaltung, wie sie von Jason Beaird in seinem Artikel "The Beautiful Principles of Webdesign" [Beai07] beschrieben werden. Beaird betont, dass unterschiedliche Größen, Formen, Farbtönungen und Platzierungen, trotz asymmetrischer Elemente, ein harmonisches Gesamtbild erzeugen können.
2. Einheitlichkeit und Konsistenz: Um die Benutzererfahrung zu optimieren, wird auf wiederholte Elemente gesetzt, die eine konsistente und leicht erkennbare Gestaltung ermöglichen [Beai07].
3. Kontraste und Hervorhebung: Kontraste werden gezielt verwendet, um wichtige Informationen hervorzuheben. Dieses Prinzip wird von Jason Beaird in seinem Artikel zu Webdesign-Kriterien [Beai07] betont.
4. Symmetrie: In den meisten Entwürfen des Projekts wird eine symmetrische Anordnung der Elemente angewendet, basierend auf Erkenntnissen aus der Studie von Kerstin Müller und Dr. Martin Schrepp [MüSc14]. Diese Studie zeigt, dass Symmetrie im Allgemeinen als positiv empfunden wird und eine angenehme Wirkung auf die Nutzer hat.
5. Farbdesign und Inspiration aus etablierten Designs: Die Farbpalette wird an bewährten Farbpaletten ausgerichtet, wie es in der Standardfarbpalette von Power BI [Magg23] der Fall ist. Somit wird die Inspiration aus anderen etablierten Designs berücksichtigt, um bewährte Gestaltungselemente in das eigene Produkt zu integrieren und die Benutzererfahrung zu optimieren [Beai07].
6. Abstände und Anordnung: Bei der Anordnung von Elementen wird darauf geachtet, gleichmäßige Abstände einzuhalten und eine kohärente Anordnung zu schaffen, um ein ästhetisch ansprechendes Gesamtbild zu erzeugen. Dies basiert auf allgemeinen Prinzipien des visuellen Designs [MüSc14].

## 5 Beschreibung der Lösungen

In diesem Abschnitt werden die entwickelten Widgets vorgestellt, die dazu dienen, das Controlling-Dashboard des Gewerbekundenportals der mb Support GmbH zu verbessern. Die Widgets wurden mithilfe des UX-Design-Tools Figma erstellt, welches eine leistungsstarke Plattform zur Erstellung interaktiver Prototypen und Designs bietet [Figm22].

### 5.1 Tracken des Bearbeitungsstatus von Schäden

Die Idee dazu entstand während des Interviews mit Interviewpartner zwei, der bei der Durchsicht des bereits bestehenden Controlling-Dashboards die Nützlichkeit des Widgets "Offene Schäden nach Schadentracking" erkannte.

Das aktuelle Widget ermöglicht eine Visualisierung der Verteilung aller Schäden nach Schadenstatus. Der Schadenstatus bezieht sich auf den aktuellen Bearbeitungstand des gemeldeten Schadens. Dieses Widget ist jedoch sowohl visuell als auch inhaltlich noch verbesserungswürdig. Dennoch erfüllt es einen wichtigen Zweck, da man auf einen Blick erkennen kann, in welchem Schadenstatus sich die Schäden aktuell befinden. Wenn der Bereich der Schadenanlage deutlich größer ist als der Bereich der offenen Unterlagen, könnte dies darauf hinweisen, dass der Makler mit der aktuellen Bearbeitung der Schäden nicht hinterherkommt. Eine hohe Anzahl offener Unterlagen könnte wiederum auf unerledigte Arbeit seitens der Hausverwaltung hinweisen.

Inspiriert von dem Interesse der Makler am konkreten Status der verschiedenen Schäden, werden Widgets entwickelt, die den Portalnutzern einen genauen Überblick über die letzten Schäden mit Statuswechsel bietet. Dies ermöglicht eine detaillierte Darstellung des Schadentrackings und zeigt auf, wo Handlungsbedarf besteht, um eine schnellere Schadensbearbeitung zu ermöglichen. Der Entwurf umfasst eine Liste von Objekten (Abb. 3), bei denen Schadensänderungen vorgenommen wurden. Jede Änderung wird durch ein Symbol auf der rechten Seite repräsentiert, um einen schnellen Überblick zu ermöglichen, und wird auch in Textform angezeigt, wenn der Nutzer über das jeweilige Objekt schwebt. Dies hilft, Unklarheiten zu vermeiden. Zudem wird die zugehörige Vertragsnummer angezeigt, um eine genaue Zuordnung der Objekte zu ermöglichen.

Die Interviews ergaben, dass für die Nutzer im Real Estate-Bereich der Objektname im Vordergrund stehen sollte, da dies die wichtigste Eigenschaft für sie ist. Beim Klicken auf einen dieser Einträge öffnet sich die eigentliche Hauptfunktionalität.


*Abb. 3: Widget mit der Liste der Objekte, bei denen sich der Bearbeitungsstatus geändert hat*

Das Design des detaillierten Schadentrackings (Abb. 4) hat seine Inspiration aus der Sendungsverfolgung verschiedener Paketlieferdienste gezogen. Die Farben spiegeln die Firmenfarben der mb Support GmbH wider. Die Symbole repräsentieren die verschiedenen Status, wobei der aktuelle Status durch einen dezenten Farbverlauf im Symbol hervorgehoben wird. Dadurch wird der Fokus direkt auf den aktuellen Status gelenkt und der wichtigste Punkt im Diagramm markiert. Wenn ein Status eintritt, wird dieser mit Uhrzeit und Datum versehen, um den zeitlichen Verlauf nachvollziehen zu können. Die Symbole repräsentieren auch, ob eine E-Mail oder eine Aufgabe zu dem jeweiligen Schaden angelegt wurde. Das Portal bietet die Möglichkeit, E-Mails und Aufgaben mit den einzelnen Schadensfällen zu verknüpfen, sodass sie eindeutig im zeitlichen Verlauf zugeordnet werden können. So wird schnell klar, ob es noch zusätzliche Informationen zu dem jeweiligen Schaden gibt.


*Abb. 4: detailliertes Schaden-Bearbeitungsstatus-Tracking-Widget*

## 5.2 Ungewöhnlich hohe Schadenfrequenz

Das nächste Widget bezieht sich ebenfalls auf Schäden und dient der Analyse und Identifizierung von Trends anhand ungewöhnlich hoher Schadensfrequenzen in bestimmten Regionen. Die Schadensfrequenzen werden laut der Versicherungsmathematik [Deut18] als das Verhältnis der Anzahl der Schäden zur Anzahl der Versicherungsverträge innerhalb einer bestimmten Zeitspanne definiert. Diese Unterscheidung ist von großer Bedeutung, da die Verträge in Relation zu den aufgetretenen Schäden betrachten müssen. In Regionen, in denen weniger Objekte versichert sind, ist es nachvollziehbar, dass auch weniger Schäden auftreten. Zur visuellen Darstellung werden drei verschiedene Widgets kombiniert, um das volle Potenzial zur Entscheidungsfindung zu entfalten.

Diese Kombination besteht aus einem gestapelten Säulendiagramm, einem einfachen Säulendiagramm und einem geographischen Diagramm, das eine Karte von Deutschland zeigt (Abb. 5). Durch einzelne Punkte auf der Karte werden Regionen mit ungewöhnlich hohen Schadenfrequenzen repräsentiert.

*Abb. 5: Umfassende Visualisierung der außergewöhnlich hohen Schadenhäufigkeiten in allen Widgets*

Das obere Säulendiagramm zeigt den prozentualen Anteil der ungewöhnlich hohen Schadenfrequenzen in den jeweiligen Regionen, untergliedert in die jeweiligen Anteile der Schadensursachen. Dadurch lassen sich nicht nur Region identifizieren, sondern auch gleich bestimmen, was die Ursachen dafür sind. Das darunter befindliche Säulendiagramm stellt die prozentuale Verteilung der häufigsten Schadensursachen dar und fungiert gleichzeitig als Legende für das obere Diagramm. Diese Darstellung bringt den Vorteil, dass die Nutzer des Controlling-Dashboards auf einen Blick erkennen können, welche Schadensursachen am häufigsten auftreten und wie sie in Bezug auf die Gesamtheit der Schäden verteilt sind. Dadurch wird die Interpretation der Daten erleichtert und eine schnellere Identifizierung von Trends oder Schwerpunkten ermöglicht. Es werden nur die Regionen dargestellt, in denen eine ungewöhnlich hohe Schadenfrequenz ermittelt wurde.

Zusätzlich bietet das gestapelte Säulendiagramm zur prozentualen Verteilung der Schaden-Region eine Drill-Down-Funktion. Durch Klicken auf eine Säule im Diagramm oder auf ein Bundesland auf der Karte öffnet sich eine detaillierte Ansicht, die sich auf die konkreten Postleitzahlen der betroffenen Regionen

innerhalb des Bundeslandes konzentriert. Zum Beispiel beim Klicken auf die Säule, die Bayern repräsentiert, oder den entsprechenden Bereich auf der Karte wird eine Ansicht geöffnet, die auf Bayern beschränkt ist und nicht auf gesamt Deutschland. Gleichzeitig ändert sich das Diagramm mit den Schadensursachen, indem es die Berechnung der durchschnittlich hohen Schadenfrequenzen nun auf das betreffende Bundesland einschränkt. Diese ermöglicht eine detaillierte Ansicht der Datensätze, was zu einer genaueren Identifizierung des Problems beitragen kann. Die Hauptfunktionalität der Karte ist durch ihre Übersichtlichkeit zu ermöglichen, auf einen Blick gleich die betroffenen Bereiche zu identifizieren. Die drei Diagramme zusammen ermöglichen es dem Nutzer, verschiedene Zusammenhänge zwischen Regionen und Schäden zu betrachten und Überlegungen anzustellen auf mögliche Verbindungen.

### 5.3 Verteilung Objekttypen

Auf dem Gewerbekundenportal existiert eine Ansicht, die die Verteilung der Vertragsarten darstellt. Allerdings ist diese in den meisten Fällen je nach Versicherungsmaklerbereich sehr ähnlich und daher wenig nützlich. Aus den Interviews ergab sich jedoch, dass ein Diagramm, das die Verteilung der Objekttypen zeigt, sinnvoller wäre. Unter Objekttypen versteht man vom Nutzer generierte Kategorien, mit denen Objekte unterteilt werden können. Zum Beispiel könnte der Nutzer die Kategorie "Einfamilienhaus" erstellen und die entsprechenden Objekte mit diesem Objekttyp kennzeichnen. Als Darstellungsform wurde eine Treemap gewählt, da sie die komplexe Datenstruktur anschaulich darstellt und Platz spart. Durch die anpassbaren Rechtecke in der Treemap können Unterschiede und Muster in der Verteilung schnell erkannt werden. Die hierarchische Struktur der Daten wird effektiv visualisiert, was zu einer übersichtlichen Darstellung führt. Je größer das Rechteck, desto größer ist der prozentuale Anteil des Objekttyps an der Gesamtzahl der Objekte [LauboJ].

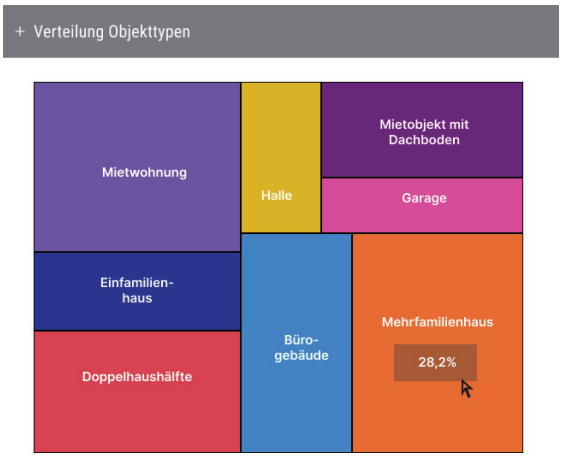



*Abb. 6: Verteilung Objekttypen-Widget*

Durch die Visualisierung der prozentualen Anteile der Objekttypen können Mitarbeiter im Real Estate Muster und Trends erkennen, die ihnen bei der Informationsgewinnung und fundierten Entscheidungsfindung helfen können. Das Widget erleichtert somit die effiziente Datenanalyse und unterstützt dabei wertvolle Erkenntnisse aus den verwalteten Objekten zu gewinnen.

## 6 Evaluierung

In diesem Kapitel dieser Arbeit steht die Evaluierung des Controlling-Dashboard-Entwurfs im Fokus. Die Umfrage, die ein zentrales Element dieses Prozesses darstellt, zielt darauf ab, Feedback von potenziellen Nutzern zu sammeln.
Dies geschieht durch die Bereitstellung anonymer Fragebögen, die eine umfassende Analyse von verschiedenen Aspekten des Dashboards ermöglichen. Der Evaluierungsprozess deckt die gesamte Bandbreite von der Entwicklung der Fragebögen bis zur umfassenden Auswertung und Interpretation der erhobenen Daten ab. Zusätzlich wird überprüft, ob die definierten Anforderungen erfüllt wurden.

### 6.1 Auswertung Fragebögen

Die Umfrage wurde unter Einbeziehung von 31 Nutzern des Gewerbekundenportals durchgeführt und basierte auf einem Fragebogen, der eine Mischung aus geschlossenen und offenen Fragen verwendete. Geschlossene Fragen setzten auf Skalen, in diesem Fall von 1 (trifft nicht zu) bis 5 (trifft voll zu), um Meinungen zu quantifizieren und die Verständlichkeit zu erhöhen, während offene Fragen den Nutzern die Möglichkeit gaben, individuelle Anmerkungen und Kommentare abzugeben [Bild12]. Die Umfrage wurde mithilfe von Google Forms durchgeführt, einem Online-Tool zur Erstellung und Verwaltung von Umfragen [GoogoJ]. Die Analyse der Umfrage-

ergebnisse stützte sich auf deskriptive Statistik, wobei Mittelwerte, Mediane und Standardabweichungen verwendet wurden, um die Datenlage zu verstehen [NovuoJ].

Die Fragen der Umfrage waren breit gefächert und deckten verschiedene Aspekte des Dashboards ab, darunter Ästhetik, Verständlichkeit, Nützlichkeit, Sinnhaftigkeit der Interaktionsmöglichkeiten und die wahrgenommene Steigerung der Attraktivität. Die Auswertung der Umfrageergebnisse lieferte folgende Ergebnisse.

Ästhetik: Die ästhetische Bewertung des Designs zeigte gemischte Meinungen bei den Teilnehmern. Die Durchschnittswerte lagen zwischen 2,9 und 4,06, wobei die Standardabweichungen eine höhere Variation in den Antworten zeigten. Dies deutet darauf hin, dass die Meinungen der Teilnehmer bezüglich des Designs stärker voneinander abweichen können.

Verständlichkeit der Diagramme: Die Teilnehmer bewerteten die Verständlichkeit der Diagramme insgesamt positiv. Der Durchschnitt lag zwischen 3,16 und 4,09, und die geringen Standardabweichungen wiesen auf eine geringe Variation in den Antworten hin. Dies deutet auf eine allgemeine Zustimmung zur Verständlichkeit hin.

Nützlichkeit der Widgets: Die Ergebnisse zeigten, dass die Teilnehmer die Widgets insgesamt als nützlich empfanden.
Die Durchschnittsbewertungen reichten von 2,84 bis 3,87, und auch hier waren die Standardabweichungen moderat. Dies deutet auf eine positive Einschätzung hin.

Bewertung der Interaktionsmöglichkeiten: Die Analyse der Umfrageergebnisse zur Sinnhaftigkeit der Interaktionsmöglichkeiten ergab, dass die Teilnehmer die gebotenen Optionen im Allgemeinen als sinnvoll erachteten. Der Durchschnitt lag bei 3,84 und die Standardabweichung zeigte moderate Abweichungen in den Antworten.

Wie sehr konnte die Attraktivität gesteigert werden?: Die Ergebnisse zeigten, dass alle Teilnehmer zu einem gewissen Maß eine gesteigerte Attraktivität durch das neue Dashboard empfanden. Der Durchschnitt lag bei 2,97, und der Median bei 3. Dies deutet darauf hin, dass die Mehrheit der Teilnehmer eine mäßige bis moderate Steigerung der Attraktivität erwartete, wobei die Standardabweichung von 0,75 darauf hinweist, dass die Antworten insgesamt nicht stark variierten und eine gewisse Einheitlichkeit in der Wahrnehmung herrschte.

Die offenen Fragen fokussierten sich vor allem auf das Verteilungsobjekttypen-Widget, was die Notwendigkeit einer Überprüfung seiner Darstellung und möglicher alternativer Ansätze aufzeigte, da die Nutzer teilweise das Diagramm als unverständlich wahrnahmen. Einige Teilnehmer äußerten auch Kritik an der Farbauswahl im Dashboard.

In den verschiedenen Kategorien erhielt das Widget zur Verfolgung des Bearbeitungsstatus von Schäden durchweg die höchste positive Resonanz, während das Widget zur Verteilung der Objekttypen trotz einer tendenziell positiven Bewertung die niedrigste Bewertung erhielt. Insgesamt bieten die Umfrageergebnisse wertvolle Einblicke in die Nutzerperspektive und dienen als Grundlage für mögliche Verbesserungen des Controlling-Dashboards.

### 6.2 Vergleich mit gestellten Anforderungen

Im Bereich der funktionalen Anforderungen (Tab. 1) standen die Darstellung von Schadenhäufigkeiten, die Visualisierung des Schaden-Bearbeitungsstatus und die Schaffung hoher Interaktivität bei der Datenvisualisierung im Fokus. Diese Anforderungen wurden erfolgreich umgesetzt.

Schadenhäufigkeiten darstellen: Dies wurde durch das Schadenhäufigkeit nach Regionen-Widget realisiert, das es ermöglicht, Schadenhäufigkeiten auf regionaler Ebene anschaulich darzustellen.

Visualisierung des Schaden-Bearbeitungsstatus: Die Umsetzung erfolgte durch die Einführung der Liste der Bearbeitungsstatusänderungen und des detaillierten Bearbeitungsstatus-Tracking-Widgets. Diese Widgets bieten eine klare und aktuelle Visualisierung des Bearbeitungsstatus von Schäden.

Hohe Interaktivität bei der Datenvisualisierung: Sowohl das Schadenhäufigkeit nach Regionen-Widget als auch das Schaden-Bearbeitungsstatus-Widget zeichnen sich durch hohe Interaktivität aus, was es den Benutzern ermöglicht, die Daten nach ihren Bedürfnissen zu filtern und zu erkunden.

Fokus verstärkt auf Versicherungsobjekte: Zusätzlich wurde durch das Verteilung-Objekttypen-Widget der Fokus verstärkt auf Versicherungsobjekte gelegt, was einer funktionalen Anforderung entspricht und die Ausrichtung des Dashboards optimierte.

Im Hinblick auf die nicht-funktionalen Anforderungen (Tab. 1), die eine ansprechende visuelle Gestaltung, hohe Benutzerfreundlichkeit und schnellen Zugriff auf relevante Informationen betrafen, ergaben sich folgende Ergebnisse.

Ansprechende visuelle Gestaltung: Die Gestaltung der Widgets wurde durch Umfrageergebnisse positiv bewertet, wobei das Schaden-Bearbeitungsstatus-Widget besonders hervorragte.

Hohe Benutzerfreundlichkeit: Die Umfrage zeigte, dass die Widgets als selbsterklärend wahrgenommen wurden, obwohl einige Widgets besser bewertet wurden als andere. Dies deutet auf eine insgesamt hohe Benutzerfreundlichkeit hin.

Schneller Zugriff auf relevante Informationen: Im Fall des Schaden-Bearbeitungsstatus wurde eine messbare Beschleunigung des Datenzugriffs festgestellt, da Benutzer nun auf einfache Weise den aktuellen Bearbeitungsstatus von Schäden abrufen können.

Insgesamt verdeutlicht diese Zusammenfassung, dass die funktionalen und nicht-funktionalen Anforderungen erfolgreich umgesetzt wurden, wobei die Widgets einen positiven Beitrag zur Verbesserung der Datenvisualisierung und Benutzererfahrung leistet.

## 7 Fazit und Ausblick

Die Gestaltung benutzerfreundlicher und an den Bedürfnissen der Nutzer orientierter Anwendungen stellt einen entscheidenden Erfolgsfaktor für Produkte dar. Dieser Aspekt wurde in der Arbeit wiederholt betont.
Um die Anforderungen der Nutzer bestmöglich zu erfüllen, ist eine enge Zusammenarbeit mit den Nutzern unerlässlich. Im Verlauf dieser Arbeit wurden sämtliche Schritte des User-Centered Designprozesses durchlaufen, wobei jedoch zu betonen ist, dass es sich um einen iterativen Prozess handelt. Der letzte Abschnitt der Arbeit beinhaltet die Integration der neuen Erkenntnisse aus den Umfragen in die Prototypen. Dieser Schritt erfolgt nicht nur einmal, sondern möglicherweise mehrfach. Hierbei ist eine sorgfältige Abwägung erforderlich, welche Verbesserungen umgesetzt werden sollen und welche nicht. Eine uneingeschränkte Zustimmung zu erreichen, ist äußerst unwahrscheinlich. Daher ist eine kontinuierliche Abwägung notwendig, wobei die Meinung der Mehrheit eine größere Rolle spielt. Die Ergebnisse verdeutlichen, dass eine Mehrheit der Befragten den Widgets zustimmt. Jedoch wirft die niedrigere Bewertung des Verteilung-Objekttypen-Widgets, insbesondere im Hinblick auf die gewählte Diagrammform, die Frage auf, ob die Verwendung einer anderen Darstellungsform sinnvoller wäre.

Die Forschungsfrage dieser Arbeit zielt darauf ab, ob die Attraktivität des Gewerbekundenportals durch die Neuimplementierung gesteigert werden kann. Aus diesem Grund wurde dieser Frage im Fragebogen besondere Aufmerksamkeit geschenkt. Die durchschnittliche Bewertung von 2,97 lässt eine Tendenz zur Zustimmung erkennen, wobei zu berücksichtigen ist, dass diese Frage einen umfangreichen Kontext umfasst. Das Controlling-Dashboard verändert nicht sämtliche Aspekte des Gewerbekundenportals, beeinflusst jedoch den ersten und entscheidenden Eindruck, wie schon im Kapitel 1 erwähnt. Daher kann die Bewertung im Hinblick auf diese umfassende Fragestellung für die mb Support GmbH als zufriedenstellend betrachtet werden.

Für eine umfassende Beurteilung der Auswirkungen des neu gestalteten Controlling-Dashboards auf die Portalnutzer ist es jedoch notwendig, den Einsatz des Dashboards über einen längeren Zeitraum zu beobachten. Besonders nach einer längeren Eingewöhnungsphase der Nutzer sollte regelmäßiges Kundenfeedback eingeholt werden. Diese kontinuierliche Aufgabe plant die mb Support GmbH über einen erweiterten Zeitrahmen hinweg umzusetzen.

## Literaturverzeichnis


[AsenoJ] ASENDIA: 5 Vorteile von Tracking beim Versand. oJ, URL https://www.asendia.de/asendia-insights/5-vorteile-von-tracking-beim-versand. - abgerufen am 2023-07-15

[Beai07] BEAIRD, JASON: The Principles of Beautiful Web Design. 2007, URL https://www.home.uni-osnabrueck.de/elsner/Skripte/Material/HTML/Artikel/The%20Principles%20of%20Beautiful%20Web%20Design.pdf. - abgerufen am 2023-07-17

[Bild12] BILDUNG, BUNDESZENTRALE FÜR POLITISCHE: Fragetypen und Antworten. 2012, URL https://www.bpb.de/lernen/angebote/grafstat/grafstat-software/51677/fragetypen-und-antworten/. - abgerufen am 2023-08-11. — bpb.de

[Deut18] DEUTSCHE AKTUARVEREINIGUNG E. V.: Leitfaden_Versicherungsmathematik. 2018, URL https://aktuar.de/aktuar-werden/dav-ausbildung/qualitaetsmanagement/Documents/Leitfaden_Versicherungsmathematik.pdf. - abgerufen am 2023-08-12

[Figm22] FIGMA: Was ist Figma? 2022, URL https://webdesign.tutsplus.com/de/what-is-figma--cms-32272a. - abgerufen am 2023-07-15. — Web Design Envato Tuts+

[GoogoJ] Google Formulare: App zum Erstellen von Onlineformularen | Google Workspace. oJ, URL https://www.google.de/intl/de/forms/about/. - abgerufen am 2023-08-11

[HeinoJ] HEINI, R.: Funktionale, nicht funktionale Anforderungen. oJ, URL http://www.anforderungsmanagement.ch/in_depth_-vertiefung/funktionale_nicht_funktionale_anforder-ungen/index.html. - abgerufen am 2023-07-12

[InteoJ] THE INTERACTION DESIGN FOUNDATION: What is User Centered Design? oJ, URL https://www.interaction-de-sign.org/ literature/topics/user-centered-design. - abgerufen am 2023-07-11. — The Interaction Design Foundation

[LauboJ] LAUBHEIMER, PAGE: Treemaps: Data Visualization of Complex Hier-archies. oJ, URL https://www.nngroup.com/ articles/treemaps/. - abgerufen am 2023-08-15. — Niel-sen Norman Group

[Magg23] MAGGIESMSFT: Verwenden von Berichtdesigns in Power BI Desktop - Power BI. 2023, URL https://learn.microsoft.com/de-de/power-bi/create-reports/desktop-report-themes. - abgerufen am 2023-07-30

[MüSc14] MÜLLER, KERSTIN ; SCHREPP, MARTIN: Einfluss von Layout-Prinzi-pien auf die ästhetische Wahrnehmung von Web-Seiten: On the Influence of Layout Heuristics on the Aesthetic Perception of Web Pages. In: i-com Bd. 13, Oldenbourg Wissenschaftsverlag (2014), Nr. 2, S. 38–46

[NovuoJ] NOVUSTAT: Deskriptive Statis-tik. oJ, URL https://novustat.com/statistik-glossar/deskriptive-statistik.html. - abgerufen am 2023-08-22. — Statistik Service

[PlätoJ] PLÄTZ, SOPHIA: Leitfaden: Kunden-in-terviews im Produktmanagement. oJ, URL https://dmexco.com/de/stories/leitfaden-best-practices-fuer-kundeninterviews-im-produktmanagement/. - abgerufen am 2023-08-12. — DMEXCO

[Poll20a] Controlling-Berichte professio-nell gestalten: Reportings aussagekräftig erstellen und visualisieren. In: POLLMANN, R.: Controlling-Berichte professionell gestalten: Reportings aussa-gekräftig erstellen und visualisieren. 1. Auflage. Freiburg München Stuttgart : Haufe Group, 2020 — ISBN 978-3-648-13048-3, S. 104

[Poll20b] Controlling-Berichte professio-nell gestalten: Reportings aussagekräftig erstellen und visualisieren. In: POLLMANN, R.: Controlling-Berichte professionell gestalten: Reportings aus-sagekräftig erstellen und visualisieren. 1. Auflage. Freiburg München Stuttgart : Haufe Group, 2020 — ISBN 978-3-648-13048-3, S. 151

[Poll20c] Controlling-Berichte professio-nell gestalten: Reportings aussagekräftig erstellen und visualisieren. In: POLLMANN, R.: Controlling-Berichte professionell gestalten: Reportings aussa-gekräftig erstellen und visualisieren. 1. Auflage. Frei-burg München Stuttgart : Haufe Group, 2020 — ISBN 978-3-648-13048-3, S. 90

[Poll20d] Controlling-Berichte professio-nell gestalten: Reportings aussagekräftig erstellen und visualisieren. In: POLLMANN, R.: Controlling-Berichte professio-nell gestalten: Reportings aussa-gekräftig erstellen und visualisieren. 1. Auflage. Freiburg München Stuttgart : Haufe Group, 2020 — ISBN 978-3-648-13048-3, S. 94

[SeoboJ] SEOBILITY: User Centered De-sign - Definition und Erklärung - Seobility Wiki. oJ, URL https://www.seobility.net/de/wiki/User_Cen-tered_Design. - abgerufen am 2023-08-21

[Supp18] MB SUPPORT GMBH: Unter-nehmen | Standardsoftware für Versicherungsmak-ler, Assekuradeure und Spezialversicherer. 2018, URL https://mbsupport.de/unternehmen/. - abgeru-fen am 2023-07-29

[TibcoJ] TIBCO SOFT-WARE: Was ist ein geografisches Dia-gramm? oJ, URL https://www.tibco.com/de/reference-center/what-is-a-geographical-chart. - abgerufen am 2023-08-02. — TIBCO Software

# Use Case-basierter Vergleich zwischen Case-Centric- und Object-Centric Process Mining im Supply Chain Resilience Management

David Messal
Hochschule Pforzheim
Tiefenbronnerstr. 65
75175 Pforzheim
messalda@hs-pforz-heim.de

Frank Morelli
Hochschule Pforzheim
Tiefenbronnerstr. 65
75175 Pforzheim
frank.morelli@hs-pforz-heim.de

Frank Schätter
Hochschule Pforzheim
Tiefenbronnerstr. 65
75175 Pforzheim
frank.schaetter@hs-pforz-heim.de

Florian Haas
Hochschule Pforzheim
Tiefenbronnerstr. 65
75175 Pforzheim
florian.haas@hs-pforz-heim.de

**ABSTRACT**
Die zunehmende Komplexität globaler Lieferketten und deren Anfälligkeit gegenüber Störungen erfordern innovative Ansätze im Supply Chain Resilience Management (SCRM). Dieser Artikel untersucht vergleichend die Anwendbarkeit von Case-Centric Process Mining (CCPM) und Object-Centric Process Mining (OCPM) zur Analyse und Stärkung der Resilienz in Supply Chains. Während sich CCPM auf die Analyse einzelner Prozessinstanzen fokussiert, erlaubt OCPM eine multidimensionale Betrachtung komplexer Objektbeziehungen und Prozessverläufe.Auf Basis synthetischer Datensätze für Order-to-Cash- und Purchase-to-Pay-Prozesse werden beide Ansätze anhand von Use Cases evaluiert, die sich am SCOR-Modell orientieren und die Resilienz-Attribute Zuverlässigkeit, Reaktionsfähigkeit und Agilität adressieren. Die Analyse mit dem Process-Mining-Tool Celonis zeigt, dass CCPM eine schnelle Identifikation zentraler Schwachstellen ermöglicht, jedoch durch Informationsverluste bei der Aggregation limitiert ist. OCPM hingegen deckt Kettenreaktionen zwischen Objekten auf, identifiziert objektspezifische Ursachen von Prozessinstabilitäten und ermöglicht eine granularere KPI-Betrachtung ohne erneute Datenaufbereitung.

**SCHLÜSSELWÖRTER**
Case-Centric Process Mining (CCPM), Object-Centric Process Mining (OCPM), Supply Chain Resilience Management, SCOR-Modell, Prozessanalyse

## EINLEITUNG

In der heutigen datengetriebenen Weltwirtschaft hat sich Process Mining zu einer Schlüsseltechnologie entwickelt, um Geschäftsprozesse transparent abzubilden und Optimierungspotenziale zeitnah zu identifizieren. Der Erfolg des Marktführers Celonis sowie der prognostizierte globale Marktumsatz von 1,5 Milliarden US-Dollar bis Ende 2025 verdeutlichen die steigende Bedeutung dieser Technologie. Zugleich erhöht sich durch zunehmende Komplexität und globale Vernetzung die Anfälligkeit von Lieferketten gegenüber Störungen wie Naturkatastrophen, Cyberattacken und Handelskonflikten. Besonders die COVID-19-Pandemie hat deutlich gemacht, wie entscheidend die Stärkung der Resilienz globaler Lieferketten ist (vgl. Eggers et al. 2021; Kerremans et al. 2024).
In diesem Kontext etabliert sich das Object-Centric Process Mining (OCPM) zunehmend als Weiterentwicklung des traditionellen, fallbasierten Process Mining - Case-Centric Process Mining (CCPM). Im Gegensatz zum klassischen Ansatz betrachtet OCPM Prozesse im Kontext von mehreren Objekten und berücksichtigt Ereignisse, die ggf. mit mehreren Objekten in Relation stehen. Dadurch ermöglicht OCPM die Analyse komplexerer Zusammenhänge und tiefere Einblicke in Prozessabläufe (vgl. Lund et al. 2020).
Mit dem digitalen Zeitalter entstanden Techniken wie Tabellenkalkulation und Business Intelligence (BI), welche jedoch meist auf einzelne Aufgaben wie z.B. das Reporting begrenzt waren. Process Mining hingegen ermöglicht eine umfassende Analyse komplexer Ende-zu-Ende-Prozesse. Process Mining verbindet dabei zwei Wissenschaftsdizsiplinen miteinander: Data Science und Process Science. Die Verwendung von prozessbezogenen Ereignisdaten („Event Logs“) und deren automatische Umsetzung in Ist-Prozessmodelle erleichtern die Analyse von Prozessabweichungen und die Visualisierung von Engpässen (vgl. Reinkemeyer 2022; van der Aalst 2016).
Wil van der Aalst gilt als Erfinder dieser Technologie und sorgt als einer der führender Wissenschaftler in diesem Bereich für deren Weiterentwicklung wie z.B. OCPM. Process Mining wurde ursprünglich vor allem akademisch erforscht und ist seit etwa 2010 zunehmend auch in der Praxis vorzufinden, um Geschäftsprozesse transparenter zu gestalten. Unternehmen wie Siemens oder BMW nutzen diese Methode erfolgreich, um Ineffizienzen wie Doppelzahlungen von Rechnungen zu vermeiden oder die Lieferketten zu optimieren. Trotz der mittlerweile umfangreichen Nutzung kommerzieller Process-Mining-Tools (z.B. Celonis und SAP Signavio) bleibt ein Großteil ihres Potenzials noch ungenutzt (vgl. van der Aalst 2022).
Die Nutzung und vor allem die Einführung von Process Mining erfordert tiefgreifende Veränderungen der Ge-

schäftsprozesse, Organisationsstrukturen und Unternehmenskulturen. Neben der technologischen Umsetzung sind klare Ziele, Führungskompetenz und wirksames Change-Management entscheidend, um Widerstände innerhalb der Organisation zu überwinden. Unterstützt durch Cloud-Technologien, „Industrial Internet of Things“ (IioT) und zunehmend autonome, selbstlernende Systeme entwickelt sich Process Mining perspektivisch zu einer zentralen Plattform für Business Operational Intelligence. Wachstumstreiber sind dabei digitale Transformation, Künstliche Intelligenz, Hyperautomation und Nachhaltigkeit (vgl. Reinkemeyer 2022; Kerremans et al. 2024).

Durch die zunehmende Weiterentwicklung dieser Technologie wird es möglich, komplexe Prozesse evidenzbasiert zu analysieren, Prozessmodelle aus Ereignisdaten zu generieren, deren Konformität zu prüfen und Prozessverläufe vorherzusagen.

Klassisches fallbasiertes Process Mining (CCPM) betrachtet jeweils ein einzelnes Objekt isoliert, einen sogenannten „Fall“ bzw. „Case“ wie z.B. einen Kundenauftrag. CCPM stößt an Grenzen, wenn verschiedene Ereignisse mit mehreren Objekte verbunden sind. Da in realen Unternehmensprozessen meist komplexe und vielseitige Beziehungen zwischen verschiedenen Objekten und unterschiedlichen Aktivitäten vorzufinden sind, etabliert sich Object-Centric Process Mining (OCPM) zunehmend als möglicher Industriestandard. Ziel dabei ist es, eine präzisere Analyse komplexer Prozessabläufe ermöglicht (vgl. van der Aalst 2023b).

Das Ziel dieses Artikels ist es, Process Mining im Bereich SCRM zu diskutieren. Es wird eine vergleichende Analyse zwischen CCPM und OCPM durchgeführt. Der weitere Verlauf dieses Artikels gestaltet sich deshalb wie folgt: Zunächst erfolgt eine Einführung in das Konzept des Supply Chain Resilience Managements, gefolgt von einer Gegenüberstellung der wesentlichen Merkmale von traditionellem (CCPM) und objektzentriertem (OCPM) Process Mining. Anschließend werden das Forschungsdesign erläutert sowie die Ansätze anhand konkreter Use Cases, basierend auf dem SCOR-Modell, verglichen. Die Schlussbetrachtung diskutiert die Ergebnisse und fasst diese zusammen.

## GRUNDLAGEN

### Supply Chain Resilience Management

Resilienz in Lieferketten (Supply Chain Resilience) bezeichnet die Fähigkeit einer Organisation, Störungen in ihren Materialflüssen frühzeitig zu antizipieren, sich angemessen darauf vorzubereiten und wirksam auf deren Eintritt zu reagieren. Das Konzept hat in den vergangenen Jahren erheblich an wissenschaftlicher wie praktischer Relevanz gewonnen, nicht zuletzt aufgrund prominenter ökonomischer und gesellschaftlicher Krisenereignisse, die verdeutlicht haben, dass auf maximale Effizienz ausgerichtete Lieferkettenstrukturen in Krisensituationen rasch ihre Belastungsgrenzen erreichen. Aus theoretischer Perspektive wird Resilienz als die Fähigkeit eines Systems verstanden, sich nach einem disruptiven Ereignis zu regenerieren und im Idealfall langfristig eine gesteigerte Leistungsfähigkeit zu erzielen. Eine resiliente Lieferkette zeichnet sich folglich dadurch aus, dass sie nach Eintritt einer Störung die Funktionsfähigkeit wiederherstellt. Hierzu sind geeignete Managementsysteme erforderlich, welche sowohl die Intensität des Schadens als auch dessen Dauer zu begrenzen vermögen. Eine vertiefte Auseinandersetzung mit dieser Thematik – insbesondere im Lichte der Post-Corona-Krise – findet sich bei Yang et al. (2021).

Im Rahmen unserer eigenen Forschungsarbeiten befassen wir uns seit mehreren Jahren mit den Implikationen des Supply Chain Risk Management (SCRM) für Unternehmen und Logistikverantwortliche. In diesem Kontext wurden methodische Ansätze entwickelt, die eine Bewertung des strategischen Resilienzstatus einer Lieferkette auf Grundlage transaktionaler Daten ermöglichen. Hierbei erfolgt die Analyse im Wesentlichen durch die strukturierte Auswertung von Lieferscheindaten, welche die Materialflüsse zu den unmittelbaren Zulieferern (Tier-1) sowie zu den direkten Abnehmern (Tier-1) eines Unternehmens abbilden. Ein zentrales Ergebnis unserer bisherigen Forschungsarbeiten war die Definition geeigneter sogenannter Key Resilience Areas (KRAs) sowie deren empirische Anwendung im Rahmen einer Fallstudie (vgl. Schätter et al. 2022; 2023).

Aufbauend auf diesen Vorarbeiten verfolgen wir das Ziel, die Potenziale des Process Mining für das SCRM nutzbar zu machen. Insbesondere das OCPM-Konzept erweist sich in diesem Zusammenhang als vielversprechend, da es nicht lediglich die Analyse isolierter Datensätze, sondern auch die Untersuchung prozessübergreifender Zusammenhänge erlaubt. Die datengestützte Resilienzbewertung verstehen wir hierbei als einen Top-down-orientierten Ansatz: Bereits eine begrenzte Menge transaktionaler Daten (z. B. Lieferpositionen) kann genutzt werden, um strategische Verwundbarkeitspunkte innerhalb der Lieferkette auf relativ aggregierter Ebene zu identifizieren – etwa kritische Netzwerkpartner basierend auf spezifischen Teilemerkmalen oder geographischen Standorten. Auf dieser initialen Kritikalitätsbewertung aufbauend ermöglicht der Einsatz detaillierter Ereignisdaten in Verbindung mit Process-Mining-Methoden eine vertiefte Analyse potenzieller Schwachstellen sowie eine Untersuchung der damit verknüpften Prozessstrukturen.

### CCPM: Konzepte und Funktionalitäten

Im Gegensatz zu Ansätzen wie das Business Process Management (BPM) oder BI ermöglicht Process Mining eine dynamische und datenbasierte Analyse realer Prozesse, wodurch präzisere und praxisnähere Ergebnisse erzielt werden (vgl. van der Aalst 2016). Die zentrale Datenquelle im Process Mining sind sogenannte Event Logs, die detaillierte Informationen über den Ablauf von Geschäftsprozessen enthalten. Sie bestehen typischerweise aus einer Fall-ID, der ausgeführten Aktivität und einem Zeitstempel, enthalten typischerweise aber auch weitere Attribute wie Ressourcen oder Kosten. Diese Da-

ten werden automatisch in Systemen wie Enterprise Resource Planning (ERP) oder Warehouse Management System (WMS) Anwendungen erfasst und ermöglichen die präzise Rekonstruktion und Analyse realer Prozessabläufe. Eine hohe Datenqualität ist dabei essenziell, um Aktivitäten korrekt zuzuordnen, zeitlich einzuordnen und verlässliche Analysen durchzuführen. Dafür ist häufig eine Vorverarbeitung und Bereinigung der Rohdaten erforderlich (vgl. van der Aalst 2016; Suriadi et al. 2017). Abbildung 1 zeigt ein Beispiel für einen typischen Event Log mit den Attributen Case ID, Aktivität, Zeitstempel, Kundennummer und Artikelnummer. Die dargestellten Prozessverläufe für die Fälle 1001 und 1002 verdeutlichen, wie Process Mining auf Basis der Case ID unterschiedliche Varianten analysieren, Abweichungen identifizieren und Optimierungspotenziale aufzeigen kann.

| _CASE_KEY | ACTIVITY | EVENTTIME | Kundennummer | Artikelnummer |
|---|---|---|---|---|
| 1001 | Receive Order | 2023-08-18 00:00:00 | C127 | A460 |
| 1001 | Confirm Order | 2023-08-19 00:00:00 | C127 | A460 |
| 1001 | Generate Delivery Document | 2023-08-22 00:00:00 | C127 | A460 |
| 1001 | Ship Goods | 2023-08-23 00:00:00 | C127 | A460 |
| 1001 | Send Invoice | 2023-08-26 00:00:00 | C127 | A460 |
| 1001 | Clear Invoice | 2023-08-27 00:00:00 | C127 | A460 |
| 1002 | Receive Order | 2023-04-04 00:00:00 | C127 | A456 |
| 1002 | Confirm Order | 2023-04-05 00:00:00 | C127 | A456 |
| 1002 | Generate Delivery Document | 2023-04-07 00:00:00 | C127 | A456 |
| 1002 | Ship Goods | 2023-04-10 00:00:00 | C127 | A456 |
| 1002 | Send Invoice | 2023-04-11 00:00:00 | C127 | A456 |
| 1002 | Write Off Invoice | 2023-04-12 00:00:00 | C127 | A456 |
| 1002 | Send Invoice | 2023-04-13 00:00:00 | C127 | A456 |
| 1002 | Clear Invoice | 2023-04-15 00:00:00 | C127 | A456 |

*Abbildung 1: Event Log eines Order-to-Cash Prozesses*

Das traditionelle Process Mining lässt sich in insgesamt drei Haupttypen unterteilen: Discorvery, Conformance und Enhancement. *Process Discovery* ist die am häufigsten genutzte Technik im Process Mining und ermöglicht die automatische Ableitung von Prozessmodellen aus Event Logs, ohne dass der Anwender Vorkenntnisse des Prozesses benötigt. Sie schafft vollständige Transparenz über reale Abläufe und zeigt den tatsächlichen IST-Prozess, deckt Abweichungen, Engpässe und Ineffizienzen auf und unterstützt so gezielte Optimierungsmaßnahmen (vgl. van der Aalst et al. 2012; van der Aalst 2016; Peters und Nauroth 2019). *Konformitätsprüfungen* im Process Mining ermöglichen den Vergleich zwischen realen Prozessabläufen und bestehenden Modellvorgaben, um Abweichungen gezielt zu identifizieren. Diese Erkenntnisse erhöhen den Wert von Prozessmodellen, indem sie deren Anpassung erleichtern und Optimierungspotenziale aufzeigen, z.B. durch die Ableitung eines neuen Soll-Modells basierend auf dem sogenannten Happy Path (häufigste Prozessvariante) (vgl. Carmona et al. 2022; Peters und Nauroth 2019). Der *Enhancement-Ansatz* im Process Mining zielt darauf ab, bestehende Prozessmodelle mithilfe von Event-Log-Daten gezielt zu optimieren oder zu erweitern. Dabei wird entweder das Modell durch zusätzliche Perspektiven wie Ressourcen oder Kennzahl-Daten angereichert bzw. so angepasst, dass es die realen Abläufe präziser widerspiegelt. Ziel ist ein robusteres, leistungsfähigeres Modell, das näher an der tatsächlichen Prozessrealität liegt (vgl. Leoni 2022; van der Aalst et al. 2012; van der Aalst 2016). Van der Aalst erweitert 2022 das bestehende Process-Mining-Framework um vier weitere Typen: Performance Analysis zur Bewertung von Effizienz und Zeit, Comparative Process Mining zum Benchmarking unterschiedlicher Prozesse, Predictive Process Mining zur Vorhersage und Optimierung künftiger Abläufe sowie Action-Oriented Process Mining zur Ableitung konkreter Maßnahmen und Automatisierung repetitiver Aufgaben (vgl. van der Aalst 2022).

**CCPM: Grenzen des traditionellen Ansatzes**

In den letzten Jahren hat sich Process Mining als datenbasierter Ansatz etabliert, um reale Prozessabläufe automatisiert aus Ereignisdaten abzuleiten. Diese klassische Vorgehensweise stößt dabei jedoch an Grenzen, da diese impliziert, dass jedes Ereignis nur einem einzelnen Fall zugeordnet ist. Eine Annahme, die in der Praxis häufig nicht zutrifft, da Ereignisse häufig mehrere Objekte betreffen. Dies führt zu Problemen wie Konvergenz und Divergenz, die Analyseergebnisse verzerren und die Aussagekraft der Modelle einschränken. In komplexen Prozessen, wie einem Purchase-to-Pay Prozess (P2P), sind Ereignisse oft mit mehreren Objekten verbunden (z.B. kann eine Bestellung mit mehreren Rechnungen und Zahlungen verknüpft sein), was zu fehlerhaften Zuordnungen, Duplikaten und dem Verlust wichtiger Abhängigkeiten führt. Dadurch wird eine umfassende Analyse realer Prozessstrukturen erheblich eingeschränkt (vgl. Adams und van der Aalst 2021; Berti et al. 2023a; van der Aalst 2019). Ein weiteres zentrales Problem traditioneller Process-Mining-Ansätze ist das sogenannte „Flattening“, bei dem man komplexe, objektzentrierte Ereignisdaten in lineare Abläufe überführt, um sie analysierbar zu machen. Dabei gehen jedoch wichtige Informationen verloren oder werden verfälscht, beispielsweise durch fehlende, doppelte oder falsch angeordnete Ereignisse, was die Aussagekraft und Genauigkeit der Analyse deutlich einschränkt (vgl. Adams et al. 2022). Zur Bewältigung komplexer, interaktiver Prozesse im Process Mining wurden Ansätze wie Proclets und XOC entwickelt, die jedoch an Komplexität und Datenvolumen scheiterten. Van der Aalst schlägt daher OCPM und erweiterte, objektzentrierte Event Logs vor, die mehrere Objekte und deren Beziehungen erfassen und so eine realistischere und flexiblere Prozessanalyse ermöglichen (vgl. van der Aalst 2019).

**OCPM: Konzepte und Funktionalitäten**

Object-Centric Process Mining (OCPM) stellt einen weiterentwickelten Ansatz dar, der die Einschränkungen klassischer, fallbasierter Process-Mining-Methoden überwindet. Anstatt Ereignisse einem einzelnen Fall zuzuordnen, arbeitet OCPM direkt mit den tatsächlichen Objekten und Ereignissen, wie sie in ERP Systemen, Customer Relationship Management (CRM) Systemen oder Manufacuring Execution Systemen (MES) auftreten. Diese Systeme enthalten häufig komplexe eins-zu-viele- oder viele-zu-viele-Beziehungen, die von traditionellen Methoden nur unzureichend abgebildet werden. Um diese Komplexität zu handhaben, ist bei klassischen Ansätzen oft eine aufwändige Datenumwandlung nötig,

die zu Informationsverlusten und Analysefehlern führen kann. OCPM nutzt hingegen objektzentrierte Eventdaten (OCED), die reale Prozessstrukturen präziser erfassen und somit eine genauere und praxisnähere Analyse ermöglichen (vgl. Berti et al. 2023b). Durch die Betrachtung von vielfältigen Objektbeziehungen wird eine deutlich differenziertere Prozessanalyse ermöglicht, da Ereignisse nicht mehr auf eine einzelne Instanz beschränkt sind, sondern mehreren, miteinander verknüpften Objekttypen zugeordnet werden können. Dadurch entsteht ein komplexes Prozessmodell, das über die klassische Betrachtung von Aktivitäten und Zeitstempeln hinaus auch die Beziehungen zwischen Objekten abbildet. Diese erweiterte Sichtweise reduziert nicht nur den Aufwand der Datenaufbereitung, sondern liefert auch tiefere Einblicke in komplexe Abläufe (vgl. van der Aalst 2023a). Um OCPM technisch umzusetzen, wird das Object-Centric Event Log (OCEL) Format verwendet, das im Gegensatz zu eXtensible Event Stream (XES) mehrere Objekte pro Ereignis unterstützt und somit die angesprochene Probleme wie Konvergenz und Divergenz vermeidet. OCEL ist zudem speichereffizient, unterstützt moderne Datenformate wie JSON und XML und bietet Schnittstellen für gängige Programmiersprachen zur Verarbeitung großer Datenmengen (vgl. Ghahfarokhi et al. 2021).

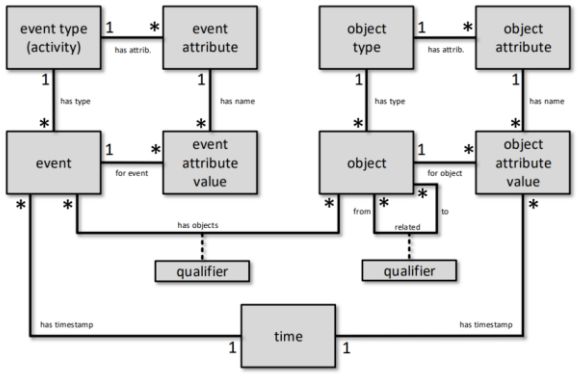



*Abbildung 2: OCEL 2.0 Metamodell (Berti et al. 2023b)*

Abbildung 2 zeigt das Metamodell von OCEL 2.0, das die Grundlage für OCPM bildet. Ereignisse (Events) repräsentieren Aktivitäten im Prozess, etwa eine Bestellung oder Zahlung, und sind jeweils mit einem Zeitstempel sowie weiteren Attributen (Event Attributes) versehen. Diese Events werden nach Event-Typen kategorisiert. Eine zentrale Rolle spielen zudem Objekte (Objects), die physische oder abstrakte Einheiten wie Produkte, Rechnungen oder Verträge (Object Types) darstellen und eigene Attribute besitzen. OCPM berücksichtigt zwei wesentliche Beziehungen: die Event-zu-Objekt-Beziehung (E2O), bei der ein Event mit mehreren Objekten verknüpft ist, und die Objekt-zu-Objekt-Beziehung (O2O), die Beziehungen zwischen Objekten unabhängig von Ereignissen abbildet (vgl. Berti et al. 2023b). O2O-Beziehungen verbinden Objekte dauerhaft miteinander und bilden dabei statische Strukturen wie Hierarchien oder Stücklisten ab. Im Gegensatz dazu sind E2O-Beziehungen dynamisch, da sie durch die Teilnahme von Objekten an zeitlich markierten Ereignissen entstehen. Während Ereignisattribute sich auf ein einzelnes, bereits zeitlich erfasstes Event beziehen, besitzen Objektattribute eigene Zeitstempel, da sich deren Werte – etwa Preis oder Gewicht – im Zeitverlauf ändern können. Pro Objekt und Zeitpunkt ist dabei nur ein Attributwert erlaubt, um die Konsistenz sicherzustellen (vgl. van der Aalst 2023b).

Auch im OCPM gibt es verschiedenen Typen von Process Mining. Der objektzentrierte Ansatz zur *Process Discovery* nutzt OCELs, um für jeden Objekttyp separate Prozessmodelle zu generieren, die anschließend zu einem übergreifenden Modell zusammengeführt werden können. Dabei ermöglichen Metriken wie die Anzahl beteiligter Objekte die Erstellung objektzentrierter Directly-Follows-Graphs (DFGs) oder Petri-Netze (OCPNs), in denen Aktivitäten durch gemeinsame Objekte verbunden sind. Dieser Ansatz erlaubt eine flexible Analyse komplexer Prozesse mit mehreren Objekttypen (vgl. van der Aalst 2023a). Alternativ existieren Methoden wie die Nutzung eines führenden Objekttyps oder ereignis-objektbasierte Graphen, um Konvergenz- und Divergenzprobleme zu vermeiden (vgl. van der Aalst 2023b). Bei der objektzentrierten *Konformitätsprüfung* verwendet man ein Modell (z.B. OC-DFGs oder OCPNs), um die Übereinstimmung mit dem Event Log zu prüfen. Dabei werden zunächst flache Modelle je Objekttyp analysiert und anschließend überprüft, ob Aktivitäten gemäß den im Modell definierten Kardinalitätsregeln korrekt ausgeführt wurden. Das Ergebnis, etwa in Form eines Recall-Werts, zeigt, wie gut das beobachtete Verhalten dem Modell entspricht (vgl. van der Aalst 2023a, 2023b). Die object-zentrierte *Performance Analyse* zielt darauf ab, Engpässe in Prozessen mit mehreren interagierenden Objekten zu identifizieren. Da traditionelle Ansätze lineare Fallsequenzen voraussetzen, werden in OCPM neue Metriken wie Synchronisations-, Pooling-, Verzögerungs- und Flusszeit eingesetzt, um komplexe, überlappende Abläufe realitätsnah zu bewerten. Methoden wie der Ansatz von Park et al. (2022) nutzen OCELs und OCPNs, um präzise Kennzahlen zu berechnen, ohne die Prozessstruktur zu verfälschen (vgl. Park et al. 2022).

Process Mining geht über klassische Anwendungen wie Discovery, Conformance und Performance hinaus und umfasst auch datenbezogene Qualitätsanalysen sowie den Einsatz von Machine Learning (vgl. van der Aalst 2023a). Im OCPM gewinnt insbesondere *Object-Centric Operational Support* (OCOS) an Bedeutung, das proaktiv Vorhersagen, Empfehlungen und Korrekturen bereitstellt, um Prozesse im laufenden Betrieb zu optimieren. Durch die Einbindung mehrerer Objekttypen steigt dabei die Prognosegenauigkeit. Technologien wie Machine Learning und Generative AI unterstützen die Analyse, ersetzen jedoch nicht die zentralen Methoden des Process Mining (vgl. van der Aalst 2024).

**OCPM: Herausforderungen**

Die Fallstudie von Berti et al. 2023a zeigt, dass die Implementierung von OCPM in einem realen P2P-Prozess

erhebliche Herausforderungen bei der Datenextraktion und -verarbeitung mit sich bringt. Besonders die Gewinnung objektzentrierter Event Logs aus einem SAP-System erwies sich in diesem Kontext aufgrund komplexer und nicht standardisierter Datenstrukturen als aufwändig und erforderte tiefes Prozessverständnis sowie gezielte technische Methoden (vgl. Berti et al. 2023a).
Zudem stößt OCPM bei der Vorhersage komplexer Prozesse an Grenzen, da viele Ansätze nur einzelne Prozesse isoliert betrachten. Dies beinhaltet beispielsweise mögliche Zusammenhänge zwischen Purchase-to-Pay und Order-to-Cash-Prozessen. Um präzisere Prognosen, beispielsweise für das Verhalten von Lieferketten, zu ermöglichen, müssen Interaktionen zwischen Objekten und deren Attributen stärker in Vorhersagemodelle einbezogen werden. Zudem bestehen Herausforderungen bei der Skalierbarkeit sowie der korrekten Datenaufbereitung und Synchronisation aus unterschiedlichen Prozessen (vgl. Galanti et al. 2023).
OCPM hat in den letzten Jahren deutlich an Aufmerksamkeit gewonnen, unter anderem durch Fortschritte wie das OCEL-Format und neue Open-Source-Tools. Dennoch bestehen weiterhin Herausforderungen, insbesondere bei der Skalierbarkeit, der Extraktion objektzentrierter Daten aus ERP- und CRM-Systemen sowie der Modellierung und Visualisierung komplexer Prozessmodelle (vgl. Berti et al. 2023c).

### Vergleich zwischen CCPM und OCPM

Wie in Abbildung 3 dargestellt unterscheiden sich CCPM und OCPM erheblich, wobei OCPM viele Schwächen von CCPM eliminieren kann.

| Aspekt | CCPM | OCPM |
|---|---|---|
| **Datenextraktion und Ereignisprotokolle** | Erfordert wiederholte und zeitaufwändige Datenextraktion für verschiedene Perspektiven.<br>Nutzt klassische Event Logs (XES). | Einmalige Datenextraktion, die flexible Perspektiven und Analysen ermöglicht.<br><br>Nutzt Object-Centric Event Logs (OCEL). |
| **Prozessmodellierung** | Modelliert Prozesse auf Basis einzelner Fälle (Cases), wobei jedes Ereignis genau einem Fall zugeordnet wird. Berücksichtigt nur Beziehungen zwischen Aktivitäten eines einzigen Prozesses. | Modelliert Prozesse basierend auf mehreren Objekttypen, wobei Ereignisse mehreren Objekten zugeordnet werden können. Berücksichtigt sowohl Event-zu-Objekt- als auch Objekt-zu-Objekt-Beziehungen. |
| **Prozessanalyse** | Konzentriert sich auf die Analyse von einzelnen Prozessinstanzen.<br><br>Schwierigkeiten bei der Analyse komplexer Prozesse, bei denen Ereignisse mehrere Objekte betreffen. | Geeignet für die Analyse komplexer Prozesse mit mehreren Objekten und Beziehungen, ohne Informationsverlust durch „Flattening". Betrachtet Prozesse umfassender, da mehrere Objekte und deren Interaktionen gleichzeitig analysiert werden können. |

*Abbildung 3: Vergleich CCPM und OCPM*

## FORSCHUNGSDESIGN UND DATENERHEBUNG

Der Praxisteil dieser Arbeit basiert auf mehreren Fallstudie, die den Einsatz CCPM und OCPM im Kontext des SCRM untersucht. Ziel ist es, anhand praxisnaher Szenarien die Eignung beider Ansätze zur Analyse und Optimierung von Supply-Chain-Prozessen zu bewerten. Die Fallstudie folgt dem sechsstufigen Vorgehen nach Yin 2009 und soll eine Brücke zwischen Theorie und realer Anwendung schlagen. Der Aufbau der Fallstudien gliedert sich wie folgt:

1. **Planung:** Ziel ist der Vergleich von CCPM und OCPM im Kontext des SCRMs, basierend auf konkreten Use Cases und definierten Forschungsfragen.
2. **Design:** Aufbau der Fallstudie anhand von O2C- und P2P-Prozessen, orientiert am SCOR-Modell und verschiedenen Personas, um praxisnahe Szenarien zu erzeugen.
3. **Vorbereitung:** Entwicklung realistischer Use Cases, Einarbeitung in Celonis und Erstellung analysierbarer Datensätze mit Fokus auf SCRM-relevante Ereignisse.
4. **Datensammlung:** Verwendung synthetischer (mit python generiert) und verfügbarer Datensätze (CCPM und OCPM) zur Abbildung realitätsnaher Prozessverläufe inklusive Störungen.
5. **Analyse:** Vergleich der beiden Ansätze im Process Mining Tool von Celonis anhand der definierten Use Cases, mit Fokus auf Aussagekraft, Detailtiefe und Resilienz-Kriterien (KPIs).
6. **Präsentation:** Strukturierte Darstellung der Ergebnisse mittels Tabellen und Visualisierungen, Bewertung der Ansätze im Hinblick auf ihre Eignung zur Stärkung der Supply Chain Resilienz.

Die Auswahl geeigneter Datensätze für die Case Study erweist sich als herausfordernd, da kaum öffentlich zugänglichen Daten verfügbar sind. Daher wurde im konkreten Fall für den Case-Centric Purchase-to-Pay-Prozess (P2P) ein Celonis-Demo-Datensatz[1] genutzt, der über eine kostenfreie akademische Lizenz zugänglich ist. Mithilfe des Celonis Data Generators erfolgte die Erstellung synthetischer Daten in Python, um den Datensatz zu erweitern. Zielsetzung war dabei, eine Vergleichbarkeit zwischen den beiden Ansätzen sicherzustellen. Hierzu wurden die Case-Centric und Object-Centric Datensätze so gestaltet, dass sie ähnliche Prozessverläufe und Ereignisse im O2C- und P2P-Kontext abbilden.

### Case-Centric Datensätze

Purchase-to-Pay

Der Case-Centric P2P-Datensatz wurde aus der Celonis P2P-Demo entnommen und basiert auf Daten eines SAP ECC-Systems, was sich an den charakteristischen Tabellennamen erkennen lässt. Dass dazugehörige Datenmodell ist in Abbildung 4 zu finden. Der gesamte Beschaffungsprozess, von der Bestellung über Lieferung und Rechnungsstellung bis zur Zahlung, ist darin abgebildet. Die zentrale Case-Tabelle ist EKPO (Purchasing Document Item), in der jeder Eintrag eine eindeutige Kombination aus Bestellnummer (EBELN) und Positionsnummer (EBELP) darstellt. Diese Tabelle enthält alle relevanten Fallinformationen wie Mandant, Materialnummer oder Artikelpositionen und dient als Grundlage

[1] *Verfügbar mit einer kostenlose Registrierung für die Celonis Academic License: https://signup.celonis.com/ui/sign-up/get-started*

zur Prozessdarstellung.Weitere wichtige Informationen wie Bestellkopf (EKKO), Materialstammdaten (MARA) oder Rechnungen (RBKP) sind über 1:n-Beziehungen mit der EKPO verknüpft. So lassen sich beispielsweise anhand der Materialnummer zusätzliche Merkmale wie Gewicht oder Materialgruppe zuordnen. Diese strukturierte Verknüpfung ermöglicht eine „abgeflachte" Darstellung des gesamten P2P-Prozesses auf Fallbasis. Der eigentliche Prozessablauf wird in der Event-Tabelle _CEL_P2P_ACTIVITIES dokumentiert. Diese enthält die Aktivitäten zu jedem Fall und ist über den _CASE_KEY mit der EKPO-Tabelle verbunden. Der _CASE_KEY fungiert dabei als Primärschlüssel in der Case-Tabelle und als Fremdschlüssel in der Event-Tabelle. So lässt sich jeder Bestellung eine chronologische Abfolge von Ereignissen zuweisen. Der Datensatz umfasst etwa eine Million Fälle und bildet damit eine umfangreiche Grundlage für die Analyse typischer P2P-Prozesse.

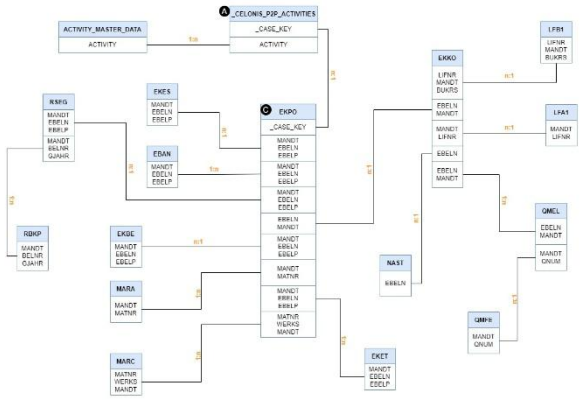



*Abbildung 4: Case-Centric P2P-Modell (erstellt mit draw.io)*

Order-to-Cash

Beim Case-Centric O2C-Datensatz handelt es sich um eine abgeflachte Version des zuvor erstellten Object-Centric O2C-Datensatzes. Mithilfe eines sogenannten Flattening-Verfahrens wurde der objektzentrierte Datensatz auf die Ebene der *Sales Order Items* reduziert, um ihn für die Analyse mit Case-Centric Process Mining nutzbar zu machen. Dabei erfolgte die Entwicklung zweier Skripte mit Python: Ein Skript generierte ein Event Log basierend auf den Sales Order Items, das andere erstellte eine zugehörige Case-Tabelle mit ergänzenden Informationen wie Produkt, Kunde oder Versandart. Die grundlegende Struktur ähnelt der des Case-Centric P2P-Datensatzes.

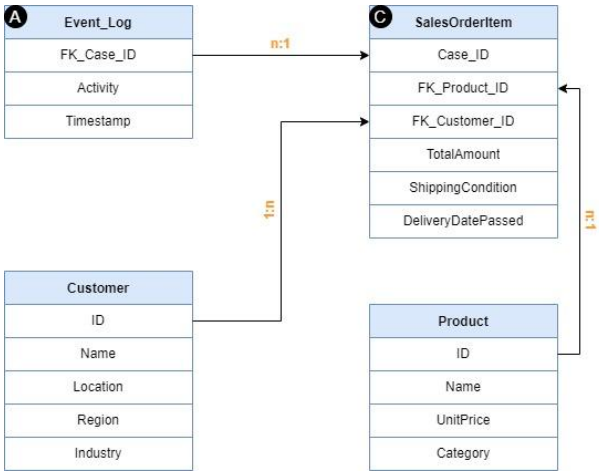



*Abbildung 5: Case-Centric O2C-Datenmodell (erstellt mit draw.io)*

Abbildung 5 zeigt das Datenmodell des O2C-Prozesses. Es existieren zwei zentrale Tabellen: eine Aktivitätentabelle („A") mit den klassischen Attributen *Case ID*, *Aktivität* und *Zeitstempel*, sowie die Case-Tabelle („C"), in der Informationen zu Produkten, Kunden und Versandbedingungen gespeichert sind. Kunden- und Produkttabellen ließen sich aus dem objektzentrierten Modell übernehmen und sind per 1:n-Beziehung mit der Case-Tabelle verbunden. Durch die Reduktion auf eine einzelne Objektebene (Sales Order Item) sind jedoch gewisse Informationsverluste entstanden. Beispielsweise stammt die Versandart ursprünglich aus den *Delivery Items*, von denen es mehrere je Sales Order Item geben kann. In der abgeflachten Version konnte daher nur ein Wert übernommen werden. Das bedeutet, dass Mehrdeutigkeiten oder Varianten wegfallen.

## Object-Centric Datenmodelle

Purchase-to-Pay

Der Object-Centric Datensatz für den P2P-Prozess wurde mithilfe des Demo Data Generators auf der Celonis Academic Plattform erstellt. Dabei ließen sich Parameter wie Anzahl der Fälle, Zeiträume sowie spezifische Events, Objekte und Attribute individuell festlegen. Verschiedene Prozessvarianten wie etwa „Change Price" oder „Change Quantity" wurden mit unterschiedlichen Wahrscheinlichkeiten definiert und teilweise kombiniert, um praxisnahe Abläufe zu simulieren.

Das Datenmodell (siehe Abbildung 6) umfasst mehrere zentrale Objekte: *Purchase Requisition Item*, *Purchase Order*, *Purchase Order Item*, *Vendor Invoice* und *Vendor Account Credit Item*. Auf das Objekt *Vendor Invoice Item* wurde zur Vereinfachung verzichtet. Die Struktur sieht unter anderem vor, dass eine *Purchase Order* mehrere *Order Items* enthält, wobei ein *Order Item* wiederum mit mehreren *Vendor Invoices* verknüpft sein kann. Zudem kann eine *Vendor Invoice* mehrere *Vendor Account Credit Items* umfassen. Eine Besonderheit ist die m:n-Beziehung zwischen *Purchase Requisition Items* und *Purchase Order Items*, die über eine separate Zwischentabelle realisiert wurde.

Alle Objekte werden im Datenmodell als eigenständige Case-Tabellen („C“) geführt, mit Ausnahme der Relationstabelle, die lediglich die Verknüpfungen zwischen den Objekten enthält. Dies unterscheidet sich grundlegend vom Case-Centric Ansatz, bei dem die Daten auf eine einzige Fallstruktur abgeflacht sind. Durch die Objektzentrierung kann man die Daten im P2P-Prozess aus verschiedenen Perspektiven betrachten. Dies bildet die Voraussetzung für eine differenzierte und flexible Analyse komplexer Prozessbeziehungen. Zusätzlich werden die E2O-Relationen verwendet, um Ereignisse mit einem oder mehreren Objekten zu verknüpfen, was eine dynamische und realitätsnahe Prozessabbildung ermöglicht.

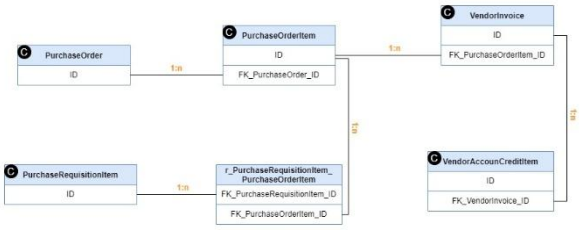



*Abbildung 6: O2O-Beziehungen im Object-Centric P2P-Prozess (erstellt mit draw.io)*

Eine grafische Darstellung aller E2O-Beziehungen im Datenmodell wäre wenig anschaulich. Stattdessen wird in Tabelle 1 beispielhaft auf einige E2O-Beziehungen eingegangen.

*Tabelle 1: E2O-Beziehungen im Object-Centric P2P-Modell*

| Events | Relation | Objekte |
|---|---|---|
| **BlockPurchaseOrderItem** | 1:1 | PurchaseOrderItem |
| **CancelPurchaseOrder** | 1:1 | PurchaseOrder |
| | 1:n | PurchaseOrderItem |
| **ChangeApprovalRequest** | 1:1 | PurchaseRequisitionItem |
| **ChangePrice** | 1:1 | PurchaseOrderItem |
| **ChangeQuantity** | 1:1 | PurchaseOrderItem |
| **ChangePurchaseOrderItem** | 1:1 | PurchaseOrderItem |
| **ClearVendorInvoice** | 1:1 | VendorInvoice |
| | 1:1 | PurchaseOrderItem |

Im Object-Centric Datenmodell lässt sich jedem Event eindeutig ein oder mehreren Objekten zuordnen. Die „Has Many“-Beziehung impliziert dabei nicht zwingend das Vorhandensein mehrerer Objekte, sondern eröffnet lediglich die Möglichkeit, dass ein Event mit mehreren Objekten verknüpft ist. Im Gegensatz zum Case-Centric Ansatz, der ausschließlich Events entlang einer gewählten Dimension bzw. Case-Tabelle abbildet, erlaubt der Object-Centric Ansatz die Modellierung von Beziehungen zwischen Events und unterschiedlichen Objekten. Dadurch können Interaktionen differenzierter analysiert und beispielsweise Prozessverzögerungen auf spezifische Objektentitäten zurückgeführt werden.

Order-to-Cash

Ein weiterer Datensatz wurde synthetisch mithilfe von Python erstellt und orientiert sich an den typischen Objekten und Events eines O2C-Prozesses. Dabei wird kein klassisches Event Log benötigt; vielmehr basieren die Daten auf Objekttabellen mit Zeitstempeln, die Änderungen erfassen und so eine OCPM-Analyse ermöglichen. Für die Modellierung wurden die Hauptobjekte *Sales Order, Sales Order Item, Delivery Item* und *Customer Invoice* berücksichtigt und um *Production Order* ergänzt, um einen Make-to-Order-Prozesses abzubilden. Zusätzlich erfolgte die Integration von *Customer* und *Product*, um weitere Informationen bereitzustellen. Das Datenmodell wird in zwei Schritten (O2O- und E2O-Beziehungen) beschrieben und durch eine Attributebene konkretisiert. Die zugrunde gelegten Annahmen lauten: (1) Eine Sales Order kann mehrere Sales Order Items enthalten, (2) ein Sales Order Item geht ggf. in mehrere Delivery Items über, (3) jede Sales Order führt zu genau einer Customer Invoice, und (4) eine Production Order ist, falls erforderlich, mit einem Sales Order Item verknüpft.

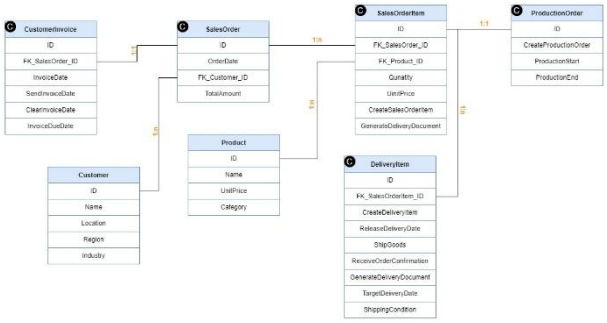



*Abbildung 7: Object-Centric O2C-Datenmodell (erstellt mit draw.io)*

Auch in diesem Datenmodell existieren mehrere Case-Tabellen, während *Customer* und *Product* nur Zusatzinformationen enthalten. Eine *Production Order* wird nur ausgelöst, wenn ein Produkt für ein *Sales Order Item* hergestellt werden muss. Zur Erhöhung der Realitätsnähe wurden realistische Zeitintervalle und Ereignisfolgen modelliert und durch Ausreißer ergänzt. Neben den Objekttabellen erfolgte die Einführung von Objektänderungstabellen, welche die Abweichungsevents wie „Change Price“ oder „Change Quantity“ enthalten. Diese folgen einer einheitlichen Struktur (*ChangeNumber, ObjectID, ChangeDate, ChangeType, OldValue, NewValue*) und stehen in einem 1:n-Verhältnis zu den Haupttabellen, sodass mehrere Änderungen je Objekt möglich sind. Änderungen wirken sich zudem auf abhängige Objekte aus, etwa wenn eine Preisänderung Anpassungen in *SalesOrderItems* und *SalesOrder* erfordert. Das Auftreten solcher Events ist probabilistisch modelliert, wobei Prozessverzögerungen nachfolgender Events berücksichtigt wurden. Durch die Einbeziehung der Objektänderungstabellen ergibt sich nun folgendes Datenmodell:

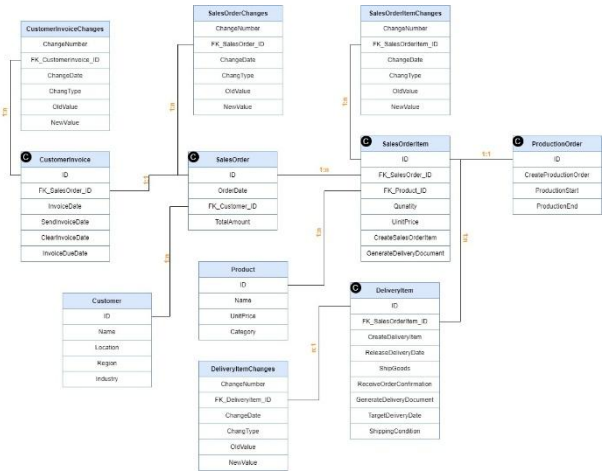



*Abbildung 8: Erweitertes Object-Centric O2C-Datenmodell (erstellt mit draw.io)*

Da eine vollständige Abbildung der E2O-Beziehungen wie beim P2P-Datenmodell zu undurchsichtig und komplex werden würde, geht die Tabelle 2 ebenfalls nur beispielhaft einige der E2E-Beziehungen ein.

*Tabelle 2: E2O-Beziehungen im Object-Centric O2C-Prozess*

| Events | Relation | Objekte |
|---|---|---|
| ApproveCreditCheck | 1:1 | SalesOrder |
| CancelOrder | 1:1 | SalesOrder |
| ChangeConfirmedDeliv-ery Date | 1:1 | SalesOrderItem |
| ChangeApprovalRe-quest | 1:1 | DeliveryItem |
| ChangePrice | 1:1 | SalesOrder |
| ChangeQuantity | 1:1 | SalesOrderItem |
| ChangeQuantity | 1:1 | SalesOrderItem |
| ClearInvoice | 1:1 | CustomerIn-voice |
| CreateCustomer Invoice | 1:1 | SalesOrder |
| | 1:n | SalesOrderItem |
| | 1:n | DeliveryItem |
| | 1:1 | CustomerIn-voice |
| CreateSalesOrder | 1:1 | SalesOrder |

Analog zum vorherigen P2P-Fall basiert der Case-Centric O2C Datensatz auf dem Object-Centric O2C Datensatz. Mithilfe eines Flattening-Verfahrens wurde der Object-Centric O2C-Datensatz auf die Dimension *Sales Order Item* heruntergebrochen. Zur Vergleichbarkeit von CCPM und OCPM erfolgte daraus die Ableitung sowohl eines Case-Centric Event Logs als auch einer Case-Tabelle. Da die Generierung probabilistisch erfolgt, sind die resultierenden Datensätze nicht replizierbar und variieren bei jeder Ausführung.

## FALLBASIERTER CCPM/OCPM VERGLEICH

Die Use Cases untersuchen, inwiefern Process-Mining-Techniken zur Stärkung der Resilienz von Supply Chains beitragen können. Resilienz wird dabei durch die Fähigkeit charakterisiert, flexibel auf Störungen zu reagieren und sich schnell zu erholen. Angesichts globaler Unsicherheiten und zunehmender Komplexität gewinnt dieser Aspekt an Bedeutung. Die Use Cases orientieren sich daher an den Dimensionen des SCOR-Modells.
Das 1996 ursprünglich entwickelte SCOR-Modell dient der Verbesserung von Supply-Chain-Prozessen und kombiniert Methodik, Diagnose- sowie Benchmarking-Tools. Es integriert Geschäftsprozesse, Kennzahlen, Best Practices und Technologien, um Kommunikation und Effizienz innerhalb der Supply Chain zu fördern. Das aktuelle Modell (Version 14) konstituiert sich aus sieben Kernprozessen (Orchestrate, Plan, Order, Source, Transform, Fulfill, Return). Es wird kontinuierlich weiterentwickelt und erlaubt eine flexible Anpassung an branchenspezifische Anforderungen durch Analysen auf verschiedenen Ebenen. Das SCOR-Modell bildet neben dem Prozessreferenzmodell die Kernbereichen Performance, Prozesse, Praktiken und Menschen ab. Für die Use Cases steht der Bereich Performance im Fokus, insbesondere die Resilienz-Attribute Zuverlässigkeit, Reaktionsfähigkeit und Agilität. Diese beschreiben die konsistente Erfüllung von Lieferungen, die Geschwindigkeit der Reaktion auf Kundenanforderungen sowie die Flexibilität gegenüber externen Veränderungen. (vgl. Association for Supply Chain Management (ASCM) 2022).
Die im nachfolgenden dargestellten Use Cases untersuchen Schwachstellen in den Resilienz-Attributen und vergleichen Case- mit Object-Centric Process Mining. Ziel ist es, Handlungsempfehlungen zur Steigerung der Supply-Chain-Resilienz auf Basis von KPIs und Personas abzuleiten. Die im durchgeführte Analyse wird mit dem Process Mining Tool von Celonis durchgeführt. Dabei ist anzumerken, dass die Analysen nur einen beispielhaften Auszug von möglichen Analysen zeigt.

### Use Case 1: Reaktionsfähigkeit und Anpassungsfähigkeit bei Störungseriegnissen

Ziel dieses Fallbeispiels ist es, die Perspektive eines *Supply Chain Managers* einzunehmen. Das heißt es wird die Fähigkeit der Supply Chain zur schnellen Reaktion auf Störungen (z. B. Lieferverzögerungen) und zur Anpassung an veränderte Nachfragebedingungen analysiert. Als Grundlage fungieren sowohl der Case-Centric als auch der Object-Centric O2C-Datensatz, um unterschiedliche Perspektiven auf Reaktionsfähigkeit und Anpassungsmechanismen zu erötern. Die Bewertung erfolgt anhand folgender spezifischer KPIs: (1) *Perfect Customer Order Fulfillment (PCOF)* als Prozentsatz termingerechter, vollständiger und fehlerfreier Bestellungen, (2) *Order Cycle Time* als Gesamtzeit vom Auftrag bis zur vollständigen Lieferung an den Kunden, (3) *Order Disruption Cycle Time* als Durchlaufzeit bei Auftreten von Störungen oder Änderungen, (4) *Rework Rate* als Anteil der Bestellungen mit Veränderungen oder Störungen im Prozess sowie (5) *Lead Time to Adapt* als Zeitspanne, die benötigt wird, um Anpassungen in der Supply Chain umzusetzen, entweder im Durchschnitt oder bezogen auf spezifische Ereignisse wie *Change Quantity*.

CCPM Analyse

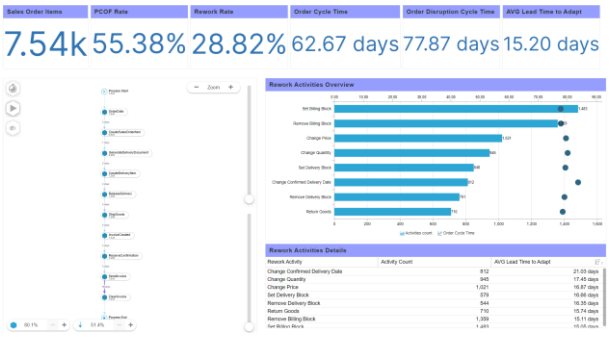



*Abbildung 9: Case-Centric KPI Dashboard*

Abbildung 9 dient als Übersicht und Einstiegspunkt für die Analyse mit CCPM. Von insgesamt 7,54 Tsd. Sales Order Items sind 2,17 Tsd. von Rework-Aktivitäten betroffen, was einer *Rework-Rate* von 28,82 % entspricht. Dieser hohe Anteil verdeutlicht die Häufigkeit von Unterbrechungen und damit eine eingeschränkte Prozessstabilität. Dies spiegelt sich auch in der PCOF-Rate wider. Mit 55,38 % werden nur knapp über die Hälfte aller Bestellungen fehlerfrei an den Kunden ausgeliefert. Die durchschnittliche *Order Cycle Time* beträgt 62,67 Tage, verlängert sich bei Störungen jedoch auf 77,87 Tage. Daraus ergibt sich eine durchschnittliche *Lead Time to Adapt* von 15,2 Tagen, die die vorhandene Flexibilität quantifiziert, zugleich aber erhebliches Verbesserungspotenzial aufzeigt. Als Hauptursache für die teilweise sehr schlechten KPI-Werten kann die Rework Rate von knapp 30% angesehen werden. Die Analyse der Rework-Aktivitäten zeigt, dass insbesondere das Setzen und Entfernen von Billing Blocks sowie Preisänderungen dominieren. Aktivitäten wie die Anpassung bestätigter Liefertermine weisen mit bis zu 21 Tagen Reaktionszeit und 83,7 Tagen Durchlaufzeit auf gravierende Engpässe hin, die das Risiko von Kettenreaktionen im Supply-Chain-Prozess erhöhen. Zur Steigerung der Anpassungsfähigkeit sind insbesondere eine effizientere Koordination zwischen den beteiligten Akteuren sowie ein frühzeitiges Monitoring potenzieller Engpässe erforderlich. Die Analyse legt nahe, dass durch Reduktion wiederkehrender Rework-Aktivitäten und gezieltes Echtzeit-Controlling Verzögerungen verringert und die Resilienz der Supply Chain gestärkt werden können.

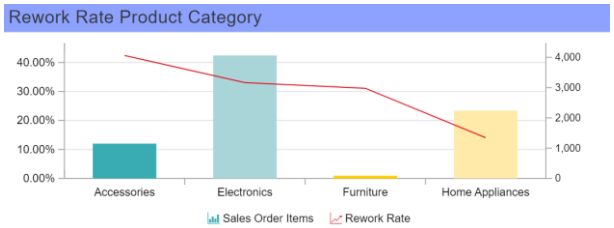



*Abbildung 10: Rework-Rate auf Produktebene*

Die Analyse auf Produktebene (Abbildung 10) zeigt deutliche Unterschiede in der Störungsanfälligkeit einzelner Kategorien. Besonders auffällig ist die Kategorie *Accessoires*, die mit einer Rework-Rate von 42,43 % den höchsten Wert aufweist. Trotz vergleichsweise geringer Bestellzahlen deutet dies auf eine überproportionale Häufigkeit von Anpassungen hin, die entweder aus höherer Komplexität in der Abwicklung oder aus verstärkten Nachfrage- und Spezifikationsänderungen resultieren. Die Kategorie *Electronics* verzeichnet mit über 4.000 Bestellungen die größte Nachfrage, ist jedoch ebenfalls durch eine hohe Rework-Rate von 33,05 % gekennzeichnet. Damit bestätigt sich, dass sowohl volumenstarke als auch kleinere Produktgruppen erheblich zur Instabilität der Prozesse beitragen können.

Die Detailanalyse in Abbildung 11 für die Produktkategorie *Accessoires*, zeigt, dass die häufigsten Rework-Ereignisse auf die Aktivitäten *Change Price* und *Change Quantity* zurückzuführen sind. Diese Ergebnisse deuten auf Optimierungspotenziale hin, die etwa durch verbesserte Abstimmungsprozesse oder den Einsatz automatisierter Anpassungsmechanismen realisiert werden könnten.

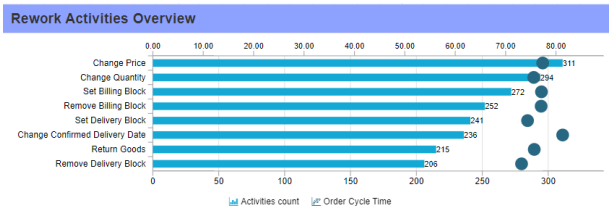



*Abbildung 11:Ursachenanalyse "Accessoires"*

Neben der Kategoriebene wird auch auf Produktebene eine hohe Störungsanfälligkeit sichtbar. Einzelne Produkte, weisen Rework-Raten von über 50 % auf, wobei P121 mit 63,89 % den Spitzenwert erreicht. Diese auffälligen Werte lassen sich auf spezifische Produktanforderungen, unregelmäßige Nachfrage oder eine erhöhte Komplexität in der Abwicklung zurückführen. Auffällig ist insbesondere die enge Verbindung zur Häufigkeit von Preis- und Mengenänderungen, die bereits in der Ursachenanalyse als wesentliche Treiber identifiziert wurden.

Rework Rate Products

| Product | Sales Order Items | Rework Rate | Order Cycle Time | PCOF Rate |
|---|---|---|---|---|
| P121 | 36 | 63.89% | 74.92 | 38.89% |
| P133 | 400 | 56.00% | 75.22 | 39.00% |
| P115 | 33 | 54.55% | 71.45 | 48.48% |
| P119 | 379 | 53.56% | 74.81 | 36.94% |
| P113 | 163 | 52.76% | 70.53 | 41.10% |
| P127 | 152 | 52.63% | 70.36 | 40.79% |
| P109 | 141 | 52.48% | 68.12 | 39.72% |
| P108 | 158 | 51.90% | 69.93 | 41.14% |

*Abbildung 12:Rework Rate Produktebene*

Damit wird deutlich, dass nicht nur ganze Kategorien, sondern auch einzelne Produkte maßgeblich zur Instabilität des Prozesses beitragen können. Dies unterstreicht die Notwendigkeit einer differenzierten Analyse- und Steuerungsebene, um gezielt Produktgruppen mit besonders hohem Rework-Risiko zu identifizieren und zu optimieren. Abbildung 12 verdeutlicht einen klaren Zusammenhang zwischen Rework-Rate, Order Cycle Time und PCOF Rate: Produkte mit hoher Änderungsanfälligkeit weisen tendenziell auch längere Durchlaufzeiten und schlechtere PCOF Raten auf. Die Produkte P121, P133, P115 und P119 verzeichnen Rework-Raten von über 50 %. Entsprechend liegen ihre Order Cycle Times mit rund 71 bis 75 Tagen deutlich über dem Durchschnitt, während die

PCOF-Werte mit etwa 36 bis 48 % vergleichsweise niedrig ausfallen. Im Vergleich dazu liegen bei P101, P110, P122 und P120 die Rework-Raten unter 20%, während ihre Order Cycle Times (ca. 55 bis 61 Tage) deutlich kürzer und die PCOF-Raten deutlich höher sind (ca. 63 bis 75%). Diese Produkte sind weniger von Anpassungen betroffen, was auf eine stabilere Nachfrage und eine einfachere Prozesshandhabung schließen lässt. Sie können somit als Best-Practice-Beispiele dienen, um Verbesserungspotenziale für störungsanfälligere Produkte abzuleiten. Demgegenüber zeigen komplexere Produkte mit hohen Rework-Raten, längeren Durchlaufzeiten und geringer PCOF-Rate, dass kundenspezifische Anpassungen oder besondere Anforderungen zu erhöhtem Änderungsbedarf führen. Für diese Produkte erscheint eine vertiefte Ursachenanalyse sinnvoll, etwa im Hinblick auf Lagerhaltung, Lieferzeiten oder Lieferantenqualität.

Rework Rate Products

| Product | Sales Order Items | Rework Rate | Order Cycle Time | PCOF Rate |
|---|---|---|---|---|
| P121 | 36 | 63.89% | 74.92 | 38.89% |
| P133 | 400 | 56.00% | 75.22 | 39.00% |
| P115 | 33 | 54.55% | 71.45 | 48.48% |
| P119 | 379 | 53.56% | 74.81 | 36.94% |

Rework Rate Products

| Product | Sales Order Items | Rework Rate | Order Cycle Time | PCOF Rate |
|---|---|---|---|---|
| P101 | 35 | 5.71% | 59.06 | 74.29% |
| P110 | 45 | 11.11% | 55.82 | 75.56% |
| P122 | 748 | 13.10% | 61.41 | 63.77% |
| P120 | 129 | 13.18% | 55.74 | 65.89% |

*Abbildung 13: Prduktvergleich*

Insgesamt liefert die Analyse wertvolle Ansatzpunkte für Prozessoptimierungen im Order-to-Cash-Prozess. Eine gezielte Reduktion der Rework-Rate und eine effizientere Anpassung an Änderungen können wesentlich zur Steigerung von Effizienz und Stabilität in der Supply Chain beitragen.

OCPM Analyse

Die Ergebnisse auf Ebene der Sales Order Items unterscheiden sich von der Case-Centric Analyse und verdeutlichen damit einen ersten Vorteil des Object-Centric Process Mining (OCPM). Das Modell erlaubt nicht nur die Abbildung der CCPM-Ergebnisse auf Item-Ebene, sondern auch die Erweiterung der Analyse auf sämtliche Objekte im Prozess und deren Relationen. Abbildung 14 zeigt, dass insbesondere *Sales Orders*, *Delivery Items* und *Customer Invoices* mit Rework-Raten von über 30 % am häufigsten von Änderungen betroffen sind. Auffällig ist zudem, dass die *Perfect Customer Order Fulfillment Rate (PCOF)* auf Sales-Order-Ebene mit 33,37 % am niedrigsten liegt, was darauf hindeutet, dass Fehler, die zur Nichterfüllung dieser KPI führen, primär hier entstehen. Während auf Item-Ebene die Rework-Rate rund zehn Prozent niedriger liegt und die *Order Disruption Cycle Time* um 22 Tage reduziert wird, bleibt die PCOF im Vergleich zur CCPM-Analyse unverändert. Gleichzeitig verlängert sich jedoch die *Lead Time to Adapt* um sechs Tage, was auf zusätzliche Komplexität bei Anpassungen über mehrere Objekte hinweg hinweist. Besonders gravierend sind die Verzögerungen durch Änderungen auf Sales-Order-Ebene: Mit einer durchschnittlichen Disruption Cycle Time von 73,07 Tagen und einer Lead Time to Adapt von 33,03 Tagen lösen sie Kettenreaktionen aus, die den gesamten Prozess verlangsamen.

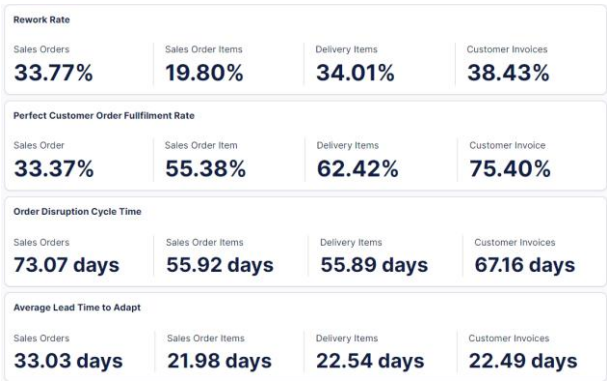



*Abbildung 14:KPIs auf Objektlevel*

Abbildung 15 verdeutlicht, dass Änderungen auf Sales-Order-Ebene unmittelbare Auswirkungen auf nachgelagerte Objektebenen haben. So führt das Event *Change Price* in mehr als 50 % der Fälle zu einer Anpassung der Lieferbestätigung auf Delivery-Item-Ebene. Damit erhöhen Preisänderungen nicht nur die Rework-Raten innerhalb der Sales Orders, sondern verstärken gleichzeitig die Störungsanfälligkeit in weiteren Prozessschritten.


*Abbildung 15:Kettenreaktion durch Aktivität "Change Price"*

Mit Hilfe des *Process Adherence Managers* von Celonis lässt sich zudem die Prozesskonformität untersuchen. Hier zeigt sich, dass insbesondere die Events *Change Price* und *Return Goods* die Order Cycle Time erheblich verlängern, um durchschnittlich 38 bzw. 54 Tage. Die Ursachenanalysen identifizieren vor allem die Produkte *P123* und *P117* als Auslöser von Preisänderungen, während *P117* zugleich am häufigsten mit Rücksendungen verbunden ist. Dieses Produkt erscheint somit als kritischer Treiber mehrerer Störungen und bietet einen zentralen Ansatzpunkt für vertiefte Analysen. Diese Erkenntnisse konnten beispielwesie nicht aus der Case-Centric Analyse gezogen werden.

Diese Ergebnisse heben das Potenzial von OCPM hervor: Im Gegensatz zur Case-zentrierten Betrachtung können hier Kettenreaktionen über verschiedene Objekte hinweg transparent gemacht und kritische Produkte mit überdurchschnittlicher Störungsanfälligkeit identifiziert werden.

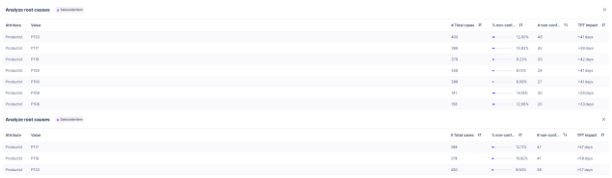


*Abbildung 16: Ursachenanalyse "Change Price" und "Return Goods"*

**Use Case 2: Lieferantenauswahl und Performance Monitoring**

Dieser Use Case betrachtet die Perspektive eines *Procurement Managers*, dessen Ziel es ist, die Lieferantenauswahl zu optimieren und die Lieferantenperformance kontinuierlich zu überwachen, um eine stabile und zuverlässige Supply Chain sicherzustellen. Dabei stehen die Pünktlichkeit und Vollständigkeit von Lieferungen, die Minimierung von Störungen sowie der Aufbau langfristiger Beziehungen zu leistungsstarken Lieferanten im Mittelpunkt. Grundlage bilden die Case- und Object-Centric P2P-Datensätze, die sowohl die Prozesssicht auf einzelne Bestellungen als auch die Beziehungen zwischen Objekten wie Bestellungen und Rechnungen abbilden. Die Bewertung erfolgt anhand folgender KPIs: (1) *Perfect Supplier Fulfillment Rate* als Anteil termingerechter, vollständiger und fehlerfreier Lieferungen, (2) *Purchase Order Cycle Time* als Zeitspanne von der Bestellung bis zur Rechnungsbegleichung, (3) *Purchase Order Disruption Cycle Time* als Durchlaufzeit bei störungsbehafteten Lieferungen, (4) *Purchase Order Rework Rate* als Anteil der von Störungen oder Anpassungen betroffenen Bestellungen sowie (5) *Supplier Agility Time* als Maß für die Geschwindigkeit, mit der Lieferanten auf Änderungen reagieren und ihre Lieferungen anpassen können.

CCPM Analyse

Der zweite Use Case adressiert die Bewertung von Zuverlässigkeit und Flexibilität der Lieferanten durch den Procurement Manager, mit dem Ziel, fundierte Entscheidungen zur Lieferantenauswahl und langfristigen Zusammenarbeit zu ermöglichen. Vorab ist anzumerken, dass kein quantitiver Vergleich zu OCPM gezogen werden kann, da sich die Datengrundlage für CCPM und OCPM voneinadner unterscheidet.

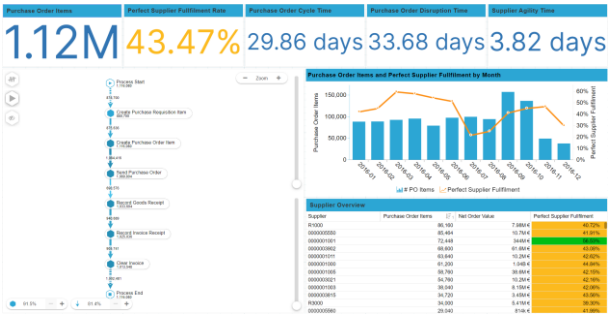



*Abbildung 17: Case-Centric KPI Dashboard*

Abbildung 17 zeigt, dass die *Perfect Supplier Fulfillment (PSF)-Rate* lediglich 43,47 % beträgt. Damit erfüllen fast 60 % der Lieferungen die Anforderungen nicht vollständig, was erhebliche Risiken für die Stabilität der Supply Chain mit sich bringt. Die *Purchase Order Disruption Time* liegt mit durchschnittlich 33,68 Tagen nur geringfügig über der regulären *Purchase Order Cycle Time*, während die *Supplier Agility Time* mit 3,82 Tagen eine schnelle Anpassungsfähigkeit signalisiert. Gleichwohl offenbart das Dashboard teils erhebliche Schwankungen in der Lieferantenperformance, die auf saisonale Einflüsse oder externe Störungen (z.B. Naturereignisse) zurückzuführen sein könnten. Über die Supplier Overview-Tabelle werden zudem Lieferanten mit besonders niedriger PSF-Rate sichtbar. Auffällig ist, dass gerade die volumenstärksten Lieferanten mit den höchsten Bestellmengen und Umsätzen durchgängig Werte unterhalb von 50 % aufweisen. Dies deutet auf ein generelles Leistungsdefizit hin, das über gezielte Maßnahmen adressiert werden muss, etwa durch die Diversifizierung des Lieferantenportfolios oder die Ergänzung zusätzlicher Partner in kritischen Zeiträumen. Im nächsten Schritt werden daher die Rework-Raten im Beschaffungsprozess analysiert, um spezifische Schwachstellen und Handlungsfelder zu identifizieren.

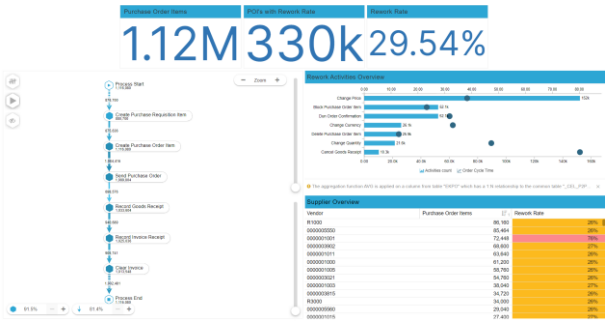



*Abbildung 18: Analyse der Rework Aktivitäten*

Die Analyse verdeutlicht eine Rework-Rate von 29,53%, was rund 330.000 *Purchase Order Items* entspricht (Abbildung 18). Mit 152.000 Vorkommen ist das Event *Change Price* für nahezu die Hälfte aller Rework-Aktivitäten verantwortlich. Während dieses Ereignis eine durchschnittliche Order Cycle Time von etwa 40 Tagen verursacht, führen *Change Quantity* und *Cancel Goods Receipt* zu erheblich längeren Durchlaufzeiten von rund 50 bzw. 80 Tagen. Damit werden erste Ansatzpunkte für vertiefte Analysen durch den Procurement Manager sichtbar.

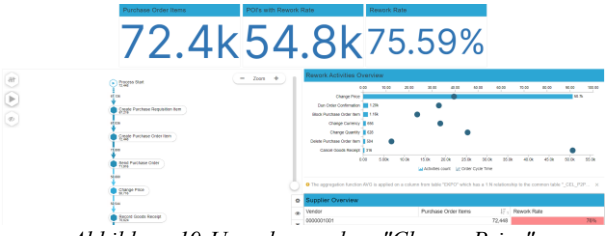
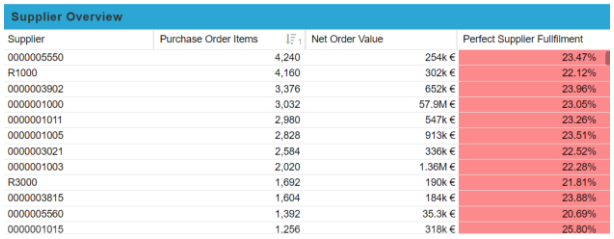



*Abbildung 19:Ursachenanalyse "Change Price"*

Besonders auffällig ist ein Lieferant mit einer Rework-Rate von 76%, der allein für etwa ein Drittel aller *Change Price*-Events verantwortlich ist (Abbildung 19). Gleichwohl liegt seine *PSF-Rate* bei über 50 %, sodass hier kein unmittelbarer Zusammenhang zwischen hoher Rework-Rate und niedriger PSF-Rate nachweisbar ist. Anders verhält es sich beim Event *Block Purchase Order Item*: In diesen Fällen sinkt die PSF-Rate auf etwa 23 %, was auf eine klare Kausalität zwischen Rework-Aktivitäten und Lieferantenperformance hinweist. Die Analyse der Lieferanten nach diesem Kriterium ermöglicht es, gezielt diejenigen mit den meisten *Purchase Order Items* und

damit dem größten Einfluss auf den Prozess zu identifizieren. Damit zeigt sich, dass CCPM ein wirksames Instrument ist, um schnell Transparenz über Prozessschwächen und deren Verursacher zu gewinnen. Dies versetzt den Procurement Manager in die Lage, gezielt auf kritische Lieferanten einzuwirken. Im nächsten Schritt erfolgt eine OCPM-Analyse des P2P-Prozesses, um zusätzliche Erkenntnisse zu erschließen (Abbildung 20).



Supplier Overview

| Supplier | Purchase Order Items | Net Order Value | Perfect Supplier Fulfilment |
|---|---|---|---|
| 0000005550 | 4,240 | 254k € | 23.47% |
| R1000 | 4,160 | 302k € | 22.12% |
| 0000003902 | 3,376 | 652k € | 23.96% |
| 0000001000 | 3,032 | 57.9M € | 23.05% |
| 0000001011 | 2,980 | 547k € | 23.26% |
| 0000001005 | 2,828 | 913k € | 23.51% |
| 0000003021 | 2,584 | 336k € | 22.52% |
| 0000001003 | 2,020 | 1.36M € | 22.28% |
| R3000 | 1,692 | 190k € | 21.81% |
| 0000003815 | 1,604 | 184k € | 23.88% |
| 0000005560 | 1,392 | 35.3k € | 20.69% |
| 0000001015 | 1,256 | 318k € | 25.80% |

*Abbildung 20: Lieferantenanalyse "Block Purchase Order Item"*

OCPM Analyse

Im Gegensatz zur Case-Centric Analyse liegt der Fokus im Object-Centric Ansatz auf den Wechselwirkungen zwischen den Prozessobjekten. Abbildung 21 zeigt, dass insbesondere *Purchase Order Items* (43,06 %) und *Vendor Invoices* (26,92 %) von hohen Rework Raten betroffen sind. Dies deutet auf Lieferantenprobleme oder interne Prozessschwierigkeiten hin, etwa Abweichungen zwischen gelieferten Waren und abgerechneten Beträgen. Entsprechend sollte der Procurement Manager prüfen, ob bestimmte Lieferanten regelmäßig Korrekturen verursachen. Die *Disruption Cycle Time* beträgt bei *Purchase Order Items* 39,10 Tage und bei *Vendor Invoices* 31,08 Tage, während die *Agility Time* bei Vendor Invoices mit 18,22 Tagen auffallend hoch ist. Dies unterstreicht, dass gerade Rechnungsprozesse durch Störungen erheblich verzögert werden und sowohl den Cashflow als auch Lieferantenbeziehungen belasten können (Abbildung 21)

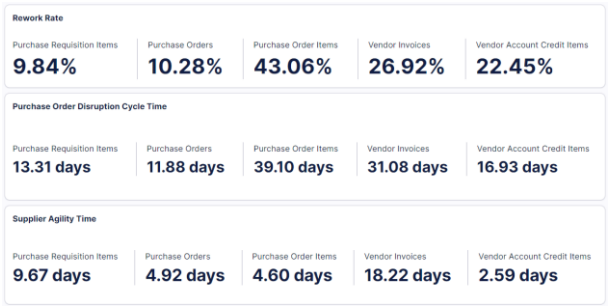



*Abbildung 21: KPIs auf Objektebene*

Die hohe Rework-Rate der *Purchase Order Items* impliziert zudem, dass nachgelagerte Objekte wie *Vendor Invoices* systematisch von Verzögerungen betroffen sind. Zur Überprüfung wurde eine Filterung nach Mengenänderungen (*Change Quantity*) auf der Ebene der *Purchase Order Items* vorgenommen (Abbildung 22).


| Supplier | Vendor Invoices | Rework Rate Vendor I... | PSF-Rate |
|---|---|---|---|
| V009 | 11 | 69.60% | 27.27% |
| V006 | 47 | 65.70% | 29.79% |
| V005 | 14 | 65.19% | 21.43% |
| V012 | 36 | 61.90% | 36.11% |
| V018 | 56 | 54.60% | 33.93% |
| V016 | 20 | 54.55% | 30.00% |
| V017 | 49 | 51.88% | 40.82% |
| V001 | 29 | 48.16% | 44.83% |

*Abbildung 22:Ursachenanalyse Lieferanten*

Die Analyse zeigt, dass sich die Rework-Rate der *Vendor Invoices* dadurch auf 40,85 % erhöht. Auffällig ist, dass vor allem die Lieferanten *V009*, *V006* und *V005* betroffen sind, deren PSF-Raten jeweils unter 30 % liegen. Dies weist darauf hin, dass Mengenänderungen im weiteren Verlauf häufig zu Rechnungskorrekturen führen.
Die Analyse nachgelagerter Prozesse (Abbildung 23) verdeutlicht darüber hinaus eine Kettenreaktion: Auf Ebene der *Vendor Account Credit Items* steigt die Rework-Rate auf 44,05%, während sich gleichzeitig die *Disruption Cycle Time* verdoppelt und die *Supplier Agility Time* um das Siebenfache verlängert. Damit wird klar, dass Störungen auf der *Purchase-Order-Item*-Ebene Instabilitäten in mehreren nachfolgenden Prozessschritten auslösen können. Die Erkenntnisse legen nahe, dass eine Reduktion von Mengenänderungen bereits auf Objektebene wesentlich zur Stabilisierung der gesamten Prozesskette beitragen und die Effizienz der Supply Chain nachhaltig verbessern kann.

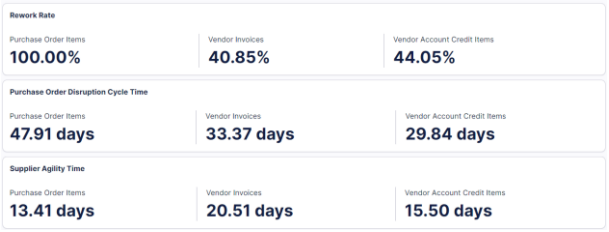



*Abbildung 23: Kettenreaktion durch "Change Quantity"*

## DISKUSSION UND FAZIT

Die vergleichende Analyse von CCPM und OCPM) im Kontext des SCRM zeigt deutliche Unterschiede in ihrer analytischen Tiefe und Aussagekraft. Während CCPM eine aggregierte, fallzentrierte Sicht bietet, ermöglicht OCPM eine differenzierte Betrachtung der Prozesskomplexität über mehrere Objekte hinweg. Auf Basis der analysierten Use Cases konnten die jeweiligen Stärken und Limitationen der Ansätze klar herausgearbeitet werden.
CCPM erwies sich insbesondere als wertvoll für eine schnelle und übergeordnete Einschätzung der Prozessleistung. Die Analysen lieferten eine konsolidierte Sicht auf zentrale Kennzahlen wie Order Cycle Time, Rework Rate oder Lead Time to Adapt, wodurch Schwachstellen effizient identifiziert und priorisiert werden konnten. Im ersten Use Case wurde deutlich, dass CCPM auf einen

Blick zentrale Anpassungsprobleme im Order-to-Cash-Prozess offenlegt, etwa durch eine erhöhte Rework-Rate von 28,82 Prozent und eine deutlich verlängerte Order Disruption Cycle Time. Auch im zweiten Use Case, der die Bewertung der Lieferantenperformance fokussiert, bot CCPM eine prägnante Einschätzung der Supplier Reliability anhand der Perfect Supplier Fulfillment Rate. Solche Ergebnisse sind besonders für ein erstes Screening und für strategische Entscheidungen nützlich.

In allen Use Cases wurde jedoch deutlich, dass CCPM durch die fallzentrierte Aggregation Informationsverluste erzeugt. Aktivitäten verschiedener Objekte werden auf eine einzige Fallebene verdichtet, wodurch Ursachen und Wechselwirkungen zwischen Objekten verdeckt bleiben. Dies zeigt sich beispielsweise im ersten Use Case in der Abweichung der Rework-Raten. Während die CCPM-Analyse eine Rework-Rate von knapp 30 Prozent auswies, lag diese auf der Ebene einzelner Objekte im OCPM zum Teil deutlich niedriger, etwa bei 19,8 Prozent für Sales Order Items. Diese Differenzen entstehen durch die Flattening-Logik von CCPM, die sämtliche Objektaktivitäten in einem Log bündelt und dadurch die Aussagekraft der Kennzahlen verändert. OCPM ermöglicht dagegen eine mehrdimensionale und beziehungsorientierte Sichtweise auf Prozesse. Die Betrachtung mehrerer Objekte und ihrer Interaktionen offenbart, wie lokale Änderungen Kettenreaktionen im Gesamtprozess auslösen können. Im ersten Use Case zeigte sich, dass Preisänderungen auf Sales-Order-Ebene in mehr als 50 Prozent der Fälle Änderungen auf Delivery-Item-Ebene nach sich ziehen. Ebenso wurde deutlich, dass hohe Rework-Raten bei Purchase Order Items systematisch zu Verzögerungen bei Vendor Invoices und Vendor Account Credit Items führen, wie im zweiten Use Case zu sehen war. Diese Ketteneffekte bleiben in einer rein fallzentrierten Betrachtung verborgen. Ein weiterer Mehrwert von OCPM liegt in der gezielten Identifikation kritischer Objekte, Produkte und Lieferanten. Durch die differenzierte KPI-Betrachtung lassen sich spezifische Engpässe lokalisieren, etwa Produkte mit außergewöhnlich hohen Rework-Raten und verlängerten Durchlaufzeiten oder Lieferanten mit überdurchschnittlicher Änderungsanfälligkeit. Dadurch entstehen konkrete Ansatzpunkte für operative Maßnahmen wie die Verbesserung von Abstimmungsprozessen, die Anpassung von Lieferantenstrategien oder die gezielte Automatisierung besonders kritischer Prozessschritte. Die Ergebnisse zeigen außerdem, dass OCPM nicht nur als Ergänzung, sondern in vielen Bereichen auch als Ersatz für CCPM dienen kann. Da OCPM die gleichen Kennzahlen wie CCPM generieren kann, jedoch auf granularerer Ebene und ohne Aggregationsverluste, ermöglicht es eine umfassendere Analyse mit höherer Aussagekraft. Der initiale Implementierungsaufwand für OCPM ist zwar höher, da die Einrichtung eines objektzentrierten Datenmodells eine tiefere Systemintegration und spezielles Fachwissen erfordert. Dieser Aufwand ist jedoch einmalig. Neue Perspektiven können anschließend ohne erneute Logextraktion erzeugt werden, was langfristig einen deutlichen Effizienzgewinn bietet.

Zusammenfassend verdeutlichen die Use Cases, dass CCPM seine Stärke in der schnellen und holistischen Problemlokalisierung hat, während OCPM tiefere Einblicke in Ursachen, Wechselwirkungen und Optimierungspotenziale ermöglicht. Gerade für die Stärkung der Supply Chain Resilience im Hinblick auf Zuverlässigkeit, Reaktionsfähigkeit und Agilität liefert OCPM entscheidende Mehrwerte, da es eine präzisere Steuerung und Anpassung auf mehreren Objektebenen erlaubt.

Die Analyse hat gezeigt, dass CCPM und OCPM im Kontext des Supply Chain Resilience Managements unterschiedliche, sich ergänzende Stärken aufweisen. CCPM bietet einen schnellen und übersichtlichen Zugang zur Prozesslandschaft und ermöglicht die effiziente Identifikation zentraler Schwachstellen. OCPM erweitert diese Perspektive durch eine differenzierte Betrachtung mehrerer Objekte und ihrer Wechselwirkungen und liefert dadurch tiefere Einblicke in komplexe Prozesszusammenhänge. Die in den Use Cases gewonnenen Ergebnisse verdeutlichen, dass OCPM in der Lage ist, Kettenreaktionen und objektspezifische Ursachen von Prozessinstabilitäten sichtbar zu machen, die in CCPM-Analysen verborgen bleiben. Dies schafft eine fundierte Grundlage, um gezielte Maßnahmen zur Steigerung von Zuverlässigkeit, Reaktionsfähigkeit und Agilität innerhalb der Supply Chain abzuleiten. Durch die flexible Perspektivenauswahl lassen sich verschiedene Objekte und Prozessebenen ohne erneute Datenaufbereitung analysieren, was langfristig die Effizienz und Aussagekraft von Process-Mining-Analysen erhöht.

Gleichzeitig weist die Studie methodische Grenzen auf. Die Analysen basieren ausschließlich auf simulierten Daten, wodurch die Übertragbarkeit der Ergebnisse auf reale Unternehmensprozesse eingeschränkt ist. Zudem wurde der Einsatz der Methoden in einem kontrollierten Analyseumfeld durchgeführt, was mögliche Herausforderungen bei der Integration in bestehende IT-Systeme, Datenqualität und Change-Management-Prozesse nur eingeschränkt berücksichtigt. Zukünftige Forschung sollte daher die Anwendung von CCPM und OCPM mit realen Unternehmensdaten untersuchen, um die praktischen Potenziale und Grenzen der Ansätze unter realen Bedingungen zu validieren. Darüber hinaus bietet die Integration von KI-Methoden und prädiktiven Modellen vielversprechende Möglichkeiten, Resilienzindikatoren proaktiv zu steuern und Prozesse adaptiver zu gestalten.

## LITERATUR


Adams, Jan Niklas; Park, Gyunam; Levich, Sergej; Schuster, Daniel; van der Aalst, Wil M. P. (2022): A Framework for Extracting and Encoding Features from Object-Centric Event Data. In: Javier Troya, Brahim Medjahed, Mario Piattini, Lina Yao, Pablo Fernández und Antonio Ruiz-Cortés (Hg.): Service-Oriented Computing, Bd. 13740. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), S. 36–53.

Adams, Jan Niklas; van der Aalst, Wil M.P. (2021): Precision and Fitness in Object-Centric Process Mining. In: 2021 3rd International Conference on Process

Mining (ICPM). 2021 3rd International Conference on Process Mining (ICPM). Eindhoven, Netherlands, 31.10.2021 - 04.11.2021: IEEE, S. 128–135.

Association for Supply Chain Management (ASCM) (2022): Supply Chain Operations Reference Model. SCOR Digital Standard, SCOR Version 14.0. Online verfügbar unter https://scor.ascm.org/api/files/24?v=1730907429253.

Berti, Alessandro; Jessen, Urszula; Park, Gyunam; Rafiei, Majid; van der Aalst, Wil M. P. (2023a): Analyzing interconnected processes: using Object-Centric process mining to analyze procurement processes. In: Int J Data Sci Anal. DOI: 10.1007/s41060-023-00427-3.

Berti, Alessandro; Koren, Istvan; Adams, Jan Niklas; Park, Gyunam; Knopp, Benedikt; Graves, Nina et al. (2023b): OCEL (Object-Centric Event Log) 2.0 Speci-fication. Online verfügbar unter www.ocel-standard.org/2.0/ocel20_specification.pdf.

Berti, Alessandro; Montali, Marco; van der Aalst, Wil M. P. (2023c): Advancements and Challenges in Object-Centric Process Mining: A Systematic Literature Review.

Carmona, Josep; van Dongen, Boudewijn; Weidlich, Matthias (2022): Conformance Checking: Foundations, Milestones and Challenges. In: Wil M. P. van der Aalst und Josep Carmona (Hg.): Process Mining Handbook, Bd. 448. Cham: Springer International Publishing (Lecture Notes in Business Information Processing), S. 155–190.

Eggers, Julia; Hein, Andreas; Böhm, Markus; Krcmar, Helmut (2021): No Longer Out of Sight, No Longer Out of Mind? How Organizations Engage with Process Mining-Induced Transparency to Achieve Increased Process Awareness. In: Bus Inf Syst Eng 63 (5), S. 491–510. DOI: 10.1007/s12599-021-00715-x.

Galanti, Riccardo; Leoni, Massimiliano de; Navarin, Nicolò; Marazzi, Alan (2023): Object-Centric process predictive analytics. In: Expert Systems with Applica-tions 213, S. 119173. DOI: 10.1016/j.eswa.2022.119173.

Ghahfarokhi, Anahita Farhang; Park, Gyunam; Berti, Alessandro; van der Aalst, Wil M. P. (2021): OCEL: A Standard for Object-Centric Event Logs. In: Ladjel Bellatreche, Marlon Dumas, Panagiotis Karras, Raimundas Matulevičius, Ahmed Awad, Matthias Weidlich et al. (Hg.): New Trends in Database and Information Systems, Bd. 1450. Cham: Springer International Publishing (Communications in Computer and Information Science), S. 169–175.

Kerremans, Marc; Sugden, David; Duffy, Nick (2024): Magic Quadrant for Process Mining Plat-forms. Hg. v. Gartner. Gartner Research. Online verfügbar unter https://www.gartner.com/en/documents/5391263.

Leoni, Massimiliano de (2022): Foundations of Pro-cess Enhancement. In: Wil M. P. van der Aalst und Josep Carmona (Hg.): Process Mining Handbook, Bd. 448. Cham: Springer International Publishing (Lecture Notes in Business Information Processing), S. 243–273.

Lund, Susan; Manyika, James; Woetzel, Jonathan; Barriball, Ed; Krishnan, Mekala; Alicke, Knut et al. (2020): Risk, resilience, and rebalancing in global value chains (Full Report). McKinsey Global Institute. Online verfügbar unter https://www.mckinsey.com/business-functions/operations/our-insights/risk-resilience-and-rebalancing-in-global-value-chains.

Park, Gyunam; Adams, Jan Niklas; van der Aalst, Wil M. P. (2022): OPerA: Object-Centric Performance Analysis. In: Jolita Ralyté, Sharma Chakravarthy, Mukesh Mohania, Manfred A. Jeusfeld und Kama-lakar Karlapalem (Hg.): Conceptual Modeling, Bd. 13607. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 281–292.

Peters, Ralf; Nauroth, Markus (2019): Process-Mining. Wiesbaden: Springer Fachmedien Wiesbaden.

Reinkemeyer, Lars (2022): Status and Future of Pro-cess Mining: From Process Discovery to Process Execution. In: Wil M. P. van der Aalst und Josep Carmona (Hg.): Process Mining Handbook, Bd. 448. Cham: Springer International Publishing (Lecture Notes in Business Information Processing), S. 405–415.

Suriadi, S.; Andrews, R.; Hofstede, A.H.M. ter; Wynn, M. T. (2017): Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. In: Information Systems 64, S. 132–150. DOI: 10.1016/j.is.2016.07.011.

van der Aalst, Wil (2016): Data Science in Action. In: Wil van der Aalst (Hg.): Process Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 3–23.

van der Aalst, Wil (2019): Object-Centric Process Mining: Dealing with Divergence and Convergence in Event Data. In: Peter Csaba Ölveczky und Gwen Salaün (Hg.): Software Engineering and Formal Methods, Bd. 11724. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 3–25. Online verfügbar unter https://doi.org/10.1007/978-3-030-30446-1_1.

van der Aalst, Wil; Adriansyah, Arya; Medeiros, Ana Karla Alves de; Arcieri, Franco; Baier, Thomas; Blickle, Tobias et al. (2012): Process Mining Manifesto. In: Florian Daniel, Kamel Barkaoui und Schahram Dustdar (Hg.): Business Process Management Workshops, Bd. 99. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Business Information Processing), S. 169–194.

van der Aalst, Wil M. P. (2022): Process Mining: A 360 Degree Overview. In: Wil M. P. van der Aalst und Josep van der Carmona (Hg.): Process Mining Handbook, Bd. 448. Cham: Springer International Publishing (Lecture Notes in Business Information Processing), S. 3–34.

van der Aalst, Wil M. P. (2023a): Object-Centric Process Mining: An Introduction. In: Antonio Cerone (Hg.): Formal Methods for an Informal World, Bd. 13490. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 73–105.

van der Aalst, Wil M. P. (2023b): Object-Centric Process Mining: Unraveling the Fabric of Real Processes. In:

Mathematics 11 (12), S. 2691. DOI: 10.3390/math11122691.
van der Aalst, Wil M. P. (2024): Academic Perspective: How Object-Centric Process Mining Helps to Unleash Predictive and Generative AI. In: Lars Reinkemeyer (Hg.): Process Intelligence in Action. Cham: Springer Nature Switzerland, S. 219–232.
Yin, Robert K. (2009): Case study research. Design and methods. 4. ed. Los Angeles, Calif.: Sage (Applied social research methods series, 5).

**KONTAKT**

David Messal hat seinen Bachelor in Wirtschaftsinformatik und Master in Information Systems an der Hochschule Pforzheim erworben und arbeitet als IT-Consultant im Bereich Technology & Process Risk.

Frank Morelli ist Professor für Wirtschaftsinformatik an der Hochschule Pforzheim. Er forscht im Bereich Business Intelligence und hat die Hochschule Pforzheim als Celonis Academic Center of Excellence etabliert..

Frank Schätter ist Professor für Supply Chain Process Management an der Hochschule Pforzheim. Er promovierte im Bereich Supply Chain Risk Management und forscht zu Supply Chain-Modellierung und -Optimierung.

Florian Haas ist Professor für Beschaffungs- und Lieferkettenmanagement an der Hochschule Pforzheim. Er verfügt über mehr als 15 Jahre praktische Erfahrung in verschiedenen Führungspositionen im Bereich Logistik bei der Robert Bosch GmbH.

# A POTENTIAL WEAKNESS OF THE GENETIC ALGORITHM IN MULTIMODAL OPTIMIZATIONS

Roman Knobloch and Jaroslav Mlynek

Technical University of Liberec, Department of Mathematics, e-mail: roman.knobloch@tul.cz, jaroslav.mlynek@tul.cz

**KEYWORDS**

Optimization, genetic algorithm, generation, crossover, mutation, selection, multimodal fitness function, random search.

**ABSTRACT**

The genetic algorithms are robust and stable evolutionary computation techniques in the sense that they mostly provide a solution. This fact is undisputedly an important positive feature, particularly in the situation when no feasible solution is available. On the other hand, this robustness and the fact that living nature utilizes analogous mechanisms for its development and propagation sometimes lead to unrealistic expectations and exaggerated faith in their capabilities. In particular, when they are used as function optimizers they sometimes tend to demonstrate also their weak points. In this article, we demonstrate the limits of the genetic algorithm in a multimodal model task. The task consists in the identification of a de Bruijn sequence of a given length between general binary sequences. Generally, the multimodality of the fitness function can lead to the stagnation of the genetic algorithm well below the desirable fitness function values which hinders the genetic algorithm from finding the optimized solution of the task under consideration.

## INTRODUCTION

During the last decades, evolutionary computation techniques have become an important tool for solving diverse and wide-ranging tasks in science, industry, technology, and economics - see for instance Mitsuo and Runwei (2000), Goldberg (2000). The steady and fast progress in the power of computational facilities contributed to the wider use of these methods. Parallel computing technology and processes made these procedures even more of a practical interest.

A group of genetic algorithms is probably the most known representative of evolutionary computation techniques. These algorithms use the principles of natural evolution as postulated by Charles Darwin and Johann Gregor Mendel in the 19th century. They are stochastic methods that incorporate the endeavour for survival and genetic inheritance. These algorithms form a population consisting of a certain number of independent individuals. Each of these individuals represents a potential solution to the task under consideration.

The operation of genetic algorithms in their standard original alternative is controlled by three characteristic operators. These are: the selection of individuals for another generation based on the value of the fitness function, cross-over, and mutation. It should also be pointed out that the first two operators - that is the selection and cross-over form the principal part of the genetic algorithm functioning. On the other hand, the role of the mutation is rather subtle and substantially different. It is a minor mechanism and its role is to keep the population of potential solutions from stagnation.

The fundamental idea standing behind the cross-over operator is that if you combine two good potential solution candidates for the given task, you can reasonably expect to obtain again good, preferably even better solution candidates. This simple concept proves prospective and fruitful in many circumstances. Unfortunately, there also exist situations, where this assumption fails or at least leads to questionable results or unconvincing substandard algorithm behaviour.

To demonstrate this potential weakness of the genetic algorithm, we choose a model task of binary vector optimization. Specifically, the task is to identify a de Bruijn sequence of a given length between general binary sequences. The task of finding a de Bruijn sequence serves exclusively as a model optimization problem. There are several diverse and more efficient methods to form de Bruijn sequences. The interested reader can consult for instance Sawada et al. (2016) or Levine (2011). This means that de Bruijn sequences themselves are not our principal aim. The primary intention is to investigate how the genetic algorithm behaves in the process of a relatively simple binary optimization problem.

We should also mention the fact that the term genetic algorithms includes nowadays a large family of similar algorithms inspired by natural selection and hereditary mechanisms. These algorithms differ in their specific implementations, selection mechanisms, cross-over, and mutation procedures. They can also include various additional mechanisms contributing to their efficiency or are useful under special circumstances. These include for instance elitism feature (keeping the best up to now found individual and contingently returning it into the subsequent

generations). All these improvements and modifications complicate in certain ways the functioning of the original genetic algorithm and additionally can prove themselves beneficial only in special circumstances.

To keep things simple and clear we used primarily the genetic algorithm in its pure original form. We utilized the conventional selection mechanisms - roulette wheel selection and tournament selection. We implemented the classical one-point crossover operator controlled by one probabilistic parameter and simple mutation also guided by one probabilistic parameter. In principle, it would not be a problem to program more sophisticated and complex features into our version of the genetic algorithm (in fact we even did it in some cases - e.g. elitism). Nevertheless, finally, we decided to keep our version of the genetic algorithm in its standard and simple form. The main reason is that our investigations, conclusions, ideas, and comments have then a broader impact and greater applicability.

## OPTIMIZATION

There is a wide class of important tasks or problems for which no suitable analytic solution strategy or algorithm exists. Many of these tasks are of great importance for science, industry, technology, or economics - see for example Mitsuo and Runwei (2000), Goldberg (2000). In general, almost any task can be formulated as an optimization problem. The search for a solution can then be interpreted as a search through a space of potential solutions. This space is sometimes called the search space. For reasonably small search spaces exhaustive techniques are applicable. For large search spaces, the exhaustive techniques are not adequate and special stochastic searching methods have to be utilized. Genetic algorithms belong to such randomized techniques.

Let us consider fitness function $f$ defined on the set of binary vectors of length $\lambda$, which assigns a value from the set of natural numbers $\mathbb{N}$ to each binary vector. Function $f$ can be symbolically expressed as

$$f : \{0,1\}^{\lambda} \mapsto \mathbb{N}.$$

Let

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_\lambda) \tag{1}$$

is a binary vector of length $\lambda$, $f(\alpha) \in \mathbb{N}$. The cardinality of the set of such binary vectors is determined by relation

$$|\{0,1\}^{\lambda}| = 2^{\lambda}. \tag{2}$$

In the following parts of the article we will concentrate on the identification of $\alpha^{max} \in L^*$, where

$$L^* = \{\alpha^* : f(\alpha^*) = max\{f(\alpha) : \alpha \in \{0,1\}^{\lambda}\}\}.$$

Vector $\alpha^{max}$ can be identified by successive calculations of the function values $f(\alpha)$ at the systematic search of all possible vectors $\alpha$ from relation (1) or by their random choice. For a large search space defined by relation (2) this procedure is not suitable and therefore we use a genetic algorithm to identify the maximum of the fitness function $f$.

## GENETIC ALGORITHM

Below we present the operation scheme of a standard genetic algorithm. In this algorithm, we call the binary vector $\alpha$ defined by relation (1) an individual.

### Genetic algorithm operation scheme

1. Create the initial generation of individuals randomly (or alternatively according to a prescribed scheme). We denote the number of individuals by symbol $n_p$.
2. Evaluate all individuals by means of the fitness function $f$.
3. Repeat until the terminating criterion is met (usually the number of generations denoted by symbol $n_g$).
   (a) Select the individuals for the next generation by means of a roulette wheel or by another mechanism favouring individuals with higher values of the objective function $f$.
   (b) Select randomly couples of individuals and perform the cross-over operation with them - see Figure 1. The cross-over procedure is controlled by cross-over probability denoted by symbol $P_{cr}$.
   (c) Submit the resulting individuals to mutation - see Figure 2. The mutation procedure is controlled by mutation probability denoted by symbol $P_m$.
   (d) Evaluate all individuals in the generation by means of the fitness function $f$.
   (e) Store the best individual in the current generation.

   Endrepeat.
4. Show the best individual.

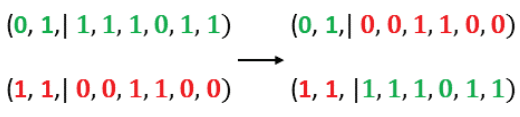



Figure 1: Principle of the one-point cross-over operation. The cross-over point is denoted by a vertical line.

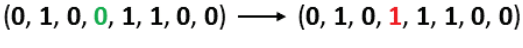



Figure 2: Principle of the mutation operation. The mutation took place at the 4th position of the binary vector.

The fitness proportionate selection, sometimes called the roulette wheel, is a standard selection mechanism used by genetic algorithms. This selection scheme, as well as some more advanced selection procedures, are described in standard introductory books on genetic algorithms, for instance Michalewicz (1999) or Simon (2013).

The part 3.(e) forms a usual part even of the standard genetic algorithm. Without it, the algorithm can lose and often loses the best-found solution candidate during the transition to the next generation.

## BINARY DE BRUIJN SEQUENCES

To test the operation of the genetic algorithm we need a suitable model task. This task should be easily scalable, so that we can test the genetic algorithm at different levels of difficulty. As a model task we chose a search for binary de Bruijn sequences of a given length.

De Bruijn sequences are interesting objects from the area of discrete mathematics relating to combinatorics and coding. As simple examples of binary de Bruijn sequences we can present for instance

$$0011$$

$$00010111.$$

The first case exhibits a cyclic sequence containing all possible variations of 2nd order composed of two symbols 0, 1. Specifically, codes 00, 01, 11, and 10. The second case gives a cycle sequence containing all possible variations of 3rd order. Specifically, codes 000, 001, 010, 101, 011, 111, 110, and 100. In the de Bruijn sequence, all codes are used just once and all possible codes are used. We should point out once more that all de Bruijn sequences are considered cyclic. This means that the codes corresponding to the 6th, 7th and 8th positions are respectively: 111, 110 and 100.

An interested reader can find more detailed information concerning de Bruijn sequences either in the original article by de Bruijn, de Bruijn (1946) or consult more up-to-date sources, for instance Sawada et al. (2016) or Levine (2011).

**Definition.** A binary de Bruijn sequence $b$ of order $\kappa$ is a string of bits $b_i \in \{0,1\}$, $b = \{b_1, \ldots, b_{2^\kappa}\}$ such that each string of length $\kappa$, $\{a_1, \ldots, a_\kappa\} \in \{0,1\}^\kappa$ occurs exactly once in $b$.

The following well-known theorem is taken from Levine (2011).

**Theorem.** Binary de Bruijn sequence of order $\kappa$ exists for all $\kappa$, and there are exactly $2^{(2^{\kappa-1})}$ of them.

Here we use the following notation concerning the de Bruijn sequences. By symbol $\kappa$ we denote the number of positions in the codes considered in de Bruijn sequences. For instance, if we are interested in de Bruijn sequences with 4-positional binary codes (specifically 0000, 0001, ..., 1111) the value of $\kappa$ equals 4. For brevity, the quantity $\kappa$ will further be termed as the code width. The symbol

Table 1: Length of de Bruijn sequences $\lambda$ and numbers of binary $\beta$ and de Bruijn $\delta$ sequences in dependence on the code width $\kappa$

| Code width $\kappa$ | Length of de Bruijn seq. $\lambda$ | Number of de Bruijn seq. $\delta$ | Number of binary seq. $\beta$ |
|---|---|---|---|
| 1 | 2 | 2 | 4 |
| 2 | 4 | 4 | 16 |
| 3 | 8 | 16 | 256 |
| 4 | 16 | 256 | 65536 |
| 5 | 32 | 65536 | $4.295 \cdot 10^9$ |
| 6 | 64 | $4.295 \cdot 10^9$ | $1.845 \cdot 10^{19}$ |
| 7 | 128 | $1.845 \cdot 10^{19}$ | $3.403 \cdot 10^{38}$ |
| 8 | 256 | $3.403 \cdot 10^{38}$ | $1.158 \cdot 10^{77}$ |
| 9 | 512 | $1.158 \cdot 10^{77}$ | $1.341 \cdot 10^{154}$ |
| 10 | 1024 | $1.341 \cdot 10^{154}$ | $1.798 \cdot 10^{308}$ |

$\lambda$ stands for the length of the corresponding de Bruijn sequences. Obviously, we have

$$\lambda = 2^\kappa, \tag{3}$$

since there are exactly $2^\kappa$ different binary codes with $\kappa$ positions.

By $\beta$ we denote the number of all possible binary sequences of a specified length.The length will always be the same as the length of the corresponding de Bruijn sequences. Because we know that the length of a de Bruijn sequence is $2^\kappa$, we automatically have for the length of the de Bruijn sequence $\lambda = 2^\kappa$. This means we can immediately determine the value of $\beta$,

$$\beta = 2^\lambda = 2^{(2^\kappa)}. \tag{4}$$

Symbol $\delta$ we use for the number of de Bruijn sequences characterized by the same value of $\kappa$. In conformity with the preceding theorem, we have for the value $\delta$ the following relation

$$\delta = 2^{(2^{\kappa-1})}. \tag{5}$$

The straightforward comparison of expressions (4) and (5) gives us a surprisingly simple relation between quantities $\beta$ and $\delta$

$$\delta = 2^{(2^{\kappa-1})} = 2^{\frac{2^\kappa}{2}} = (2^{(2^\kappa)})^{\frac{1}{2}} = \beta^{\frac{1}{2}} = \sqrt{\beta}.$$

To provide the reader with a more quantitative idea about the numbers of de Bruijn sequences with respect to the numbers of all possible binary sequences of the same length we include these quantities for several small values of $\kappa$ in the following Table 1.

## RANDOM SEARCH PROBABILITIES

For our further considerations it will be useful to have a clear idea about probabilities of identification of one specific unique de Bruijn sequence and of finding an arbitrary

de Bruijn sequence by a random choice. By a random choice, we mean filling up a sequence with $2^\kappa$ positions by symbols 0 and 1 randomly. That is the probability of putting 0 or 1 into any position is exactly equal to $\frac{1}{2}$. It is, of course, to expect that hitting the specific (fully determined) de Bruijn sequence is much less probable than hitting an arbitrary de Bruijn sequence available, since there are many different de Bruijn sequences, specifically according to (5) there are exactly $2^{(2^{\kappa-1})} = \sqrt{2^\lambda}$ of them.

**Probability of identifiing the specific de Bruijn sequence**

We choose one specific de Bruijn sequence. This is a unique specimen we want to find. The probability of getting this specific de Bruijn sequence by one random choice is given by relation

$$\frac{1}{2^\lambda}.$$

This fact indicates that the probability that we miss this specific individual at one random try is

$$1 - \frac{1}{2^\lambda}.$$

The probability of missing the specific individual at $n$ random tries is

$$(1 - \frac{1}{2^\lambda})^n.$$

The probability of hitting the specific individual at $n$ random tries is then

$$1 - (1 - \frac{1}{2^\lambda})^n.$$

Now, let us suppose that we want this probability to be bigger than a fixed quantity $\rho$,

$$1 - (1 - \frac{1}{2^\lambda})^n > \rho.$$

Using simple modifications we can easily express $n$

$$n > \frac{\ln(1-\rho)}{\ln(1 - \frac{1}{2^\lambda})}. \tag{6}$$

The quantity $n$ determines the number of random tries ensuring to hit the specific bit sequence with probability bigger than $\rho$. The relation (6) represents the lower estimate for $n$. The formula (6) can be further simplified using a Taylor expansion of the natural logarithm function in the vicinity of 1 - see Rudin (1976). The Taylor expansion gives

$$\ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \dots. \tag{7}$$

This means that for $x$ small we have an approximation $\ln(1+x) \approx x$ and for $x < 0$ we even have relation $|x| < |\ln(1+x)|$. This indicates that the approximation does not spoil the lower estimate property. That is

$$n > \frac{\ln(1-\rho)}{\ln(1 - \frac{1}{2^\lambda})} \approx -2^\lambda \ln(1-\rho). \tag{8}$$

To illustrate the meaning of the formula (8) let us suppose the probability value $\rho = \frac{1}{2}$. This assumption gives the lower estimate of the quantity $n_{s,0.5}$ which is the number of random choices that guarantees to achieve the specific de Bruijn sequence with probability greater than $\frac{1}{2}$

$$n_{s,0.5} > 2^\lambda \ln 2. \tag{9}$$

This means that to get the specific individual by random choices with probability $\rho = \frac{1}{2}$, we have to generate at least $2^\lambda \ln 2$ individuals, where $2^\lambda$ is the total number of different bit sequences. The relation (9) represents a rigorous lower estimate. The value of $\ln 2 \simeq 0.693148$. In practical terms, we should generate at least $70\%$ of the total number of all possible bit sequences to have the probability of hitting the specific bit sequence bigger than $50\%$.

**Probability of getting an arbitrary de Bruijn sequence**

This case is analogous to the previous part. For the definite code width $\kappa$ we have according to (5) the following number of various de Bruijn sequences $\delta = 2^{(2^{\kappa-1})}$. This number determines the relative frequency of de Bruijn sequences among general binary sequences. The total number of all binary sequences with a length $2^\kappa$ is according to (4) $\beta = 2^{(2^\kappa)}$. To keep formulas in a simpler form, we use the relation (3) $\lambda = 2^\kappa$. The probability of getting an arbitrary de Bruijn sequence at a random choice is then equal to the ratio of positive cases to all possible cases. In our circumstances, it induces that the probability of getting an arbitrary de Bruijn sequence is

$$\rho = \frac{2^{\lambda^{\frac{1}{2}}}}{2^\lambda} = \frac{1}{2^{\frac{\lambda}{2}}}.$$

We can carry out the same considerations as in the previous part to obtain the number $n$ of random choices that secure finding an arbitrary de Bruijn sequence with probability greater than $\rho$. So, we get

$$n > \frac{\ln(1-\rho)}{\ln(1 - \frac{1}{2^{\frac{\lambda}{2}}})}.$$

Using the Taylor expansion (7) gives

$$n > -2^{\frac{\lambda}{2}} \ln(1-\rho)$$

and putting $\rho = \frac{1}{2}$, we get the final formula

$$n_{a,0.5} > 2^{\frac{\lambda}{2}} \ln 2, \tag{10}$$

where the symbol $n_{a,0.5}$ denotes he minimal number of random choices ensuring to identify an arbitrary de Bruijn sequence with probability greater than $\frac{1}{2}$. The values $n_{a,0.5}$ and $n_{s,0.5}$ given by relations (9) and (10) for small values of $\kappa$ and $\lambda$ are summarized in Table 2. The notation $\lceil x \rceil$ of a real number $x$ is defined as $\lceil x \rceil = min\{p \in \mathbb{Z}, p \geq x\}$, where $\mathbb{Z}$ denotes the set of all integer numbers.

Table 2: Numbers of random choices ensuring identifying any de Bruijn sequence $\lceil n_{a,0.5} \rceil$ and the specific de Bruijn sequence $\lceil \mathrm{n}_{s,0.5} \rceil$ with probability greater than $\frac{1}{2}$ in dependence on the code width $\kappa$

| Code width $\kappa$ | de Bruijn seq. length $\lambda$ | Number of rand. choices $\lceil n_{a,0.5} \rceil$ | Number of rand. choices $\lceil n_{s,0.5} \rceil$ |
|---|---|---|---|
| 1 | 2 | 2 | 3 |
| 2 | 4 | 3 | 12 |
| 3 | 8 | 12 | 178 |
| 4 | 16 | 178 | 45427 |
| 5 | 32 | 45427 | $2.098 \cdot 10^{9}$ |
| 6 | 64 | $2.098 \cdot 10^{9}$ | $1.279 \cdot 10^{19}$ |
| 7 | 128 | $1.279 \cdot 10^{19}$ | $2.359 \cdot 10^{38}$ |
| 8 | 256 | $2.359 \cdot 10^{38}$ | $8.026 \cdot 10^{76}$ |
| 9 | 512 | $8.026 \cdot 10^{76}$ | $9.294 \cdot 10^{153}$ |
| 10 | 1024 | $9.294 \cdot 10^{153}$ | $1.246 \cdot 10^{308}$ |

## GENETIC ALGORITHM TESTING

We tested the operation of the standard genetic algorithm in two configurations:

1. **Identification of a specific fixed de Bruijn sequence.**

   In this case, we select one specific de Bruijn sequence and subsequently try to identify this specific de Bruijn sequence by the genetic algorithm. Under such circumstances, it is obvious that this task has a unique solution. Technically, we compare binary individuals with the selected de Bruijn sequence that is further termed as the specimen. If the individual has the same binary value at a given position, the value of the fitness function is incremented by 1.

   Fitness function $f_1$ is defined by the following prescription. Let $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\lambda)$ is the given fixed de Bruijn sequence of length $\lambda$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_\lambda)$ an arbitrary binary vector of length $\lambda$. Then

   $$f_1(\alpha) = \sum_{i=1}^{\lambda} p_i,$$

   where $p_i = 1$ for $\alpha_i = \gamma_i$ and $p_i = 0$ for $\alpha_i \neq \gamma_i$.

   In this way, individuals similar to the specimen are considered better (have a higher value of the fitness function) and they have higher chances to proceed to subsequent generations. In this way, the genetic algorithm population evolves relatively quickly and successfully and without any apparent obstacles. The specimen is identified relatively fast, of course, the actual number of generations necessary to identify the specimen is dependent on the length of the specimen. This is natural since the length determines the general difficulty of this optimization task.

2. **Identification of an arbitrary de Bruijn sequence.**

   In this case, we seek for an arbitrary de Bruijn sequence of the given length. The solution to this task is evidently not unique since according to (5) there are exactly $\delta = 2^{(2^{\kappa-1})}$ of such de Bruijn sequences. This indicates that from the perspective of random search, this task should be much easier since there are plenty of solutions waiting to be found. The fitness function is constructed in the following way. We traverse cyclically through the given bit sequence representing an individual. Each group of $\kappa$ digits is interpreted as a number expressed in the binary system. That is each individual obtains a sequence of numbers. We check this sequence and each number is counted only once that is we remove all duplicities. The fitness function value is then set to the quantity of unique numbers in the sequence.

   Fitness function $f_2$ is defined by the following prescription. Let us have binary vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_\lambda)$, where $\lambda = 2^\kappa$. Then

   $$f_2(\alpha) = s,$$

   where the value $s$ is determined by the number of different binary codes of length $\kappa$ contained in vector $\alpha$. According to the Theorem above and relation (5) the number of different de Bruijn sequences is $\delta = 2^{(2^{\kappa-1})}$.

   When we seek for a de Bruijn sequence of length $\lambda = 2^\kappa$, the maximal possible fitness is equal to the length $\lambda$ of the binary sequence and it is obvious that such an individual corresponds definitely to a certain de Bruijn sequence. This confirms that the fitness function is constructed correctly.

## EXPERIMENTAL RESULTS

We present, in Tables 3, 4, and 5, a limited selection from the large pool of experimental results. The intention is to provide the reader with quantitative information regarding the behaviour of the genetic algorithm in the model task with different sizes. For each size of the task, the genetic algorithm was run 10 times. We used two selection strategies, standard roulette wheel selection and tournament selection with tournament size 4. The average result and standard deviation are stated in the last two rows of the tables.

Table 3 illustrates the identification of any de Bruijn sequence with length $\lambda = 16$. In the computation the following algorithm parameters were used: number of individuals $n_p = 20$, number of generations $n_g = 40$, cross-over probability $P_{cr} = 0.99$, mutation probability $P_m = 0.06$.

The algorithm attained the target and solved the task without problems even with small values $n_p = 20$ and $n_g = 40$.

Table 4 summarizes the identification of any de Bruijn sequence with length $\lambda = 64$. In the computation the following algorithm parameters were used: number of

Table 3: Search for a de Bruijn sequence with 16 positions. The maximal value of the fitness function is 16.

| Exp. No. | Roulette selection | Tournament selection |
|---|---|---|
| 1 | 16 | 16 |
| 2 | 16 | 16 |
| 3 | 16 | 16 |
| 4 | 16 | 16 |
| 5 | 16 | 16 |
| 6 | 16 | 16 |
| 7 | 16 | 16 |
| 8 | 16 | 16 |
| 9 | 16 | 16 |
| 10 | 16 | 16 |
| Average | 16.0 | 16.0 |
| Stand. deviation | 0.0 | 0.0 |

Table 4: Search for a de Bruijn sequence with 64 positions. The maximal value of the fitness function is 64.

| Exp. No. | Roulette selection | Tournament selection |
|---|---|---|
| 1 | 56 | 64 |
| 2 | 58 | 64 |
| 3 | 56 | 63 |
| 4 | 56 | 64 |
| 5 | 57 | 64 |
| 6 | 58 | 64 |
| 7 | 56 | 63 |
| 8 | 57 | 64 |
| 9 | 57 | 64 |
| 10 | 56 | 61 |
| Average | 56.7 | 63.7 |
| Stand. deviation | 0.823 | 0.972 |

individuals $n_p = 200$, number of generations $n_g = 800$, cross-over probability $P_{cr} = 0.99$, mutation probability $P_m = 0.02$.

Despite higher values $n_p$ and $n_g$, the roulette wheel selection gives evidently insufficient results. The target value is $64$. The tournament selection gives much better and acceptable results. The reason is that the tournament selection exerts generally much higher selection pressure in comparison with the roulette wheel selection. Another advantage of the tournament selection is that it is easily and suitably quantifiable by the tournament size parameter.

Table 5 illustrates the identification of any de Bruijn sequence with length $\lambda = 256$. In the computation the following algorithm parameters were used: number of individuals $n_p = 200$, number of generations $n_g = 800$, cross-over probability $P_{cr} = 0.99$, mutation probability $P_m = 0.004$.

From Table 5 it is apparent that the algorithm gives much better results using tournament selection. The target value now is $256$ and the results of the algorithm with the roulette wheel selection are extremely low. But even the results of the algorithm using tournament selection are

Table 5: Search for a de Bruijn sequence with 256 positions. The maximal value of the fitness function is 256.

| Exp. No. | Roulette selection | Tournament selection |
|---|---|---|
| 1 | 195 | 236 |
| 2 | 198 | 233 |
| 3 | 196 | 234 |
| 4 | 196 | 238 |
| 5 | 193 | 236 |
| 6 | 197 | 236 |
| 7 | 193 | 240 |
| 8 | 197 | 236 |
| 9 | 194 | 240 |
| 10 | 195 | 235 |
| Average | 195.4 | 236.4 |
| Stand. deviation | 0.823 | 0.972 |

well below the target value. Additionally, from the course of the computation, it was apparent that the algorithm exhibited symptoms of stagnation. The best-attained values did not even change when the parameters $n_p$ and $n_g$ were increased.

The behaviour of the genetic algorithm was completely different in the situation when the task was to identify one specific in advance chosen de Bruijn sequence. In this case, the specific target de Bruijn sequence was always identified without problems, even with both the selection mechanisms.

## GENETIC ALGORITHM PARADOX

In general, we face two diametrically different sorts of behaviour of the genetic algorithm. When in the first case we look for the specific in advance given de Bruijn sequence, the genetic algorithm behaves efficiently and finds the solution without any problems up to $\kappa = 10$ inducing $\lambda = 1024$.

On the other hand, in the second task, we look for any of the de Bruijn sequences out of the number $2^{(2^{\kappa-1})}$, the algorithm solves efficiently only relatively small tasks corresponding to $\kappa = 4$ and $\kappa = 5$ corresponding to $\lambda = 16$ and $\lambda = 32$. For bigger tasks $\kappa = 6$ and $\kappa = 7$ corresponding to $\lambda = 64$ and $\lambda = 128$ the algorithm becomes inefficient and without special modifications (e.g. tournament selection) does not attain the maximum of the fitness function. Using the genetic algorithm for tasks with values of $\kappa >= 8$ corresponding to $\lambda >= 256$ in these circumstances seems rather questionable even with special modifications - see Table 5. We recorded obvious stagnation of the algorithm well below the target value. It is also apparent that the situation gets worse with the increasing length of the de Bruijn sequence.

The reasons standing behind this curious behaviour are of great importance. We came to the conclusion that the source of this behaviour is the multimodality of the fitness function. When we consider a multimodal task, the usual

assumption of the genetic algorithm that two prospective potential solutions combine by the cross-over operation to other prospective potential solutions fails. Combining these individuals can lead to individuals with much worse value of the fitness function. This fact finally leads to the stagnation of the genetic algorithm far from the maxima of the fitness function.

## CONCLUSIONS

In this article, we handle the topic of convergence of the genetic algorithm to the maximum of the fitness function. This theme is important from theoretical as well as practical reasons. If a scientist considers using the genetic algorithm, he should have a concrete idea regarding what he can expect. The topic is even more urgent since genetic algorithms are relatively demanding concerning computational resources.

In the next part, we describe the identification of de Bruijn sequences by the genetic algorithm. We present a limited selection of numerical results from the large pool of experimental data - see Tables 3, 4, and 5. We come to an interesting result that for the genetic algorithm is much easier to find the fixed de Bruijn sequence than to identify an arbitrary de Bruijn sequence. This leads to the conclusion that the genetic algorithm is not suitable for the optimization of multimodal tasks without any further modifications of the algorithm or the fitness function. The reason for this pathological behaviour are the fluctuations of the algorithm between maxima of the fitness function caused by the cross-over feature of the algorithm. These fluctuations finally lead to the stagnation of the algorithm far from maxima of the fitness function.

There are three principal points ensuing from our investigations concerning the binary optimizations.

1. In general, the standard genetic algorithm can be used as a robust solver providing solutions to diverse intricate binary optimization tasks. Nevertheless, the quality of the solution depends considerably on the modality of the fitness function of the task.

2. For unimodal fitness functions the genetic algorithm can be considered a relatively good optimizer providing good results and converging relatively reliably to the maximum of the fitness function.

3. For multimodal fitness functions the situation is different. The genetic algorithm can be negatively influenced by many fitness function maxima. The convergence process can stagnate because of the algorithm fluctuations between fitness function maxima.

Since we tested the genetic algorithm for this article exclusively on the tasks concerning the de Bruijn sequences identification, future research should continue and focus on other binary and also non-binary optimization tasks.

## References


N. G. de Bruijn. A combinatorial problem. *Indag. Math. (N.S.)*, 49:758–764, 1946.

D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 2000.

L. Levine. Algebraic combinatorics. Lecture Notes, 21 18.312, Cornell University, Ithaca, NY, US, 2011.

Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Springer-Verlag Berlin Heidelberg New York, 1999.

G. Mitsuo and Ch. Runwei. *Genetic Algorithms and Engineering Optimization*. Joh Wiley Sons, Inc., 605 Third Avenue, New York, 2000.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Education, London, 1976.

J. Sawada, A. Williams, and D. Wong. A surprisingly simple de Bruijn sequence construction. *Discrete Mathematics*, 339:127–131, 2016.

D. Simon. *Evolutionary Optimization Algorithms, Biologically-Inspired and Population-Based Approaches to Computer Intelligence*. Willey, Published by John Wiley Sons, Inc., Hoboken, New Jersey, 2013.


## AUTHOR BIOGRAPHIES


**ROMAN KNOBLOCH** was born in Turnov, Czechoslovakia. After high school, he studied mathematics and physics at the Faculty of Mathematics and Physics, Charles University in Prague. He is occupied as an assistant professor at the Department of Mathematics of the Technical University in Liberec, the Czech Republic. His principal areas of interest are: mathematical modeling, up-to-date optimization techniques, continuum mechanics, and bit fields generation. His e-mail address is: roman.knobloch@tul.cz
ORCID: 0000-0003-2427-3479



**JAROSLAV MLYNEK** was born in Trnava, Czechoslovakia and he studied numerical mathematics at Charles University in Prague at the Faculty of Mathematics and Physics. His professional activities are focused on the study of convergence of evolutionary algorithms, especially differential algorithms. He also deals with the optimization of fiber winding on polymer composite frames and the optimization of heating of thin-walled metal molds in the production of artificial leather. He currently holds the position of an associate professor at the Technical University of Liberec. His e-mail address is: jaroslav.mlynek@tul.cz
ORCID: 0000-0002-3386-6738

# ON THE LOCATION-INDEPENDENT RECONSTRUCTION OF PHOTOSYNTHETICALLY ACTIVE RADIATION IN THE WATER COLUMN USING NEURAL NETWORKS

Martin Maximilian Kumm[1], Christoph Tholen[2], Lars Nolle[1,2] and Frederic Stahl[2]

[1]Jade University of Applied Science, Department of Engineering Sciences
Wilhelmshaven, Germany, e-mail: {martin.kumm|lars.nolle}@jade-hs.de
[2]German Research Center for Artificial Intelligence (DFKI), Marine Perception
Oldenburg, Germany, e-mail: {christoph.tholen|frederic_theodor.stahl}@dfki.de

**KEYWORDS**

BGC-Argo float, photosynthetically active radiation prediction, artificial neural network, machine learning, generalisation.

**ABSTRACT**

Accurate reconstruction of photosynthetically active radiation (PAR) in aquatic environments is critical for understanding primary production and ecosystem dynamics. This study evaluates the generalisation abilities of artificial neural networks (ANNs) for location-independent PAR reconstruction using data from BGC-Argo floats. The proposed ANN model is trained on datasets from multiple geographic regions and validated against independent test data from diverse oceanic locations. Comparisons with multiple linear regression (MLR) and regression tree (RT) models demonstrate that the ANN consistently achieves superior predictive accuracy, with $R^2$ values exceeding 0.97 in most test cases. The results indicate that neural networks can effectively generalise across different marine environments, even in regions with distinct optical properties. Notably, the ANN outperforms alternative models except in one test case, highlighting the potential influence of regional environmental factors. This study underscores the potential of machine learning techniques to enhance bio-optical sensor configurations and reduce the necessity for dedicated PAR sensors.

## INTRODUCTION

The dramatic increase in environmental challenges, ranging from climate change and sea level rise to pollution, deoxygenation, and ocean warming, has driven the development of innovative methods for monitoring critical environmental parameters (Lotze, et al., 2019; Wollschläger, et al., 2021). Modern operational oceanography now relies on an array of autonomous platforms, among which Argo floats have emerged as a cornerstone (Roemmich, et al., 2019). There are over 4,000 floats deployed globally and more than 1,600 specialised biogeochemical (BGC-Argo) floats in operation since 2012. To date, over 50,000 multispectral profiles have been collected using this configuration of bands, and these profiles have been evaluated in terms of data quality (Jutard, et al., 2021; Organelli, et al., 2016; Stoer, et al., 2023). These platforms continuously collect high-resolution vertical profiles from the ocean's surface to depths of approximately 2,000 meters. Data transmission via Iridium or Argos satellite systems ensure that information is publicly and freely available through two global data assembly centers (GDAC). Data is typically available within 24 hours (see Argo website https://argo.ucsd.edu).

Initially designed with a three-sensor configuration to capture fundamental physical oceanographic properties, Argo floats have evolved with the BGC-Argo initiative to include a suite of additional physical, chemical, and bio-optical sensors (Johnson, et al., 2017; Claustre, et al., 2011). A key instrument in this expanded sensor package is the Ocean Colour Radiometer (OCR), such as the OCR-504 from SATLANTIC Inc./Sea-Bird Scientific, which measures radiometric observations at four channels. Three of these channels,380 nm, 412 nm, and 490 nm, were selected in this study for their sensitivity to variations in the underwater light field, while the fourth channel is dedicated to recording Photosynthetically Active Radiation (PAR). PAR, which integrates downward irradiance between 400 nm and 700 nm, is essential for assessing the light available for primary production in natural waters (Behrenfeld & Falkowski, 1997; Morel & André, 1991).

In parallel with these technological advancements, the rapid expansion in sensor diversity has necessitated more sophisticated data management, quality control, and analysis methods, with machine learning emerging as a particularly promising tool (Claustre, et al., 2011; Jiang, et al., 2017). Recognizing the potential for streamlining sensor configurations, the BGC-Argo community has suggested reconfiguring the OCR by omitting the dedicated PAR channel. This proposal is based on the observation that PAR measurements can be reliably reconstructed from the three remaining spectral channels 380 nm, 412 nm, and 490 nm, and pressure. Previous studies have demonstrated that techniques such as Multiple Linear Regression (MLR) (Stahl, et al., 2021), Regression Trees (RT) (Stahl, et al., 2021) and Artificial Neural Networks (ANN) (Kumm, et al., 2022; Pitarch, et

al., 2025) can successfully predict PAR sensor readings. However, most recent research has shown an unsatisfactory accuracy of the models relaying on the three wavelengths 400 nm, 412 nm and 490 nm for the Freefall Profiler dataset, which covers more geolocations and is therefore more complex (Tholen, et al., 2024).
Building on these findings, the present study evaluates an ANN model with datasets from different geolocations to demonstrate the ability of generalisation of the ANN to reconstruct the PAR values and this work improves the architecture of the ANN. An MLR and an RT were also trained as a comparison to the ANN.

## VARTICAL RADIOMETRIC MEASURMENT OF THE WATER COLUMN

Among the six essential variables measured by BGC-Argo floats, the underwater light field is a key parameter (Claustre, et al., 2019). To capture this, OCR-504 from SATLANTIC Inc./Sea-Bird Scientific is utilised, measuring downward irradiance at three specific wavelengths, 380 nm, 412 nm, and 490 nm, along with PAR, which is integrated across the 400–700 nm range (Satlinc, 2013). The arrangement of these four sensors is depicted on the right hand side of Figure 1. These three specific wavelengths were chosen due to their strong correlation with the primary variations in underwater optical properties (Xing, et al., 2012; Organelli, et al., 2016). The PAR sensor data is commonly used to estimate the amount of light available for primary production in aquatic environments (Mignot, et al., 2018).
To ensure a fair comparison of the methods, this study utilised a similar dataset for training as in (Kumm, et al., 2022; Stahl, et al., 2021). The float data was collected and made publicly available by the International Argo Program, along with contributions from national initiatives (http://www.argo.ucsd.edu and http://argo.jcommops.org).

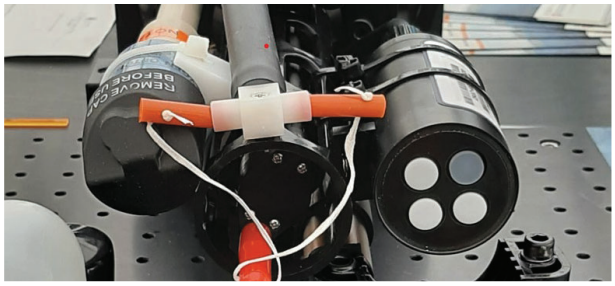


Figure 1: OCR-504 mounted on the BGC-Argo float

## TRAINING AND TEST DATA

For training our model, similar BGC-Argo floats as used by Kumm, et al. (2022) were employed. However, the dataset was larger since new data has been acquired since the original publication. All samples collected above 100 dbar were removed, as no light reaches those depths. To prevent an unbalanced dataset, an equal number of samples from each dataset was used for training. Specifically, the number of samples from the smallest dataset, the North Atlantic dataset with 2,124 samples, served as the reference. For the other datasets, 2,124 samples were randomly selected, resulting in a balanced training dataset. For each set, an 80/20 train-test split was applied individually, with 10% of the training data further reserved for validation, before merging them into training, test, and validation sets.
The aim of the publication was to demonstrate the ability of generalisation of the developed ANN with respect to geolocations. Therefore, the ANN was subsequently tested with data from eight additional floats, located at different regions of the world. Table 1 shows the different datasets from the different sites used in this work.

Table 1: Training and test data

| Identifier | Location | Samples |
|---|---|---|
| Training | | |
| WMO 7900561 | North Atlantic | 2,124 |
| WMO 7900562 | Mediterranean Sea | 2,124 |
| WMO 7900579 | Baltic Sea | 2,124 |
| WMO 7900580 | Baltic Sea | 2,124 |
| Test | | |
| WMO 7900585 | North Atlantic | 4,116 |
| WMO 7900583 | Tasmania Sea | 1,843 |
| WMO 4903711 | Caspian Sea | 544 |
| WMO 5905505 | Tasmanian Sea | 700 |
| WMO 5906661 | Indian Ocean | 1,471 |
| WMO 6902906 | South Pacific | 10,042 |
| WMO 6990638 | North Atlantic | 770 |
| WMO 7902198 | North Pacific | 230 |

Figure 2 depicts the locations of the floats. It can be seen that we deliberately tested the ANN with data, which were not included in the training process.

## ARCHITECTURE OF THE ANN

In this study, the architecture proposed by Kumm et al. (2022) was largely followed. The ANN is implemented as a feed-forward neural network consisting of three main components: an input layer, a hidden layer, and an output layer. The input layer receives the four sensor data inputs Pressure, Ed380, Ed412, and Ed490, while the output layer provides the value for PAR.
The ReLU function was used as the activation function to capture the non-linearities in the model. Based on Kumm et al. (2022), the Root Mean Squared Propagation algorithm was employed for training with a learning rate of 0.01, and the Mean Absolute Error (MAE) was chosen as the loss function.
To determine the optimal number of nodes in the hidden layer, ANNs with varying node counts from 1 to 50 were trained. For each node count, the ANN was trained 20 times to calculate the mean and standard deviation of performance. A batch size of 32 was used during training. Figure 3 depicts the development of the mean value and

the standard deviation from 13 to 50 nodes. Each ANN was trained for 1000 epochs. The $R^2$ from the ANNs with less than 13 nodes in the hidden layer were below 0.9935. This systematic variation of the node count enabled a detailed investigation of the impact of ANN topology on model performance, thereby facilitating a well-founded decision regarding the optimal ANN structure. The best $R^2$ of 0.9974 was reached with 48 nodes in the hidden layer. The standard deviation with this number of nodes was 0.0003.


Figure 2: Location of the datasets

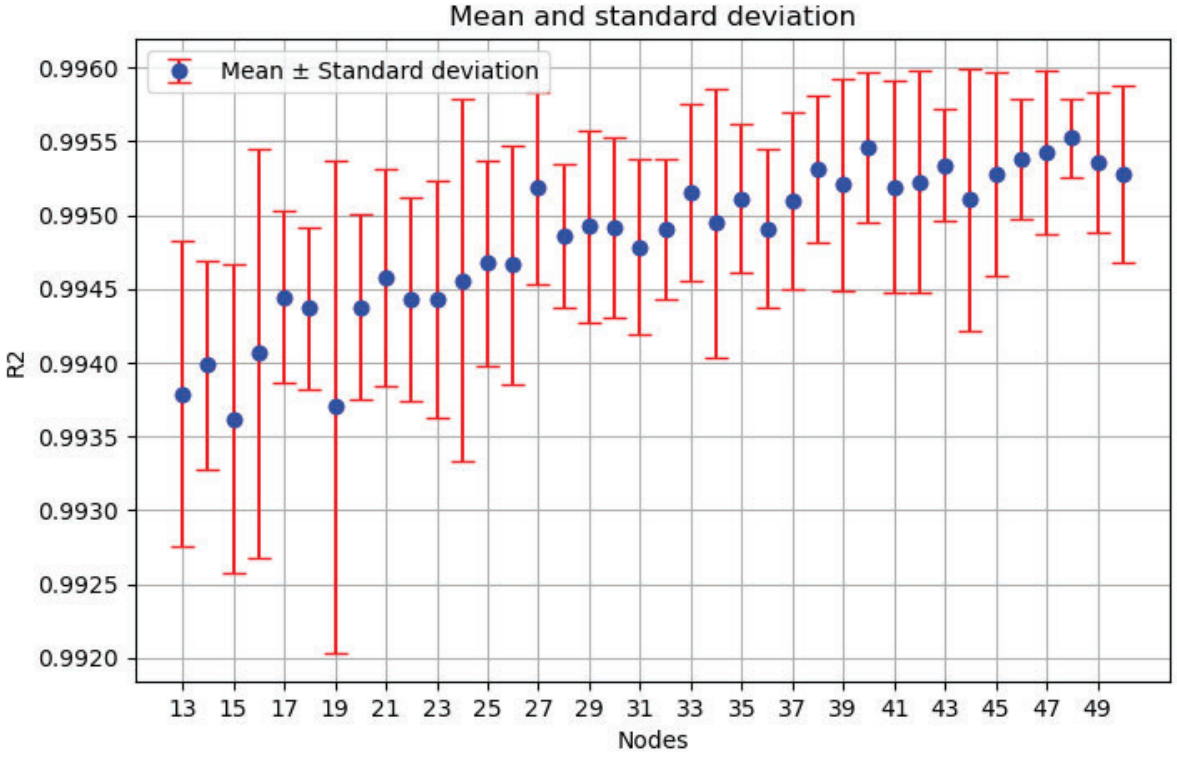



Figure 3: Mean and standard deviation of the ANN

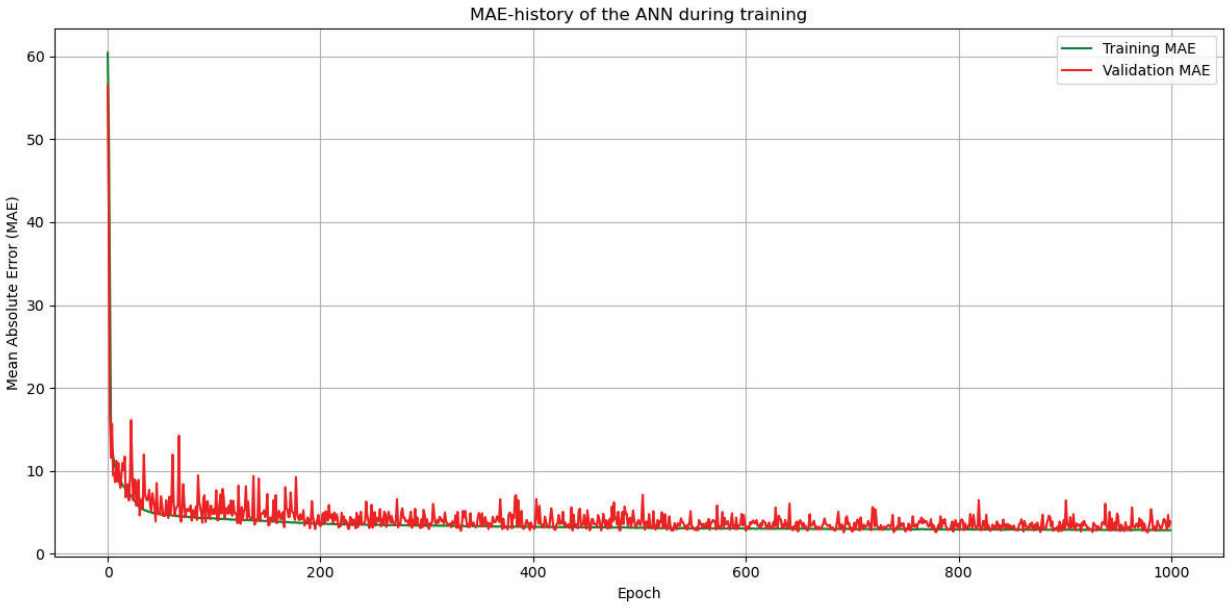



Figure 4: MAE-history of the ANN during training

## EVALUATION OF THE ANN

Figure 4 illustrates the error progression of the ANN for one training run for both the training and validation datasets over time. The graph indicates that after 1000 epochs, no significant improvement in performance is observed, suggesting that the model has reached convergence.

Furthermore, the results show no signs of overfitting. The validation error remains stable even after 1000 epochs, indicating that the model maintains its ability to generalise well to unseen data. This stability suggests that the ANN effectively learns the underlying patterns in the data.

An MLR and an RT were also trained. The exact same training data as for the ANN was used for this. A comparison of the predictions for the test set can be seen in Figure 5. It can be observed, that the three methods are able to predict the PAR.

To assess the generalisation capability of the trained ANN, the MLR and the RT, eight datasets from previously unseen locations were used for testing. The results, presented in Table 2, demonstrate that the ANN exhibits strong generalisation properties compared with MLR and RT due to his consistently higher $R^2$ value.

The coefficient of determination ($R^2$) values for all datasets range between 0.97 and 0.99, indicating a high level of agreement between predicted and observed values. However, one dataset yielded a slightly lower $R^2$ value of 0.8955.

Table 2: $R^2$ from the different models

| **Dataset** | **ANN** | **MLR** | **RT** |
|---|---|---|---|
| Training | 0.9974 | 0.9872 | 0.9923 |
| WMO 7900585 | 0.9739 | 0.9378 | 0.9130 |
| WMO 7900583 | 0.9944 | 0.9823 | 0.9864 |
| WMO 4903711 | 0.9910 | 0.8838 | 0.9318 |
| WMO 5905505 | 0.9975 | 0.9921 | 0.9944 |
| WMO 5906661 | 0.9983 | 0.9914 | 0.9823 |
| WMO 6902906 | 0.9773 | 0.9424 | 0.9701 |
| WMO 6990638 | 0.8955 | 0.9636 | 0.9513 |
| WMO 7901106 | 0.9823 | 0.9548 | 0.8867 |

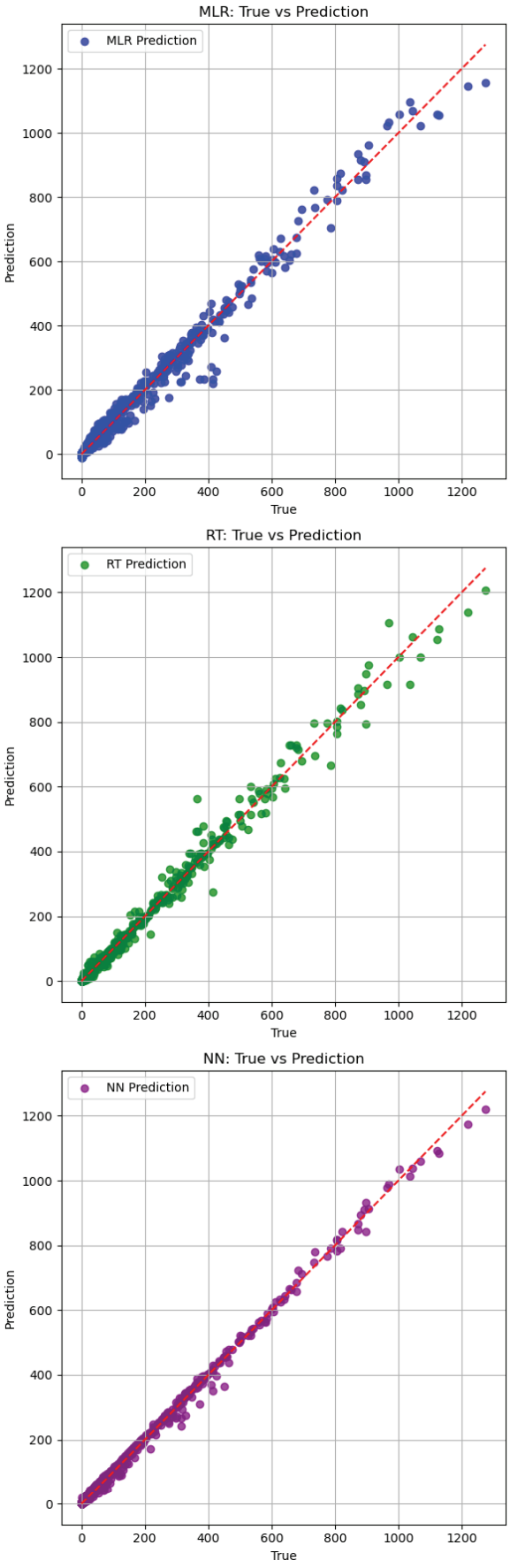



Figure 5: Prediction vs. observation for MLR, RT and ANN

## DISCUSSION AND FUTURE WORK

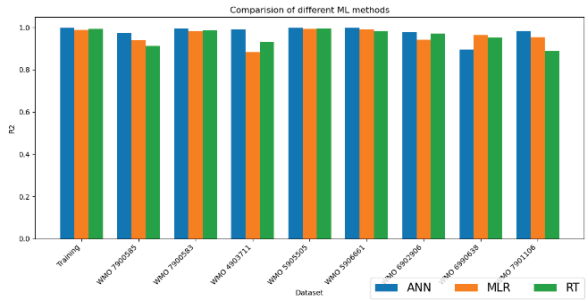



Figure 6: Comparison of the models regarding their location

As shown in Table 2 and Figure 6, the ANN outperforms the MLR and the RT on every tested site except for WMO 6990638 with 0.8955. This Argo float is located to the east of Greenland. This deviation could be attributed to the unique environmental conditions at the data collection site. In this region, highly variable light conditions, influenced by factors such as sea ice, may have introduced additional complexities that were not fully captured by the model during training. This needs to be investigated in more detail in future studies.
On the other hand, the results of the WMO 4903711 should also be emphasised. This float is located in the Caspian Sea. The Caspian Sea has a different water composition due to its geographical location. Nevertheless, the PAR values are predicted with an accuracy of 0.9910. This is a further indication of the generalisation properties of the developed ANN.

Future research should focus on further improving the predictive performance and robustness of the ANN. One potential avenue is the integration of additional spectral bands, which could enhance the model's ability to capture complex underwater optical properties. Another important aspect is the extension of the dataset to include more diverse oceanic regions, particularly areas with extreme environmental conditions, such as polar waters. This would help assess the model's adaptability to highly variable light conditions and improve its overall reliability.

## ACKNOWLEDGMENT

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

## REFERENCES

Behrenfeld, M. J. & Falkowski, P. G., 1997. A Consumer's Guide to Phytoplankton Primary Productivity Models. *Limnology and Oceanography 42.*

Claustre, H. et al., 2011. Bio-Optical Sensors on Argo Floats, In: Claustre, H. (ed.). *Reports and Monographs of the International Ocean-Colour Coordinating Group*, pp. 1-89.

Claustre, H., Johnson, K. & Takeshita, Y., 2019. Observing the Global Ocean with Biogeochemical-Argo.

Jiang, Y. et al., 2017. A machine learning approach to argo data analysis in a thermocline. *Sensors, Vol. 17, No. 10, Article 2225*.

Johnson, K. S. et al., 2017. Biogeochemical sensor performance in the SOCCOM profiling float array. *Journal of Geophysical Research: Oceans Vol. 122 Issue 8*, p. 6416–6436.

Jutard, Q., Organelli, N., Briggs, N. & al., e., 2021. Correction of Biogeochemical-Argo Radiometry for Sensor Temperature-Dependence and Drift: Protocols for a Delayed-Mode Quality Control. *Sensors 21.*

Kumm, M. M. et al., 2022. On an Artificial Neural Network Approach for Predicting Photosynthetically Active Radiation in the Water. *Artificial Intelligence XXXIX, Lecture Notes in Computer Science.*

Lotze, H. K., Tittensor, D. P., Bryndum-Buchholz, A. & al., e., 2019. Global ensemble projections reveal trophic amplification of ocean biomass declines with climate change. *Proc. Natl Acad. Sci. USA 116*, pp. 12907 - 12912.

Mignot, A., Ferrari, R. & Claustre, H., 2018. Floats with bio-optical sensors reveal what processes trigger the North Atlantic bloom. *Nature Communications*.

Morel, A. & André, J.-M., 1991. Pigment Distribution and Primary Production in the Western Mediterranean as Derived and Modeled From Coastal Zone Color Scanner Observations. *Journal of Geophysical Research: Oceans 96.*

Organelli, E. et al., 2016. A Novel Near-Real-Time Quality-Control Procedure for Radiometric Profiles Measured by Bio-Argo Floats: Protocols and Performances. *J. Atmos. Ocean. Technol. 33* , p. 937–951.

Pitarch, J., Leymarie, E., Vellucci, V. & Massi, L., 2025. Accurate estimation of photosynthetic available radiation from multispectral downwelling irradiance profiles. *Limnology and Oceanography: Methods*.

Pitarch, J. et al., 2025. Accurate estimation of photosynthetic available radiation from multispectral downwelling irradiance profiles. *Limnology and Oceanography: Methods*.

Roemmich, D., Alford, M. H., Claustre, H. & al., e., 2019. On the Future of Argo: A Global, Full-Depth, Multi-Disciplinary Array. *Frontiers in Marine Science, Vol. 6, Article 439.*.

Satlinc, 2013. Operation manual for the OCR-504. In: *SATLANTIC Operation Manual SAT-DN-00034.* s.l.:s.n., p. 66.

Stahl, F., Nolle, L., Jemai, A. & Zielinski, O., 2021. A Model for Predicting the Amount of Photosynthetically Available Radiation from BGC-ARGO Float Observations in the Water Column.

Stoer, A. C., Takeshita, Y., Maurer, T. L. & al., e., 2023. A Census of Quality-Controlled Biogeochemical-Argo Float Measurements. *Frontiers in Marine Science 10.*

Tholen, C., Nolle, L., Wollschläger, J. & Stahl, F., 2024. Model Generalisation for Predicting the Amount of Photosynthetically Available Radiation in the Water Column from Freefall Profiler Observations. *ECMS 2024 Proceedings*.

Wollschläger, J. et al., 2021. Climate Change and Light in Aquatic Ecosystems: Variability & Ecological Consequences. *Frontiers in Marine Science vol. 8*.

Xing, X. et al., 2012. Combined processing and mutual interpretation of radiometry and fluorometry from autonomous profiling Bio-Argo floats: 2. Colored dissolved organic matter absorption. *JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 117*.

## AUTHOR BIOGRAPHY

**MARTIN MAXIMILIAN KUMM** graduated from Jade University of Applied Sciences in Wilhelmshaven, Germany, with a Master degree in Mechanical Engineering in 2022. Since 2020 he is a research fellow at the Jade University of Applied Sciences responsible for the research aircraft. He also works in a joint project between Jade University of Applied Sciences, the German Research Center for Artificial Intelligence (DFKI) and marinom GmbH for the development of explainable artificial intelligence decision support systems for nautical officers.

**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.

**FREDERIC STAHL** received his Dipl.-Ing. (FH) degree in bioinformatics from the University of Applied Sciences, Weihenstephan, Germany, in 2006, and the Ph.D. degree in computer science from the University of Portsmouth, U.K., in 2010. From 2010 to 2012, he was a Senior Research Associate in the Department of Computer Science at the University of Portsmouth. In 2012, he worked as a Lecturer in the Department of Design Engineering and Computing at Bournemouth University, U.K. From 2012 to 2019, he served as a Lecturer and Associate Professor at the University of Reading, U.K. In 2019, he joined the German Research Centre for Artificial Intelligence (DFKI GmbH) as a Senior Researcher and became the Scientific Director of the DFKI Research Department Marine Perception in 2023. Since 2023, he has also been the Head of the AI Competence Center for Environment and Sustainability (DFKI4Planet). He has published over 100 articles in peer-reviewed conferences, journals, and book chapters and has been working in the fields of data science and artificial intelligence for nearly twenty years.

# EVALUATING TEMPERATURE UNIFORMITY IN PASTEURIZATION: CFD AND EXPERIMENTAL APPROACH FOR VISCOUS FLUIDS

Natalya Lysova[1], Federico Solari[1], Leonardo Restori[1] and Roberto Montanari[1]

[1]Department of Engineering for Industrial Systems and Technologies, University of Parma,
e-mail: natalya.lysova@unipr.it

**KEYWORDS**
Computational Fluid Dynamics, viscous products, heat exchanger, virtual sensor, experimental validation, industrial process optimization.

**ABSTRACT**

Ensuring uniform temperature distribution during thermal treatments is crucial for food safety and quality, particularly in the case of high-viscosity products. This study investigates temperature uniformity in a tube-in-tube heat exchanger using a combined Computational Fluid Dynamics (CFD) and experimental approach. CFD provides detailed thermal profiles, while the experimental tests carried out using concentrated orange juice (40 Brix) allow to validate the simulation results and confirm the presence of significant temperature variations across the pipe section. The results highlight the need for more precise thermal control strategies to mitigate risks of localized undertreatment or excessive heating, which can lead to food safety concerns or product degradation. A virtual sensor approach based on simulation results is proposed to this end. This solution would allow estimating internal temperature distributions based on external sensor data, acting as an essential tool for real-time process optimization that could be integrated into Digital Twin frameworks. This approach would enhance pasteurization efficiency, reduce fouling, and ensure consistent product quality. Future work will extend the investigation and application of simulation-based virtual sensors to this and other heat exchanger designs, to enhance and support the implementation of Digital Twin technologies for advanced control of food processes.

## INTRODUCTION

Thermal treatment is a fundamental process in the food industry, playing a crucial role in ensuring microbial safety while preserving the organoleptic properties of food products throughout their shelf life. Pasteurization and sterilization are the primary thermal processing methods, both of which rely on either direct or indirect heat exchange technologies.

The heat exchanger should be selected to maximize energy efficiency based on the type of product to be treated. Direct exchangers are the most efficient, but they can only be used with low-viscosity products without suspended solids and when treatment temperatures exceed 140°C. For lower treatment temperatures (typically used with high-acid products), plate heat exchangers are used. When the viscosity of the product is high, or when there are many fibers or suspended solids, tubular exchangers are to be preferred. Among them, tube-in-tube exchangers represent the most flexible, cleanable, and simplest solution, even though less energy-efficient, allowing for reduced fouling risk compared to plate heat exchangers. However, significant issues may persist, mainly related to non-uniform heating, which may result in localized undertreatment or excessive heating, with the former leading to safety issues, and the latter to browning and product degradation.

Typically, two main approaches have been explored to address these issues: the introduction of mixing devices at the exit of the heat exchanger, and the adoption of alternative, also non-thermal, pasteurization techniques. The incorporation of static or mechanical mixers can enhance uniformity and heat distribution; however, this method presents limitations when suspended solids are present, as excessive mixing may compromise their structural integrity and sensory attributes. Additionally, in the case of highly viscous products, the presence of additional components increases fouling issues generating the need for frequent and labor-intensive cleaning procedures. In contrast, alternative non-conventional heat treatments, such as Ohmic Heating and High-Pressure Pasteurization, have been investigated for their potential to enhance thermal uniformity and process efficiency on such kind of products (Pereira et al., 2010; Roobab et al., 2018). Other emerging techniques, including Pulsed Electric Field (PEF) and Ultraviolet (UV) treatments, remain impractical for viscous fluids due to insufficient penetration depth, as well as limited energy efficiency and cost-effectiveness.

Despite ongoing advancements in alternative pasteurization methods, traditional heat exchangers remain widely utilized, particularly for highly viscous food products. Enhancing the performance of these systems is imperative to achieve an optimal balance between microbial safety and the preservation of organoleptic properties (Lysova et al., 2024; Perone et al., 2021; Jha et al., 2019), but also to improve energy efficiency as some studies underline (Pereira et al., 2010). Notably, temperature distribution at the outlet of shell-and-tube heat exchangers has received limited attention in existing studies, especially concerning heat-sensitive food products. The focus has been mainly on the study of the cold spots during the treatment (Jha et al., 2019), geometry (Kola et al., 2021), and process

parameters (Lysova et al., 2022), but without paying particular attention to the temperature distribution of the product leaving the exchanger. A more comprehensive understanding of temperature gradients within these systems can facilitate improved thermal cycle management, thereby minimizing peak temperatures and optimizing treatment uniformity according to product viscosity. Furthermore, an in-depth analysis of fluid velocity profiles as a function of rheological properties could enable more precise control over the thermal treatment, leading to enhanced energy efficiency, reduced fouling, and improved product quality. Some studies in this area include (Perone et al., 2021; Jha et al., 2019; Kola et al., 2021; Tancredi et al., 2020).

Computational Fluid Dynamics (CFD) analysis can be an excellent way to simulate and investigate efficiency problems without carrying out many tests on physical plants. Indeed, the simulations provide a detailed picture of the temperature distribution within the system so sensors can be strategically placed in areas with significant temperature gradients, and simulation results could integrate sensor data to gain a complete understanding of the thermal profile (Jha et al., 2019). Indeed, using the results of CFD simulations and data collected from real sensors, virtual sensors can be created to estimate temperature in inaccessible locations where physical sensors cannot be installed (Guzmán et al., 2019; Tancredi et al., 2022) or, in the case of highly viscous fluids, where fouling could occur, and their cleaning would be impossible. As an example, by combining physical sensor data with CFD simulations, temperature information can be derived at critical points, such as the temperature at the center of the flow (Yolmeh et al., 2017; Han et al., 2015; Liu et al., 2021; Kola et al., 2021).

Results from CFD simulations provide an accurate description of thermo-fluid dynamics in pasteurization processes, but their high computational cost prevents their direct use in real-time control. To overcome this limitation, reduced-order models (ROM) can be developed using dimensionality reduction techniques or methods based on parametric interpolation. ROM preserves the essential features of the system by significantly reducing the number of equations to be solved, thus allowing fast predictions (Lysova et al., 2022). This simplified model can be integrated into predictive control strategies (MPC) to optimize real-time operating parameters such as flow, temperature, or pressure. In addition, also ROM, as well as simulation results, can act as a virtual sensor (Lysova et al., 2024) to estimate process variables that are difficult to measure directly, improving plant monitoring and efficiency. To summarize, virtual sensors could be used to estimate physical quantities of interest at specific points where physical sensors cannot be installed.

In the future, these insights could contribute to the development of cost-effective, high-precision process control solutions, seamlessly integrating with existing industrial setups through the implementation of Digital Twin technologies. Such advancements would enable real-time monitoring and adaptive control of processing parameters, thereby optimizing heat exchanger performance and ensuring superior product safety and quality.

## MATERIALS AND METHODS

### Experimental setup

The pilot-scale processing system comprises a stainless-steel tube-in-tube heat exchanger having a maximum processing capacity of 2,500 l/h. In particular, the product flows upwards through the inner tube while heating water circulates counter-currently within the outer shell. The inner tube has an internal diameter of 38 mm, thus allowing the treatment of highly viscous products, as well as formulations containing suspended particulates having a maximum size of up to 10 mm. The heat exchanger consists of six horizontally aligned heating modules, each approximately 4 meters in length, arranged in a vertically stacked configuration.

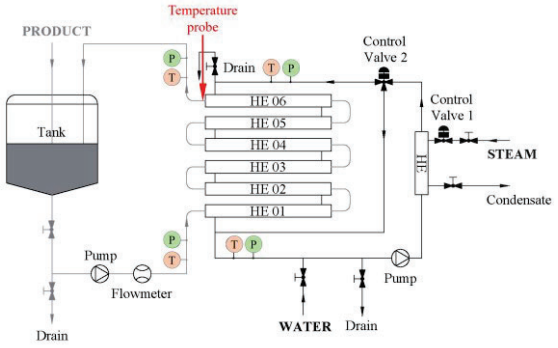



Figure 1: Schematic layout of the plant and position of the outlet temperature probe used

The inner pipes of successive modules are interconnected via 180° return bends, whereas the outer shell sections are linked through vertical flange couplings.

The heat exchanger is connected to a stainless-steel mixing tank where the product is continuously recirculated. The flow through the system is provided by a twin-screw volumetric pump having a maximum flow rate of 5 $m^3$/h. The flow rate is monitored using a mass flowmeter and can be adjusted by modulating the pump's operating frequency through an inverter.

Temperature and pressure measurements of both product and heating water are continuously monitored at the inlet and outlet sections of the heat exchanger. Temperature measurements are performed using resistance temperature detectors (T in Figure 1), whereas pressure is measured using flush diaphragm pressure transmitters (P). All the data acquired during experimental tests were recorded every 0.5 seconds with a LabView NI 9208 module.

Water temperature is adjusted via indirect steam heating in a dedicated heat exchanger, while the flow rates of both steam and water are regulated by two control valves (denoted as Control Valve 1 and Control Valve 2 in Figure 1). These valves can be adjusted both automatically, using a Proportional-Integral (PI) control system to maintain product and water temperatures near

predefined setpoints, or manually, by specifying a fixed valve opening percentage.
The facility is integrated with auxiliary subsystems supplying steam, electrical power, water, and compressed air to the various operational units. Process control is implemented via a Programmable Logic Controller (PLC), with parameter adjustments accessible through a Human-Machine Interface (HMI) on the control panel.

**Product and rheology**

The tests were carried out by using the tubular heat exchanger to heat orange juice. The circulated juice was obtained by diluting a concentrated product, initially stored in 20 L bag-in-box containers, from an initial concentration of 60 Brix degrees to 40 Brix. In particular, 100 L of juice at 60 Brix was diluted to 40 Brix by adding 50 L of water. Before beginning the experimental campaign, the two fluids were accurately mixed to ensure homogeneity. To reproduce the tests via simulation, and gain valuable insights about the experimental outcomes, the rheology of the fluid was characterized using a rotational Anton Paar MCR 102 rheometer at different shear rates ranging from 10 to 300 $s^{-1}$, capturing the significant shear rate ($\dot{\gamma}$) values usually encountered in pipe flow applications (Steffe, 1996). The obtained apparent viscosity ($\eta$) data were fitted with a power law model, to obtain the consistency ($K$) and behavior ($n$) indices characterizing the non-Newtonian behavior of the fluid according to the relation (Equation 1) described in Steffe (1996).
The analysis was performed at three temperatures, i.e., 298.15 K, 323.15 K, and 348.15 K, to capture the temperature-dependent behavior of the apparent viscosity. These tests were then used to calculate the value of the activation energy $E_a$ of the Arrhenius equation, where $\eta_\alpha$ is the apparent viscosity at the reference temperature $T_\alpha$=348.15 K, and $R$ is the universal gas constant (Equation 2).
At $T_\alpha$ the following indices were obtained: $K$= 0.018 Pa $s^n$, $n$=0.6354. Since $n$<1, a non-Newtonian shear-thinning behavior emerges, as it is common in the case of food fluids and fruit juices. Moreover, the value of $\frac{E_a}{R}$ was calculated equal to 3898 K.

$$\eta = K\dot{\gamma}^{n-1} \quad (1)$$

$$\frac{\eta}{\eta_\alpha} = \exp\frac{E_a}{R}\left(\frac{1}{T} - \frac{1}{T_\alpha}\right) \quad (2)$$

**Experimental campaign**

The experimental procedure was designed to validate the temperature profile of the product at the outlet of the heat exchanger. To this end, the tank was filled with 150 L of fruit juice at 40 Brix, at an initial temperature of approximately 25°C. The heating water was brought to an initial temperature of 75°C, then the product circulation was activated. The product inlet and outlet temperature, as well as the water inlet temperature and the product flow rate, were continuously monitored throughout the entire product heating phase.
Regarding the product inlet and outlet temperature, two PT100 RTD sensors, having a measuring range between -50°C and 100°C were used. According to the technical characteristics of the sensor, the active part of the stem in temperature detection is 2 cm at the end of the probe. In light of this, to track the temperature profile of the product at the outlet of the heat exchanger, the temperature was measured at two different points by changing the position of the temperature probe. In the first position the 2 active centimeters of the stem were placed exactly at the center of the section (1 cm above the axis and 1 cm below the axis); in the second position they were placed at the bottom of the section (avoiding direct contact with the wall so as not to influence the measurement with metal temperature).
The test ended when the product temperature approximately reached the water temperature.
Some operating snapshots that occurred during the tests were then reproduced with fluid-dynamic simulations, and the results were compared in terms of the temperature profile of the product exiting the heat exchanger.

**Simulation settings**

The CFD simulations of the plant operation were carried out with ANSYS Fluent using the model presented and validated in Lysova et al. (2022). Specifically, the geometry reproduced half of the heat exchanger, exploiting its symmetry plane, and the mesh included 10 inflation layers, with a first layer thickness of $10^{-4}$ m, on each side of the heat exchange surface. Simulation settings are summarized in Table 1.
Regarding the definition of the treated product, the apparent viscosity was described using the results of the rheological characterization performed, while the physical properties were retrieved in the literature ($\rho$ [kg/$m^3$] = 1336 – 0.5286·T (Garza and Ibarz, 2010); $\lambda$ [W/m K]=0.1064 + 0.0012·T (Zhang et al., 2011); $c_p$=3114.4 [J/kg K] (Moresi and Spinosi, 1980)).
The simulated operating point was defined to reproduce a condition encountered during the experimental campaign, in order to validate the obtained results (product entering at 300.15 K and water at 333.15 K).
To evaluate the temperature profile calculated, and compare it with experimental data, the temperature was calculated across a vertical line, passing through the center of the pipe in the same position as the temperature probe in the real plant (Figure 2). To reproduce the readings of the probe in the central position, the average of the values of the central 2 cm of the pipe was calculated, while for the bottom section, the lower 2 cm was considered, starting 0.5 mm from the wall to reproduce the experimental set-up.

Table 1: Simulation settings

| | | |
|---|---|---|
| General settings | Solver | Pressure-based Steady-state |
| | Gravity | -9.81 m/s$^2$ in the y direction (see Figure 2) |
| | Turbulence model | k-$\omega$ SST |
| Boundary conditions | Product inlet | Velocity inlet 0.55 m/s 300.15 K |
| | Water inlet | Velocity inlet 1.3 m/s 333.15 K |
| | Product and water outlets | Pressure outlet |
| | Faces on the cutting plane | Symmetry condition |
| | Walls (all) | No-slip condition |
| | Interface between water and product | Coupled two-sided steel wall with 1.5 mm thickness |

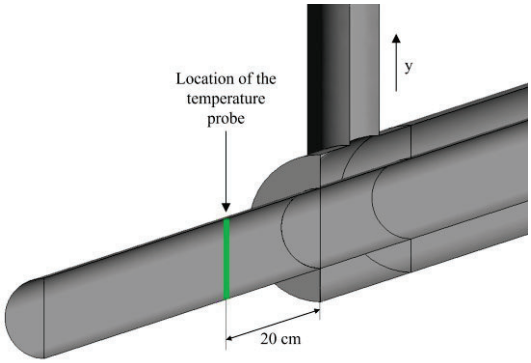



Figure 2: Line where the simulation results were evaluated, corresponding to the position of the temperature probe installed on the plant

## RESULTS AND DISCUSSION

Figure 3 presents the temperature profile calculated across the line presented in Figure 2. As expected, it can be clearly seen how there is variability across the pipe section, with higher values near the pipe walls, and lower values in the pipe center where the heat is transmitted more slowly. This behavior is due to the high concentration of the product, and therefore higher apparent viscosity, as discussed in (Lysova et al., 2024). Moreover, the top section of the pipe presents higher temperatures compared to the bottom.

The simulation was run reproducing the conditions of the experimental campaign, in order to validate the obtained results and the hypotheses about the temperature profile of the viscous product. The comparison between the simulated (in green) and experimental (in black) results can be seen in Figure 4. The error in the values is 0.23 K for the bottom measurement and 0.21 K for the central measurement, equal to 1.31% and 1.22% of the measured temperature increase in that location, respectively. Apart from the considerations about the error, which may be linked to the definition of the product characteristics and the experimental measurements, it can be observed that the trend of the temperature differences between the two locations is confirmed. This highlights the necessity of deeply investigating the heating of viscous products, without relying on a single value measured by a sensor.

To this end, the simulation results were used to investigate the temperature profile in the final section of the heat exchanger (Figures 5 and 6). In the sections where the product is still in indirect contact with the heating medium, the non-uniformity in the temperature appears particularly pronounced. Proceeding towards the outlet, as the heating ceases, the differences decrease. The heat transfer and gradual changes of the temperature profile are shown through contour plots in Figure 5, and plotted more in detail in Figure 6, where the values calculated across the pipe section at several locations, starting from the end of the heating zone, are presented.

In Figure 6a it can be seen how the non-uniformity in the temperature is particularly relevant in the heating section (0 cm plot) and in the immediately adjacent locations (5 and 10 cm plots).

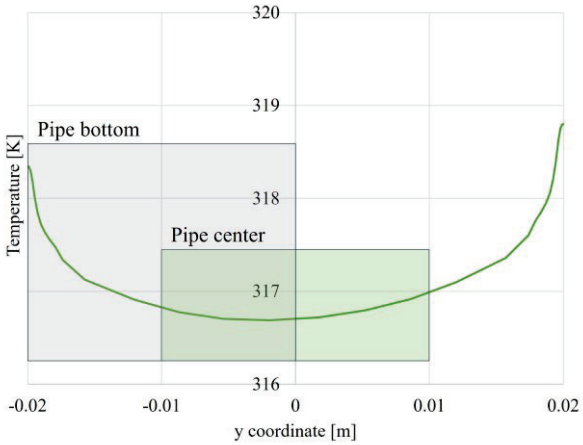



Figure 3: Temperature profile at the probe location. Highlighted in green are the data corresponding to the measurement in the pipe center, while in grey those related to the measurement of the bottom section.

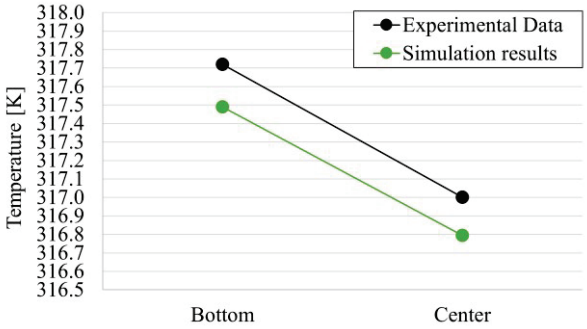



Figure 4: Comparison between the simulation results and the experimental measurements

On the other hand, the plot reported in Figure 6b shows the heat penetration from the near wall region towards the center of the pipe as the fluid moves away from the heating medium.
In particular, it is evident how the temperature in the center of the pipe increases while that in the near-wall region gradually decreases, reaching an almost uniform distribution after 0.5m.

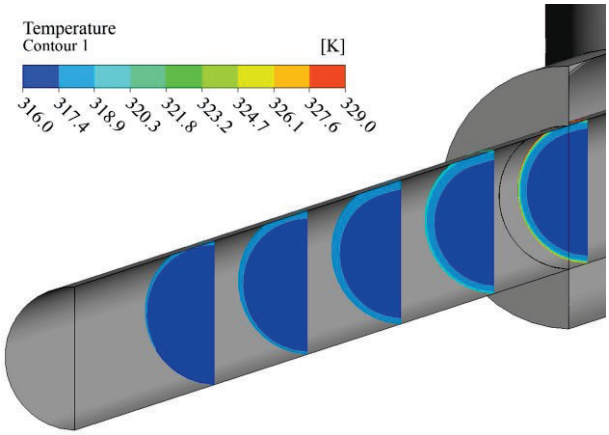



Figure 5: Contour of temperature calculated on 5 planes. The central plane is relative to the position of the probe, while those that precede and follow it are placed at a 5 cm distance between each other

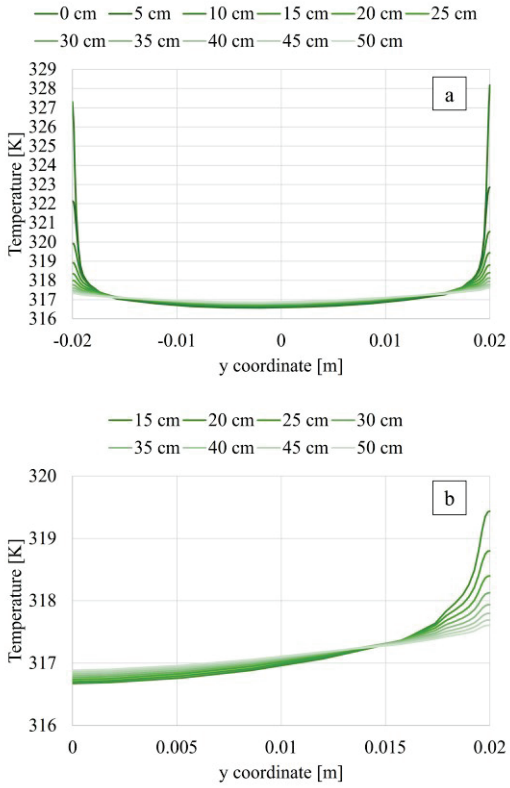



Figure 6: Temperature profiles at different distances from the end of the heating zone a) across the whole pipe section and b) across the upper half

## CONCLUSIONS

During the thermal treatment of fluid foods, especially those characterized by high viscosity, and therefore by reduced convective motions, the temperature distribution within the product during, and immediately after the heat treatment, suffers a certain degree of non-uniformity. This phenomenon impacts the performance of the process, both in terms of minimum and maximum thermal treatment achieved. The former could result in food safety issues, while the latter in problems related to fouling or food degradation.
In the present work, the fluid dynamics simulation capability in predicting the temperature profile at the outlet section of a heat exchanger was investigated. The results were validated through experimental tests conducted on an industrial-scale tube-in-tube heat exchanger using concentrated orange juice at 40 Brix. Experimental and numerical results agreed in predicting temperature variations across pipe sections located in the proximity of the outlet of the heat exchanger (i.e. the temperature in the near-wall regions was consistently higher than the temperature at the center of the section). Simulation results demonstrate that this difference is even more pronounced inside the heat exchanger, where the product is in direct contact with the hot surfaces, which are heated by the heating medium.
These results suggest the potential for the development of simulation-based virtual sensors based on the data measured by the physical sensor installed at the exit of the heat exchanger and the operating conditions. These virtual sensors would allow the real-time estimation of the temperature at each point within the heat exchanger, even where the installation of physical sensors would be very impractical, if not impossible. Moreover, the insights obtained through dedicated simulation campaigns, combined with the development of virtual sensors, would allow for the process to be adjusted to avoid localized time-temperature combinations that could cause fouling and/or sensory and organoleptic product degradation at any point of the heat exchanger.
Future research activities will focus on the development, testing, and validation of virtual sensors based on simulation results and their integration in Digital Twins and advanced control frameworks of industrial plants. In particular, the sensors will be developed using statistical techniques to interpolate simulation data calculated across a significant operating range. Finally, to test and confirm broader applicability, the described methodology will be applied to different viscous fluids (e.g., tomato paste or dairy products) and different kinds of heat exchangers (e.g., concentric channel, shell and tube, and plate heat exchangers).

## REFERENCES


Bottani, E., Vignali, G., & Carlo Tancredi, G. P. (2020). A digital twin model of a pasteurization system for food beverages: tools and architecture. *2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1–8. https://doi.org/10.1109/ice/itmc49519.2020.9198625

Garza, S., & Ibarz, A. (2010). Effect of Temperature and Concentration on the Density of Clarified Pineapple Juice. *International Journal of Food Properties*, 13(4), 913–920. https://doi.org/10.1080/10942910902919596

Guzmán, C. H., Carrera, J. L., Durán, H. A., Berumen, J., Ortiz, A. A., Guirette, O. A., Arroyo, A., Brizuela, J. A., Gómez, F., Blanco, A., Azcaray, H. R., & Hernández, M. (2018). Implementation of Virtual Sensors for Monitoring Temperature in Greenhouses Using CFD and Control. *Sensors*, 19(1), 60. https://doi.org/10.3390/s19010060

Han, H.-Z., Li, B.-X., Wu, H., & Shao, W. (2015). Multi-objective shape optimization of double pipe heat exchanger with inner corrugated tube using RSM method. *International Journal of Thermal Sciences*, 90, 173–186. https://doi.org/10.1016/j.ijthermalsci.2014.12.010

Jha, A., Moses, J. A., & Anandharamakrishnan, C. (2019). Optimizing Beverage Pasteurization Using Computational Fluid Dynamics. Preservatives and Preservation *Approaches in Beverages*, 237–271. https://doi.org/10.1016/b978-0-12-816685-7.00008-2

Kola, P. V. K. V., Pisipaty, S. K., Mendu, S. S., & Ghosh, R. (2021). Optimization of performance parameters of a double pipe heat exchanger with cut twisted tapes using CFD and RSM. *Chemical Engineering and Processing - Process Intensification*, 163, 108362. https://doi.org/10.1016/j.cep.2021.108362

Liu, S., Huang, W., Bao, Z., Zeng, T., Qiao, M., & Meng, J. (2021). Analysis, prediction and multi-objective optimization of helically coiled tube-in-tube heat exchanger with double cooling source using RSM. *International Journal of Thermal Sciences*, 159, 106568. https://doi.org/10.1016/j.ijthermalsci.2020.106568

Lysova N, Solari F, Vignali G. Optimization of an indirect heating process for food fluids through the combined use of CFD and Response Surface Methodology. Food and Bioproducts Processing 2022;131:60–76. https://doi.org/10.1016/j.fbp.2021.10.010.

Lysova, N., Solari, F., Bocelli, M., Rizzi, A., & Montanari, R. (2024). Exploring the impact of the process parameters on the thermal treatment of viscous food fluids in a tube-in-tube heat exchanger. *Procedia Computer Science*, 232, 2347–2357. https://doi.org/10.1016/j.procs.2024.02.053

Moresi, M., Spinosi, M. (1980). Engineering factors in the production of concentrated fruit juices 1. Fluid physical properties of orange juices. *International Journal of Food Science & Technology*, 15(3), 265–276. Portico. https://doi.org/10.1111/j.1365-2621.1980.tb00939.x

Pereira, R. N., & Vicente, A. A. (2010). Environmental impact of novel thermal and non-thermal technologies in food processing. *Food Research International*, 43(7), 1936–1943. https://doi.org/10.1016/j.foodres.2009.09.013

Perone, C., Romaniello, R., Leone, A., Catalano, P., & Tamborrino, A. (2021). CFD Analysis of a Tubular Heat Exchanger for the Conditioning of Olive Paste. *Applied Sciences*, 11(4), 1858. https://doi.org/10.3390/app11041858

Roobab, U., Aadil, R. M., Madni, G. M., & Bekhit, A. E. (2018). The Impact of Nonthermal Technologies on the Microbiological Quality of Juices: A Review. *Comprehensive Reviews in Food Science and Food Safety*, 17(2), 437–457. Portico. https://doi.org/10.1111/1541-4337.12336

Steffe, J.F., 1996. ch.1, Introduction to Rheology. Rheological Methods in Food Process Engineering. Freeman Press, pp. 1–91.

Tancredi, G. P. C., Vignali, G., & Bottani, E. (2022). Implementation of a remote control and system monitoring via a Digital Twin on an industrial pilot plant. *IFAC-PapersOnLine*, 55(10), 1128–1133. https://doi.org/10.1016/j.ifacol.2022.09.541

Yolmeh, M., & Jafari, S. M. (2017). Applications of Response Surface Methodology in the Food Industry Processes. *Food and Bioprocess Technology*, 10(3), 413–433. https://doi.org/10.1007/s11947-016-1855-2

Zhang, M., Zhao, H.-Z., Zhang, L.-J., Yang, L., Zhong, Z.-Y., & Chen, J.-H. (2011). Experimental Study on Thermal Conductivity of Orange Juice Using Probe System. *International Journal of Food Properties*, 14(1), 134–144. https://doi.org/10.1080/10942910903147965


## AUTHOR BIOGRAPHIES


**NATALYA LYSOVA** is a Ph.D. Candidate in Industrial Engineering 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 SOLARI** is a researcher and lecturer at the Department of Engineering for Industrial Systems and Technologies 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 57 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

**LEONARDO RESTORI** is a Master's student at the University of Parma. After a Bachelor’s Degree in Mechanical Engineering, he is now concluding his studies in Engineering for the Food Industry and this work will be part of his dissertation.

**ROBERTO MONTANARI** is a Full Professor at the Department of Engineering for Industrial Systems and Technologies - 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.

# HiPUTS: Super-Scalable Simulation of Microscopic Continuous Urban Traffic Model

Mateusz Najdek[1], Natalia Brzozowska[2], and Wojciech Turek[3]

[1]Institute of Computer Science, AGH University of Krakow, e-mail: najdek@agh.edu.pl
[2]Institute of Computer Science, AGH University of Krakow, e-mail: nbrzoza@student.agh.edu.pl
[3]Institute of Computer Science, AGH University of Krakow, e-mail: wojciech.turek@agh.edu.pl

**KEYWORDS**

HPC, Simulation, Urban traffic, Scalability, Multi-agent systems

**ABSTRACT**

The presented HiPUTS system is a new urban traffic simulator, which aims at providing (almost) linear scalability of the distributed simulation computation up to several thousand computing cores of a modern supercomputer. The simulation system supports microscopic, continuous traffic models with fully transparent distribution of computations which does not affect simulation results. In this paper we present the architecture and crucial algorithms of the simulation system and evaluate its performance and scalability. The synthetic scalability evaluation includes both strong and weak scaling experiments. The paper is summarized with a set of design guidelines, which are important for obtaining super-scalability of distributed spatial simulations.

## 1 Introduction

Urban traffic is one of the most problematic aspects of the functioning of modern cities. Despite over a hundred years of motorization development and adaptation of road infrastructure in cities to its needs, the inhabitants of the vast majority of cities and towns experience problems with the smoothness of movement in road networks, even those newly designed. Interestingly, humanity is able to manage traffic in other types of networks, like railway and telecommunications networks. The autonomy and diversity of urban traffic users, the legacy of traffic managing methods and the growing size of road networks still pose a very important and interesting scientific challenge.

One of the key methods of studying phenomena occurring in urban road systems is their modeling and simulation. The first works in this field were published in the 1950s (Gerlough 1955). The first methods were based on modeling flows, without identifying specific vehicles. Over the next decades, research was moving towards more and more detailed microscopic models (van Wageningen-Kessels et al. 2015), which allow to represent the autonomy, diversity and unpredictability of the behavior of individual traffic participants. An obvious consequence of this direction is the growing complexity of models and algorithms for simulating changes in their state over time.

The need for large computing power and memory resulted in the search for methods of parallelizing computations and using multiple computers to simulate complex models. Many attempts of developing parallel versions of urban traffic simulations have been described over the last decades (Nagel and Schleicher 1994; O'Cearbhaill and O'Mahony 2005; Kanezashi and Suzumura 2015; Horni et al. 2016; Arroyo et al. 2018). The results clearly show, that the problem of urban traffic simulation using detailed microscopic models is not straightforward to parallelize due to several reasons. The computational task requires updating one common data structure (the model), using its previous state, which complicates the synchronization between computing workers. The simulation generates large volume of results, which cannot be processed in a centralized manner. Moreover, the size of computational task assigned to each worker changes over time as the simulated cars traverse the road network.

In this work, we present the architecture, the algorithms and their implementation in a new, distributed urban traffic simulation, called HiPUTS. All its elements were designed with distribution and scalability in mind, in order to allow effective utilization of thousands of computing cores of a modern High Performance Computing (HPC) machine. The proposed solution aims at providing two crucial features: scale invariance and distribution transparency. Scale invariance in this context is understood as the invariance of the task size of a single worker when increasing the entire computation with a proportional increase in hardware resources. This feature is the basis for the scalability of a distributed system by making the size of a worker's task independent of the scale of the problem. The distribution transparency is a feature of model processing algorithm in a distributed manner. The simulation model and the algorithm of its state changes over time should not be dependent on the presence or degree of distribution. In other words, the results obtained in the sequential and parallel versions should be identical.

In the next section the analysis of existing approaches towards parallel urban traffic simulation is presented. The details of the proposed architecture and algorithms are

described in Section 3. Systematic tests of the scalability of the developed implementation are described in the Section 4. The paper is summarized with the discussion of elements crucial for obtaining good scalability, presented in Sections 5 and 6.

## 2 Existing Scalable Traffic Simulations

In the late 20th century, the necessity for parallel execution in traffic simulation became apparent. The Nagel-Schreckenberg model, a widely utilized microscopic traffic model, was initially implemented in parallel by its creator back in 1994 (Nagel and Schleicher 1994). Subsequent research by the same group (Rickert and Nagel 2001) suggested that global synchronization in parallel traffic simulation algorithms might not be necessary, which is a key finding in the context of providing scale invariance. Eliminating all centralized components from a computations is crucial for achieving scalability on HPC hardware. However, transitioning to this approach in traffic simulations posed challenges.

Several endeavors to develop a centrally synchronized parallel traffic simulation have been reported in the recent decades. Research in (O'Cearbhaill and O'Mahony 2005) demonstrated scalability on a limited scale. The idea of extending the intervals between global synchronizations, as discussed in (Kanezashi and Suzumura 2015), resulted in improved scalability but also introduced notable errors in the simulation outcomes. In another approach outlined in (Toscano et al. 2012), a sophisticated protocol was proposed to rectify errors stemming from infrequent synchronizations.

One of the earliest successful attempts for defining a scale-invariant distribution method in a traffic simulation was outlined in (Xu et al. 2015). The authors introduced an Asynchronous Synchronization Strategy, where in each computational process interacts solely with a restricted subset of processes responsible for neighboring segments of the simulated environment. This strategy facilitated linear scalability with the involvement of 32 parallel workers.

Similar distributed architecture without centralized synchronization was used and extended in (Turek 2018), where traffic simulation scaled linearly up to 19200 parallel workers executed on 800 computing nodes of a supercomputer, proving the possibility of running HPC-grade traffic simulations. The simulation was able to simulate over 11 million cars with running speed at 160 steps per second, which is one of the largest scenario run at that time in the context of traffic simulations. The practical application of the solution was very limited - it used the discrete, Nagel-Schreckenberg model with a uniform grid of roads and it was not able to represent any real-life scenarios.

Using a simplified traffic model is one of the possible approaches, when simulation of large scenarios is needed. This approach was applied in one of the most popular traffic simulation tools nowadays, MATSim (Horni et al. 2016) (Multi-Agent Transport Simulation), which uses queues of cars to represent road lanes. Simplified models can provide high efficiency, however, they sacrifice accuracy and correctness in representing traffic behaviors. This trade-off is particularly relevant when simulating complex scenarios where small details can have a substantial impact on traffic flow. In terms of improving the efficiency, up to this point the main focus in MATSim was put on improving the existing engines to work as optimally as possible but still within a single process and a single computing node, utilizing concurrency and parallelization on multiple threads and parallel queues within MATsim internal architecture (Dobler 2010). Unfortunately, such an approach revealed its obvious limitations. MATSim is full of features and additional functions added during its long development, which increase the complexity of computations. Therefore, the simulation of complex scenarios is slow anyway and parallelization seems a desired feature now. Unfortunately, at this point it is hardly possible due to internal architecture and assumptions of the system.

Another very popular traffic simulation tool is SUMO (Acosta et al. 2016) (Simulation of Urban MObility). It uses a continuous space and state models with several variants of decision and motion modelling, and many additional features. Running large scale scenarios with complex traffic models quickly exceeds the abilities of a single computer, therefore several attempts to improve overall performance of SUMO by distributing computations (Acosta et al. 2016; Chen et al. 2020) or using new strategies for synchronization (Arroyo et al. 2018) have been reported. In all cases achieving either very limited scalability or even results worse than the centralized version of the simulation. Another problem of mentioned solutions is also a growing problem of increasing error with increasing scale.

Probably the first advanced traffic simulator which considered distribution at the design stage was SMARTS(Ramamohanarao et al. 2016) (Scalable Microscopic Adaptive Road Traffic Simulator). It is able to simulate continuous traffic models, import real-life models and execute computations on many computing nodes. These promising features encouraged us to investigate it carefully and improve its scalability features in cooperation with SMARTS creators (Najdek et al. 2021). In (Zych et al. 2022) the original architecture of the simulator has been modified and the introduced improvements allowed to scale efficiently to hundreds of cores. Unfortunately, SMARTS is not well-maintained and also suffers from several issues in its assumption. The most important problem is the lack of distribution transparency, which may lead to inconsistent results. When a vehicle requires information of the proceeding vehicle in front, and the proceeding vehicle is not processed by a direct neighbor computing node, then the information cannot be obtained due to the fact that there is no communication channel between the two nodes. This causes that distribution is not transparent, but depends on initial division of the model. This is a major flaw in this approach that should as well be addressed at the design stage.

## 3 The High Performance Urban Traffic Simulator

The design process of creating a new urban traffic simulation system was guided by four main aims:

- Support for continuous traffic models with the ability of extending the solution with new models in the future,
- Distribution transparency for ensuring equivalency of results in single- and multi-node execution.
- Super-scalability achieved by scale invariance of computations executed by distributed nodes and load-balancing mechanisms.
- Ability to simulate large-scale and real-life scenarios by importing data from open map services.

These aims led us to developing HiPUTS, the High Performance Urban Traffic Simulator. HiPUTS is implemented in Java and it is available for download at `https://github.com/hiputs/hiputs`. It is a fully functional tool, which can be used for simulating large-scale scenarios, efficiently utilizing distributed computational resources. The environment models can be imported from the Open Street Maps[1] format. The traffic can be generated randomly or according to specified rules. The simulation can represent single- and multi-lane roads, vehicles of different types, crossroads with and without traffic lights and many other extensions, including visualization, results collecting and aggregation, modelling of pedestrian crossings or public transport. While these extensions demonstrate the extensibility of the solution, they are not crucial in the context of scalability and distribution transparency, discussed in this paper.

### 3.1 Simulation models

The environment model consists of two main types of elements: lanes and junctions. The model is represented in the form of a directed graph with edges representing the lanes between the nodes, which represent the junctions as shown at figure 1. Each node can be connected by either one-way (fellow) or two-way (blue) road. Imported data is converted to the graph of lanes and junctions. During the import, several filtering and optimization operations are performed, which are required due to the quality of the source data. There are two types of junctions: bends and crossroads, where bends connect one or two lanes, while crossroads connect more. Each junction has a fixed, known location and each lane represents a straight line between the junctions. As a result, each lane has a fixed, known length and each junction knows the circular order of incoming and outgoing lanes. From the vehicle's perspective it contains information about its current lane and position on it, current speed, acceleration and others, such as its length, max speed of vehicle and route for this car, so it knows where to move.

[1] https://www.openstreetmap.org

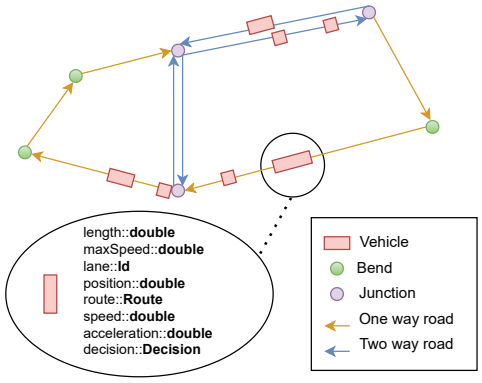



Figure 1: Representation of vehicle model and graph of environment

Each lane contains a list of cars currently present in the lane. Cars state is updated according to decisions generated by a set of so called *deciders*. Each car uses a basic car-following decider (which implements the IDM model (Treiber et al. 2000)) and a junction decider, which implements the right of way on crossroads. There are several other extensions which add the traffic model with multiple lanes, the MOBIL lane change model (Kesting et al. 2007), traffic lights on crossroads and overtaking on single-lane roads.

### 3.2 Distribution transparency

The environment model needs to be distributed between computing nodes. The road graph is divided into small area parts, which are the basis for ensuring distribution transparency and load balancing. Each computing node needs to contain information about its working area and information about adjacent parts and neighboring areas that are required for the correctness of the simulation. This approach is kind of based on a stencil pattern, but differs in the form of storing data, as the simulation data and state are not globally stored but distributed among computing units instead. Such approach allows also to be scalable in the context of data management. We are not only limited to a single node specification. Instead dividing the environment allows to work only on specific part of the simulation environment, resulting in speeding up overall computation and reducing memory usage overhead, as the result allowing also to simulate very large areas with heavy congestion.

Ensuring correctness in the distribution of the environment model is crucial. Its distribution cannot affect computation itself in any form or way, meaning that it should be transparent from the simulation point of view. Unfortunately, there are examples of existing solutions (Ramamohanarao et al. 2016) that do not address this issue so closely causing simulation to lose required information about e.g. vehicles during the communication process.

The Figure 2 presents the problem in the distributed simulation caused by the exchange of insufficient informa-

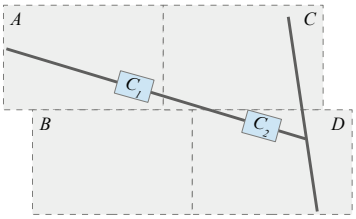



Figure 2: The problem of communication need between non-neighbour fragments

tion between adjacent nodes. Each computing node ($A$, $B$, $C$, $D$) simulates part of the environment, but $A$ and $D$ in this example are not direct neighbors. The simplistic model division boundary is shown by dashed lines in the mentioned example. Each node processes an assigned fragment of the environment and it needs to exchange information about the state of the fragment's border with the nodes responsible for adjacent fragments. However, determining how the border should be drawn is a challenge – it should be limited in order to reduce the communication overhead, however it has to provide all the necessary information. Even when the size of the border is properly estimated, errors still can occur when the arrangement of fragments causes the need for accessing data from indirect neighbors. This problem has been depicted in Fig. 2, where an improperly thought out division can cause problems. In this example, the car $C1$ is simulated in fragment $A$ by one process and needs information about car $C2$ that is simulated in fragment $D$, but it is not a direct neighboring process. From the car $C1$ point of view it should contain information about its proceeding vehicle, but due to this division it will not obtain it, as the edge on which the vehicle is moving is cut by the process $C$ (direct neighbour to $A$) and it does not contain information about $C2$ as it is only in $D$. For this reason process $A$ will not obtain information about $C2$ causing error in simulation and resulting in different outcome.

We propose to solve these kind of problems by introducing the notion of *patches*. The road network is partitioned into disjoint fragments, called patches, which should be possibly small and guarantee the following feature: every decision algorithm in the simulation is based on the state of car's current patch and its direct neighbours only (allowing to maintain a required amount of neighbors to a minimum).

The proposed algorithm for partitioning the road network into patches is to use the grid composed of regular hexagons. To define the mesh, it is necessary to provide its orientation in a specific coordinate system and the length of its edge. The edge of each hexagon has to be equal to the maximum observation range of all decision algorithms. The indexation in the grid is shown in Fig. 3.

The geometric division of the road network is therefore determined by the coordinates of point A and the length of the side of the hexagon. The next step is to apply this division to the graph structure. Each node in the network model can be uniquely assigned to the interior of one

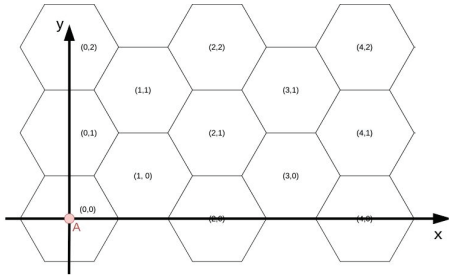



Figure 3: An example mesh of hexagons centered at point A along with their indexation

Figure 4: Exemplary division of a road network into patches

hexagon in which it is located. Such an assignment can be performed in linear time with respect to the number of vertices in the graph. An example of patches in a real-life network is shown in Fig. 4.

The proposed hexagon-based algorithm is not the only solution to the task of creating patches for a road network. However, it is definitely a simple and efficient solution for addressing this problem.

### 3.3 Parallel execution and scale invariance

The grid of patches can also be considered a graph, where each patch is a node with up to 6 neighbours. This graph is used for dividing the simulation task between the available computing nodes. Each of the patches is assigned an estimated computational cost, calculated as the total length of lanes in the patch. Then the graph of patches is partitioned using the k-means (Ahmed et al. 2020) clustering algorithm.

At the beginning of the simulation each computing node receives a list of patches to load. From the point of view of a one particular node, the patches processed by it are divided into three sets: internal patches, border patches and remote patches. The internal and border patches are updated by the node. The border patches have adjacent patches in the model, which are not processed by the current node. These adjacent patches are called the remote patches. An example of the types of patches is presented in Fig. 5

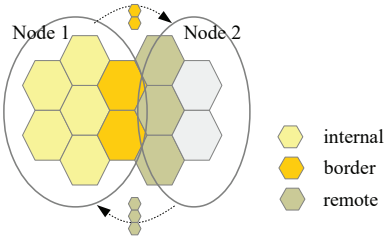
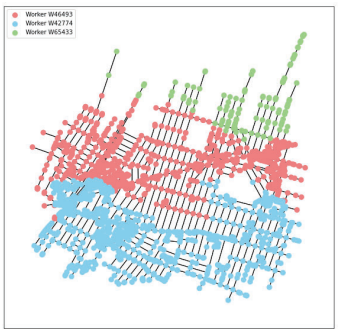


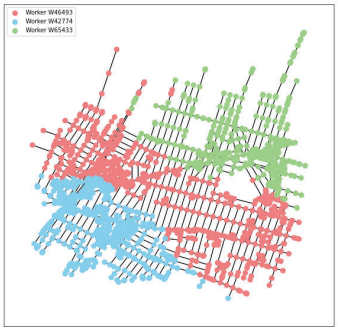

Figure 5: Internal, border and remote patches of exemplary *Node 1* and the exchange of patches' states between nodes

During a simulation step each node sends the state of border patches to the nodes responsible for adjacent remote patches. In return it receives the state of the remote patches (which are actually the border patches of the other nodes). The state of the remote patches contains all the information needed for computing decisions of all cars in the internal and border patches.

The initial partitioning of work between the available nodes cannot guarantee load balance throughout the simulation. The computational cost of updating a patch is not dependent only on the length of its lanes, but rather on the the number of cars currently present in the lanes. This dynamic load imbalance problem requires using a load balancing algorithm, which in our solution is based on exchanging patches between nodes.



(a) Exemplary traffic network during 1st simulation step



(b) Exemplary traffic network during 200th simulation step

Figure 6: Results of the load balancing mechanism on traffic network.

Each node measures the time needed for perform a simulation step computations and exchanges the value with its neighbours. When a significant difference is detected, a border patch is selected and sent to the less-occupied node. The selection of patches to send aims at preserving the integrity of the map fragments processed by the node, which is not always straightforward. The exchange is allowed once in a few steps in order to prevent rapid oscillations. In addition, only one source of transferred patches is allowed in a single step to prevent overloading the less-occupied node by sending it patches from several different nodes at a single step. These three mechanisms turned out to be sufficient to provide satisfactory balancing even in complex real-life scenarios, although there is still field for improvements in this area. A simple example of load balancing results are presented in Fig. 6a and 6b.

### 3.4 Simulation algorithm

The simulation calculations are carried out by workers who handle a specified number of simulation steps. The simulation can be divided into many workers, each one running on a separate computing node. The main simulation algorithm of a single simulation step can be described in a few stages. The overview is presented in Figure 7.

During each step, workers update vehicle parameters to reflect how they would realistically change over a time span equivalent to that step's duration. Neighboring workers may sometimes be at different stages of calculation or working on different time steps, but these discrepancies are short-lived and are resolved at local synchronization points, typically through message exchanges. This eliminates the need for centralized synchronization, which often creates bottlenecks in this type of computation. A given worker can proceed with simulation locally for the next $n + 1$ step if all of its neighbors have completed the given state of the current $n$ step and sent the state of their changes from $n$ step to the mentioned worker.

The entire algorithm is conceptually divided into two main stages that involve making decisions by vehicles located on lanes and then updating those vehicles' states. This solution was introduced to assess that simulation could be divided into fine-grained level tasks and parallelized. The first stage (decision) allows to perform calculations of traffic models for every vehicle on lane without making any physical changes to vehicles in the environment. Decision allows vehicles to store data about its state in the following $n+1$ step and not modify its current state in $n$ step. This ensures that each vehicle makes decision based on the same state of environment in $n$ step. Once all vehicles finish this stage they can be actually updated by moving them to the appropriate position on the lane and applying other changes to their state in the next iteration step of simulation, this can also be performed similarly to the previous stage in a parallel manner.

The first part of the discussed algorithm is that the vehicle determines its next state based on information from the current step. The decision is expressed by factors like positive or negative acceleration, speed, route changes,

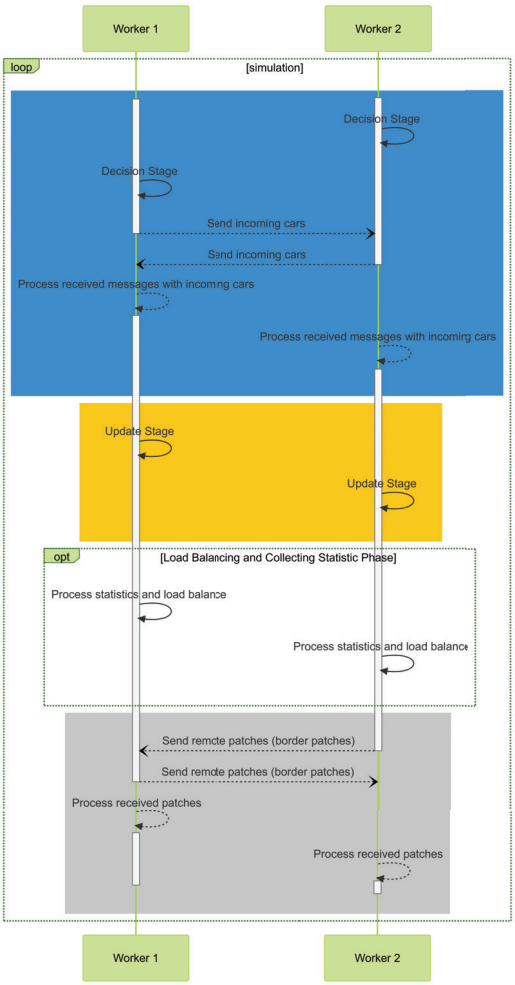



Figure 7: HiPUTS simulation algorithm overview of single step with 2 workers.

the lane the vehicle will occupy, and its position on that lane. The decision within itself involves applying different models (such as IDM, etc.) by taking into account vehicle's current state and the preceding vehicle's state. The decision is essentially a reflection of the vehicle's future state for the next iteration step and it is stored and managed by each vehicle to be used in the following part of the simulation stage (known as the update stage).

The decision part of the algorithm is divided into fine-grained tasks. Each task corresponds to a specific lane separately and is responsible for managing all vehicles on that lane within this task. Those tasks are then executed in parallel on a shared thread pool, with a size equivalent to the available computing cores on the node. This mechanism speeds up calculations by parallelizing execution of those tasks and utilizing all available computing resources – all cores of the computing node. During this stage, vehicles that decide to move to another lane are added to the buffer of their destination lane. This buffer is used to ensure synchronized access to shared lanes among different tasks in the thread pool.

Once the decision-making stage is completed, the worker immediately begins dispatching vehicles located on the boundary lanes that separate his work area from neighboring workers. When all workers finish sending to every neighbor within their work area vehicles to their remote buffers of each border lane then they await to receive all messages for this part in the background and start to process them. The message sending is done asynchronously which is shown with a dotted line in Figure 7. After messages are sent, worker loses control (what is visualized with a block on the timeline) up to the point when all incoming messages from all direct neighbors are received. The whole part including decision stage is highlighted with a blue background. Once the worker sends these messages to each of its neighbors, it awaits status updates from each neighbor. After that, it can freely proceed further to update stage of the simulation. This is due to the fact that from a single worker's perspective, it has all the required knowledge to proceed further in the simulation. The update stage involves applying previously made decision and changing the vehicles's state appropriately for this simulation step. The stored decision of each vehicle is actually applied to it, changing its current state. This is another task which is parallelized in the same manner as the decision stage.

The following part of algorithm contains an optional feature that involves load balancing and uses statistical data of monitored worker performance parameters. It was described in more detail in subsection 3.3 .

The last part involves sending the state of border patches after they are changed due to the update phase. They are shared to ensure correctness and to spread required knowledge of environment after the update stage. For this reason, these updated patches need to be sent to other computing nodes. From the perspective of another worker, it updates the state of its remote patches using the received state of neighboring border patches. Once all these stages are complete the simulation can proceed to the next iteration step of the simulation loop. This part is marked in grey in the algorithm overview shown in Figure 7.

## 4 Evaluation

In order to investigate benefits offered by the described simulator architecture, experiments based on weak and strong scalability laws were prepared. However, for examination of HiPUTS' super-scalability feature, in particular it's scale invariance requirement, an evaluation conducted in accordance with Gustafson's law (Gustafson 1988) was more essential (weak scalability). This is due to necessity to ensure equal amount of work per each worker while both, number of processors and problem size, increase. Such demand may indeed result in highly satisfying scalability outcome.

Test scenario in both experiments was based on a synthetically generated traffic network, comprising perpendicular, intersecting roads, connected at the ends to form a torus. The roads were two-way and the distance between adjacent intersections was equal to 500 meters. Each patch included 4 connected intersections placed at the vertices

of a square. A single time step represented 1 second of real time. The traffic network had a constant number of cars during simulation (cars were regenerated when they reached their destination). For each scalability experiment two tests with different car density were prepared. In the first case average distance between cars was equal to 30 meters. In the second one, the distance was 60 meters. HiPUTS used load balancing strategy of transferring road network patches, by moving them from the most workloaded neighbour to the least burdened one. Each run computed 1500 simulation time steps. The presented performance results were computed based on the last 1000 steps to exclude strongly variable overhead associated with JVM initialisation. The majority of the tests were repeated at least 10 times. The only exceptions were executions using 1 and 96 nodes during strong scalability tests with vehicles an average of 60 meters apart, which were executed 8 and 4 times respectively due to limitations in hardware availability.

Experiments were executed using Ares supercomputer located in ACK Cyfronet AGH in Krakow, Poland, which is part of the PL-Grid infrastructure. Currently Ares is ranked 442th on the TOP500 list [2]. This supercomputer is composed of Xeon Platinum 8268 24C 2.9GHz processors connected by Infiniband EDR. In total Ares offers 37 824 computing cores. Experiments were performed using up to 96 nodes, each containing 48 cores. This gives a maximum number of 4 608 computing cores used. One simulator's worker was assigned to one node.

### 4.1 Weak Scalability

One aim of evaluation was to examine simulator's scale invariance, which assumes constant size of overall problem per worker regardless of their total number. To accomplish this, problem size needed to increase proportionally with number of workers. Therefore, as mentioned previously, weak scalability experiments were more relevant in this regard.

In the following tests each worker was initially responsible for a network comprising 65 536 intersections. Number of cars was equal to 2 184 534 or 4 369 067, depending on the experiment. Overall problem size changed as the number of workers increased, thus with 4 nodes traffic network grew to 262 144 junctions and over 17 million cars in the tests with 30 meter average distance between cars.

Results of total simulation time (fig. 8) show interesting trend where the increase in mentioned time stabilises after around 12 used nodes and only slightly grows when more workers are introduced. Reason for shortest computation time, when only 1 node was used is natural and related to no communication, synchronization and load balancing overhead, which also shows fig. 9. As the number of nodes increases, each worker has to interact with a greater number of neighbours. However, this relationship is no longer accurate since the use of 12 workers. The primary reason for this is constant initial number of neighbourhoods (equal to 4) when total number of workers is 12 or greater. Therefore, discussion about scale invariance of HiPUTS should start from this point, since this is when its requirements begin to be met.

[2] https://www.top500.org/lists/top500/list/2024/06

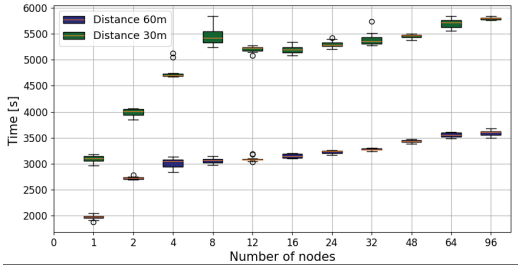



Figure 8: Total simulation time depending on used number of nodes

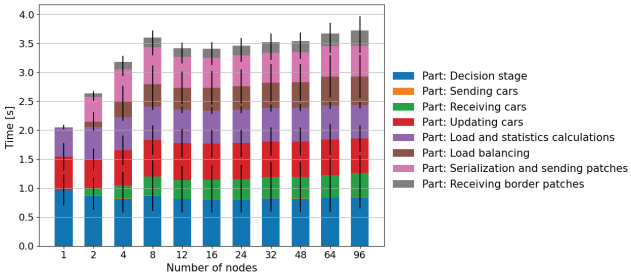
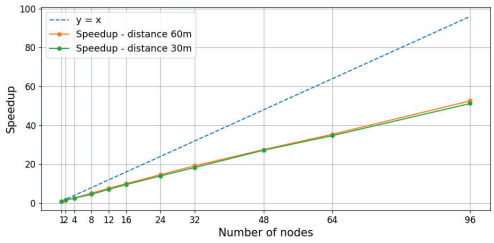



Figure 9: Average duration of one simulation step subdivided into stages (average cars distance equal to 30m)

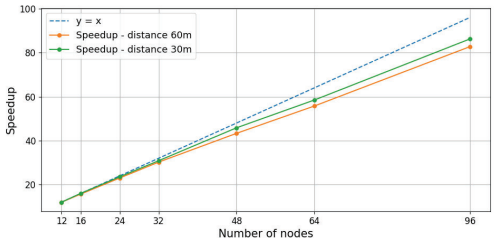
An analysis of average time spent on each part of simulation step (fig. 9) strongly suggests ensuring scale invariance by the simulator. Data show that since the use of 8 nodes, duration of particular calculation stages has remained highly close to constant, while number of nodes has increased. This observation follows that addition of another worker in HiPUTS does not affect performance of the other computing units.

The speedup achieved by HiPUTS simulator is decent, but far from ideal (fig. 10a). Metric calculation was based on the duration of execution with only 1 worker. Achieving similar simulation time while incorporating communication and synchronization processes is a challenging task. For this reason, obtaining of linear speedup while problem size increases and additional overhead occurs is difficult when related to performance on 1 node. However, calculating metrics from the point of ensuring scale invariance yields significantly different results and highlight the impact of this feature on performance (fig. 10b). Achieved speedup, obtained in relation to 12 nodes, is highly close to linear.

Described results clearly show that scale invariance ensures constant calculation time, resulting in almost linear speedup as both problem size and amount of computing power increase.




(a) in relation to single node (increased communication)




(b) in relation to 12 nodes (constant communication)

Figure 10: Speedup of HiPUTS simulator

### 4.2 Strong Scalability

Evaluation of HiPUTS performance included also experiments based on Amdahl's scalability law. During this ex-amination problem size was constant, thus parts assigned to each worker varied inversely proportional to the number of nodes used. Traffic network distributed among workers had a size of 1 048 576 junctions and contained 69 905 072 or 34 952 536 cars, depending on the task.

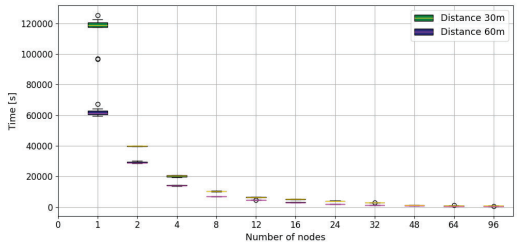



Figure 11: Total simulation time depending on used number of nodes

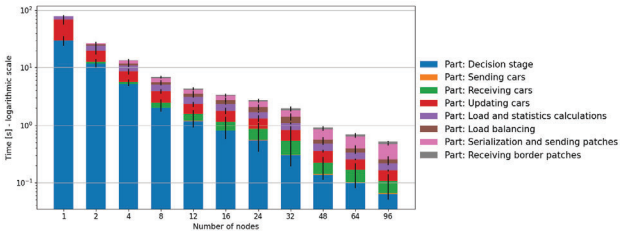



Figure 12: Average duration of one simulation step subdivided into stages (average cars distance equal to 30m)

As expected, when the number of nodes increases, the total simulation time and average duration of a simulation step decrease (fig. 11 and fig. 12).

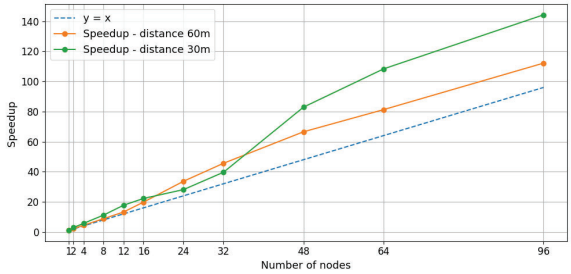



Figure 13: Speedup

However, from a certain point onwards (around 48 used nodes), the addition of more workers results in a diminishing time saving, which is consistent with Ahmdal's observations (Amdahl 1967). The decrease in computation duration between 1 and 2 nodes used is greater than linear in experiments with 30 meter distance between cars and close to linear in the other test. These results are surprising considering the additional overheads associated with communication and load balancing. Although these costs are relatively low for small number of nodes, their proportion increases as the number of nodes grows. The mentioned over-linear reduction of the simulation time results in a super-linear speedup (fig. 13).

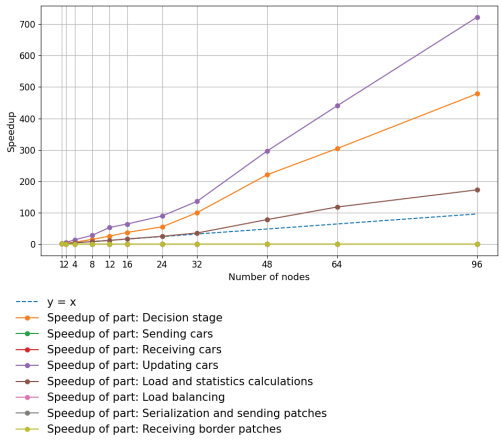



Figure 14: Speedup of each simulation part (average cars distance equal to 30m)

Analysis of the figure 14 reveals that this is due to the excellent scalability of only certain stages of the calculations - the decision stage, updating cars parameters step and calculations of load and statistics phase. All these are entirely computational stages and the two former are embarrassingly parallel tasks, which were executed in this manner. The remaining stages of simulation did not occur in the tests with 1 worker, thus they decelerate execution of the programme and are hidden on figure behind "Receiving border patches" line. The greatest speedup in experiment with 30 meters distance is achieved when the simulation was executed with more than 32 workers. Most probably this is due to the relatively small size of the problem, which resulted in fewer references to RAM and a less frequent need to reload the cache memory, than during the tests with 1 node.

Undoubtedly, HiPUTS simulator is highly scalable, mainly due to the suitable parallel implementation.

## 5 Towards Super-Scalability

Achieving proper scalability in large-scale distributed systems that iteratively modify a common data structure is a challenge. According to (Engelmann and Geist 2005), a super-scalable system must exhibit scale-invariance, meaning each computing unit maintains a constant workload under weak scaling. This includes not only computational tasks but also the volume of data to be processed, the number of communication partners, message volume, initialization and shutdown tasks, and results collection. Any violation of this principle can lead to bottlenecks and limit scalability.

There are several factors in a distributed urban traffic simulator that may violate scale-invariance. The HiPUTS system was specifically designed to address these challenges, focusing on various aspects of scalability:

*Computational task invariance* is relatively easy to ensure at the beginning of the simulation. Each patch of the model is assigned a computational cost estimate based on the total lane length and initial vehicle count. Workers are then assigned clusters of patches with similar total costs, ensuring an initially balanced workload.

*Load balancing* is necessary throughout the simulation to preserve an equal computational effort across nodes. As congestion builds up in specific areas, those areas require more computational resources. To manage this, a mechanism allows patches to be exchanged between neighboring workers. This introduces a minor communication overhead but prevents computational hotspots and ensures load distribution remains even.

*Communication invariance* is one of the most crucial factors for scalability. Even a single centralized communication point can create a bottleneck, limiting overall performance. To avoid this, HiPUTS is designed around local communication only. Each computing node interacts with a fixed number of neighbors, sending and receiving a consistent volume of messages, regardless of the overall size of the model. The only centralized step is the initial exchange of node addresses, after which synchronization occurs naturally through local interactions.

*Static model data* consists of the road network, junctions, and patches. In HiPUTS, each worker loads this data independently. While this approach does not fully comply with scale-invariance, as memory usage grows with the size of the simulation, it does not negatively impact efficiency once the data is loaded. On the contrary, it enables various optimizations, such as facilitating load balancing and distributing path planning. If memory constraints become an issue, these optimizations can be omitted.

*Path planning* can become a major scalability concern when traditional algorithms such as A* are used, as their complexity is dependent on the size of the graph. To mitigate this, alternative solutions can be applied. A simple random-walker approach allows vehicles to select their next junction randomly, eliminating the need for complex computations. When more structured or controlled traffic is required, paths can be precomputed for all vehicles and stored in files or a database. Since path generation can be parallelized, this method allows trajectories to be reused across multiple simulation runs, further improving efficiency.

*Parallelization within a computing node* is another critical factor for scalability. Even if computational loads are evenly distributed among nodes, poor multi-core utilization can degrade performance. In HiPUTS, all major simulation tasks, including inter-node communication, vehicle decision-making, and state updates, are divided into fine-grained tasks executed via a thread pool. This ensures efficient CPU utilization and prevents local computational bottlenecks.

*Graph partitioning* is an important preprocessing step required for efficient simulation. This process consists of two stages: generating the patch graph and partitioning it among computing nodes. While uniform grids simplify partitioning, real-world road networks introduce additional complexity. Fortunately, the patch graph is significantly smaller than the full road network, making partitioning a less demanding task. In HiPUTS, a k-means-based partitioning method is used, but more sophisticated techniques can be applied if needed. Since partitioning results are stored, they can be reused in subsequent simulations with the same model and computational resources.

*Results collection and processing* is another key consideration. Running a simulation without collecting output data is meaningless, but improper data aggregation can create performance bottlenecks. To ensure scalability, HiPUTS employs a decentralized approach in which each node independently records results and writes them to files. While this prevents real-time global statistics computation, it eliminates centralization and maintains performance. The collected data can then be analyzed post-simulation.

The design choices outlined above enabled HiPUTS to achieve a high level of scalability. Many of these solutions are not specific to traffic simulation and can be generalized to other large-scale distributed simulations, serving as valuable guidelines for building super-scalable systems.

## 6 Conclusions and Further Work

The created HiPUTS system demonstrates the possibility of building super-scalable simulations of microscopic continuous urban traffic models. The design of the system allowed achieving scale invariance and distribution transparency in this complex computational problem.

The proposed architectural concepts and algorithms can be used in other simulation tools, also outside the field of traffic simulation.

The next steps in the HiPUTS development will be the verification of its features during simulations of large-scale real-life scenarios. Detailed simulation models together with the ability of simulating large cases very fast may open new areas of applications for this type of tools.

**Acknowledgements**

The research presented in this paper was funded by the National Science Centre, Poland, under the grant no. 2019/35/O/ST6/01806. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016691

**References**

A. Acosta, J. Espinosa, and J. Espinosa. Distributed simulation in sumo revisited: Strategies for network partitioning and border edges management. In *Proceedings of the 4th SUMO User Conference*, pages 61–71, 2016.

M. Ahmed, R. Seraj, and S.M.S. Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

G.M Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference*, pages 483–485, 1967.

N. Arroyo, A. Acosta, J. Espinosa, and J. Espinosa. A new strategy for synchronizing traffic flow on a distributed simulation using sumo. *EPiC series in engineering*, 2:152–161, 2018.

H. Chen, K. Yang, S.G. Rizzo, G. Vantini, P. Taylor, X. Ma, and S. Chawla. Qarsumo: a parallel, congestion-optimized traffic simulator. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 578–588, 2020.

C. Dobler. Implementation of a time step based parallel queue simulation in matsim. In *10th Swiss Transport Research Conference. Ascona*, 2010.

C. Engelmann and A. Geist. Super-scalable algorithms for computing on 100,000 processors. In *International Conference on Computational Science*, pages 313–321. Springer, 2005.

D.L. Gerlough. *Simulation of Freeway Traffic on a General-purpose Discrete Variable Computer*. University of California, Los Angeles, 1955.

J. Gustafson. Reevaluating Amdahl's Law. *Commun. ACM*, 31(5):532–533, 1988. ISSN 0001-0782.

A. Horni, K. Nagel, and K.W Axhausen. Introducing matsim. In *The multi-agent transport simulation MATSim*, pages 3–7. Ubiquity Press, 2016.

H. Kanezashi and T. Suzumura. Performance optimization for agent-based traffic simulation by dynamic agent assignment. In *2015 Winter Simulation Conference (WSC)*, pages 757–766. IEEE, 2015.

A. Kesting, M. Treiber, and D. Helbing. General lane-changing model MOBIL for car-following models. *J. of Transportation Research Board*, 1999(1):86–94, 2007.

K Nagel and A Schleicher. Microscopic traffic modeling on parallel high performance computers. *Parallel Computing*, 20(1):125 – 146, 1994. ISSN 0167-8191.

M. Najdek, H. Xie, and W. Turek. Scaling simulation of continuous urban traffic model for high performance computing system. In *International Conference on Computational Science*, pages 256–263. Springer, 2021.

E. A O'Cearbhaill and M. O'Mahony. Parallel implementation of a transportation network model. *Journal of Parallel and Distributed Computing*, 65(1):1–14, 2005.

K. Ramamohanarao, H. Xie, L. Kulik, S. Karunasekera, E. Tanin, R. Zhang, and E.B. Khunayn. SMARTS: Scalable microscopic adaptive road traffic simulator. *ACM Trans. on Intelligent Systems and Technology (TIST)*, 8(2):1–22, 2016.

M. Rickert and K. Nagel. Dynamic traffic assignment on parallel computers in transims. *Future Generation Computer Systems*, 17(5):637 – 648, 2001. ISSN 0167-739X.

L. Toscano, G. D'Angelo, and M. Marzolla. Parallel discrete event simulation with erlang. In *Proceedings of the 1st ACM SIGPLAN workshop on Functional high-performance computing*, pages 83–92, 2012.

M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

W. Turek. Erlang-based desynchronized urban traffic simulation for high-performance computing systems. *Future Generation Computer Systems*, 79:645–652, 2018.

F. van Wageningen-Kessels, H. Van Lint, K. Vuik, and S. Hoogendoorn. Genealogy of traffic flow models. *EURO Journal on Transportation and Logistics*, 4 (4):445–473, 2015.

Y. Xu, W. Cai, H. Aydt, M. Lees, and D. Zehe. An asynchronous synchronization strategy for parallel large-scale agent-based traffic simulations. In *Proceedings of the 3rd ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pages 259–269, 2015.

M. Zych, M. Najdek, M. Paciorek, and W. Turek. Distributed architecture for highly scalable urban traffic simulation. In *International Conference on Computational Science*, pages 517–530. Springer, 2022.

# INCREMENTAL TRANSITION SYSTEM CONSTRUCTION WITH REFINEMENT FOR REGION-BASED PROCESS MINING

Shengrui Peng[1] and Helena Szczerbicka[1]

[1]L3S Research Center, Leibniz University Hannover, e-mail: {peng, hsz}@l3s.de

**KEYWORDS**

Region-based Process Mining, Incremental Algorithm, Transition System, Configuration Refinement.

**ABSTRACT**

Process discovery focuses on deriving process models from event logs to facilitate the analysis of complex workflows. Region-based approaches are known for their ability to effectively capture concurrency and handle invisible transitions, but they suffer from high computational demands and reliance on predefined state information. This paper introduces an incremental algorithm to construct *transition systems* (TS) from event logs, integrating a refinement strategy to optimize the representation of the state. The proposed method significantly reduces the complexity of the transition system by configuring state abstractions using past, future, or hybrid horizons. A refinement strategy is used to determine optimal configurations for state representation, ensuring model quality through precision and simplicity metrics. The approach is evaluated using a case study that demonstrates its effectiveness in improving scalability and efficiency in region-based process discovery.

## 1 INTRODUCTION

*Process mining* is a multifaceted field that bridges several disciplines, aiming at constructing precise process models from *event log* data to enable in-depth process analysis with the help of simulation. It is widely used to support applications in manufacturing and services where processes need to be continuously improved and adapted. *Event logs* can be understood as sequential records of operations performed. Modern information systems generate massive amounts of data, which poses significant challenges in efficiently constructing process models from these large datasets. Various *discovery algorithms* have been developed to address this challenge (Augusto et al., 2022), producing distinct types of process models, including *transition systems*, *Petri nets*, and *Business Process Modeling Notation* (van der Aalst, 2011). Among these approaches, the *region-based process discovery* approach is known to produce robust models capable of handling complex behaviors such as concurrency and invisible transitions. Petri net (PN) proposed in Petri (1962) is used to represent the process model specifically due to their solid mathematical foundations and ability to effectively represent concurrent processes.

Compared to other process discovery approaches, the main limitation of region-based algorithms is their high computational cost when handling large or highly variable logs. Typically, these algorithms need a transition system (TS) as an intermediate representation of the event log, from which a PN is derived using the *theory of regions* (Desel and Reisig, 1996). While many works focus on improving scalability and computational efficiency, a fundamental question is often overlooked, i.e., how to find a suitable (or even optimal) TS from a given event log. Since event logs generally do not provide state information, the first task is to extract it. Different abstractions can be applied to define states through their past (events before the state), future (events after the state), or both (van der Aalst et al., 2010). However, the modelers usually decide how these abstractions are combined. Thus, a systematic strategy is needed to identify a suitable configuration for representing states.

In this work, we propose an incremental algorithm for constructing transition systems from event logs to boost region-based algorithms. With abstractions, the equivalent states in *sub-TS*s can be identified and merged automatically, potentially reducing the size of resulting *super-TSs*. A *sub-TS* is a sequential unit TS derived from a single trace, and a *super-TS* is the TS constructed by integrating the *sub-TSs*. We then use two metrics, namely precision and simplicity, to evaluate and compare the *super-TS*s under different configurations. Finally, we introduce a refinement strategy to configure the abstractions.

This work is organized as follows. The section 2 reviews the literature, points out challenges and highlights the necessity of our approach. The section 3 covers fundamental definitions and concepts, especially abstractions applicable to event logs. The section 4 introduces our incremental algorithm and details the configuration refinement strategy. A numerical case study illustrating the efficacy of the proposed approach appears in section 5. Finally, section 6 provides conclusions and suggests future research directions.

## 2 LITERATURE REVIEW

The region-based process discovery approach originates from the Petri net (PN) synthesis problem proposed by

Communications of the ECMS, Volume 39, Issue 1,
Proceedings, ©ECMS Marco Scarpa, Salvatore Cavalieri,
Salvatore Serrano, Fabrizio De Vita (Editors) 2025
ISBN: 978-3-937 436-86-9/978-3-937 436-85-2
(CD) ISSN 2522-2414

(Desel and Reisig, 1996). The objective is to generate an *elementary PN* from a *(labeled) transition system* (TS), i.e., a concurrency model derived from sequential observations, using the *theory of regions* proposed in (Ehrenfeucht and Rozenberg, 1990). An *elementary PN* solves this synthesis problem *iff.* its reachability graph is isomorphic to the given TS. (Cortadella et al., 1995) also presented an early method for generating a safe PN from a (labeled) TS. Later, (Cortadella et al., 1998) introduced a broader class of (labeled) TS, called the excitation-closed TS. The solution is no longer restricted to elementary PN. Subsequently, region-based approaches are adapted for the *process mining problem* (see Definition 2). Typically, the event log is transformed into a TS, and then a PN is derived. One major issue in this approach is extracting state information from event logs, which provide only sequences of activities. In other words, the state information is not directly available from an event log.

An instance-based multi-step approach is proposed in (van Dongen and van der Aalst, 2004, 2005), focusing on constructing models at the instance level (i.e., *sub-TS*) and then aggregating them into an overall process model (i.e., *super-TS*). This enables analyzing individual instances without requiring a complete process model. Constructing models at the instance level avoids scalability issues, even with large data sets. However, this approach relies on causal relations from the event log, so instance-graph accuracy depends on log quality and detail. As an improvement, (van Dongen et al., 2007) propose an incremental approach to learn a $PN$ from event data iteratively. This iterative strategy reduces memory requirements, making it feasible for large logs. However, while memory usage is reduced, computation time increases because all regions must be computed before identifying the minimal ones. As noted in (Solé and Carmona, 2012), this approach trades space for time.

The two-step approach in (van der Aalst et al., 2010) details how event logs are transformed into a TS and how a PN is derived from it. Unlike the earlier-mentioned iterative approach, which builds the TS incrementally, this method constructs a complete TS in one step, so it needs representative logs. The authors also discuss representing states as sequences of activities, applying various abstractions to the event log. However, they restricted the approach to *elementary* PNs. (Shershakov et al., 2017) address the issue of evaluating the transition systems generated by different configurations. The authors essentially use simplicity and precision as metrics. However, the author only considers the past (or history) of a state, and a numerical search is used to find suitable configurations.

Another issue with region-based approaches, as stated in (Augusto et al., 2022), is their high computational cost. Recent works address scalability constraints and propose efficient methods for constructing a TS from large event logs. For example, (Teren et al., 2023a,b) introduce a framework to decompose a PN into synchronizing sub-systems. In (Peng and Szczerbicka, 2024), the authors enhance the region-based process discovery algorithm by proposing an advanced minimal region generation approach that is twice as fast as the original.

Typically, these works assume that the TSs are already extracted from event logs, so they focus on improving the efficiency of learning PNs via the *theory of regions*. However, accurately extracting state information and selecting a suitable TS representation remain crucial for the final PN. Thus, simply stating how to extract state information is insufficient. It is also necessary to provide methods for evaluating different configurations, helping to find the most suitable TS representation at a moderate computational cost.

## 3 PRELIMINARIES

In this section, we explore the foundational concepts vital for the focus of this paper: incremental construction of a transition system from an event log and refining the configurations. Each topic is selected to offer a comprehensive theoretical background that enhances understanding of our proposed algorithm.

### 3.1 Region-based Process Discovery

*Process mining* is an interdisciplinary field that constructs detailed process models from *event logs* to allow in-depth analysis. As modern information systems generate vast amounts of data, efficiently derive process models from these datasets remains a challenge. Various *discovery algorithms* have been developed to address this, producing models such as *transition systems* (TS), *Petri nets* (PN), and *Business Process Modeling Notation* (van der Aalst, 2011). Region-based approaches typically use TS as an intermediate representation of event logs, from which a PN is derived. The definition of a (labeled) transition system is provided in Definition 3. Due to space limitations, details of PN and its derivation via region-based approaches are referred to in the literature cited in section 2.

**Definition 1** (Language and Overapproximation (Solé and Carmona, 2012))**.** *The **language** of a system, denoted by $\mathcal{L}(\mathsf{S})$, is the set of all possible sequences feasible from the initial state. For two systems $\mathsf{S}_1$ and $\mathsf{S}_2$. $\mathsf{S}_2$ is called an **overapproximation** of $\mathsf{S}_1$, when $\mathcal{L}(\mathsf{S}_1) \subseteq \mathcal{L}(\mathsf{S}_2)$.*

**Definition 2** (Process Mining Problem (Bergenthum et al., 2007))**.** *The objectives of the process mining problem can be articulated as follows: given an event log $\mathsf{L}$ from system $\mathsf{S}$, we are searching for a preferably small finite marked PN, s.t. 1) $\mathcal{L}(\mathsf{S}) \subset \mathcal{L}(\mathsf{PN})$; and 2) $\mathcal{L}(\mathsf{PN}) \setminus \mathcal{L}(\mathsf{S})$ is minimal, which implies that the learned $\mathsf{PN}$ should be the tightest overapproximation over $\mathsf{S}$.*

**Definition 3** ((Labeled) Transition System)**.** *A transition system (TS) is a quadruple $\mathsf{TS} = (S, E, F, s_0)$, where $S$ is the set of states, $E$ is the set of events, $F : S \times E \times S$ is the set of **labeled** arcs, $s_0 \in S$ is the initial state.*

### 3.2 From Event Log to Transition System

The size of the extracted transition systems is highly related to the way the states of the process are represented. First, we define the trace and the event log:

**Definition 4** (Trace and Event Log (van der Aalst, 2011))**.** *Given the universe of activities $\mathcal{A}$. A **trace** $\sigma =< a_1, \ldots, a_n > \in \mathcal{A}^*$ is a sequence of activities, where $\mathcal{A}^*$ is the Kleene closure of the set $\mathcal{A}$. An event log is a multiset of traces, i.e. $\mathsf{L} \in \mathcal{B}(\mathcal{A}^*)$.*

One of the issues in transforming an event log into a transition system is that the event logs contain various information about events, but there is no information about the states. This is typically a challenge for constructing a transition system, since it is a state-based framework. Thus, identification of the states is crucial. In the literature, a state can be identified by considering the ***past/future (P/F) strategy***: ① *past*, i.e., what has happened before the state; ② *future*, i.e., what will happen after the state; and ③ *hybrid*, i.e., consider both past and future (van der Aalst et al., 2010). In the rest of the work, we use P/F-1, P/F-2, and P/F-3 to represent the three different strategies accordingly. Take the trace $\sigma_1$ in Figure 1 as an example. After activity $c$ is performed, the process has entered the state $s$. Based on the strategies stated above, the state $s$ can be expressed as (a-b-c, -), (- , d-e-f-g) or ( a-b-c, d-e-f-g) considering the P/F-1, the P/F-2, or P/F-3 respectively. A state is written in the form of a tuple, where the first entry represents the past of the state, and the second the future. The symbol '-' is used as a placeholder when the past or future is not used or when the state has no past or future. Thus, in this work, the state and its equivalence are defined as follows.

**Definition 5** (State and Equivalence)**.** *Its past and / or future give a state with or without considering their orders, written as a tuple $s = (s_{past}, s_{future})$. The states $s$ and $s'$ are considered equivalent when their past and future are equivalent, i.e., $s = s'$, iff. $s_{past} = s'_{past}$ and $s_{future} = s'_{future}$.*

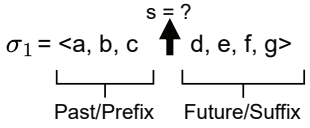



Figure 1: Example of state representation.

The *P/F strategy* sets the baseline for the state representation, which can be defined by ***horizons***. However, it needs to be further polished to fit the intended process. For this purpose, several other *abstractions* were discussed in (van der Aalst et al., 2010), i.e., the ***length***, the ***projection***, the ***frequency***, the ***order***, and the ***grouping***.

The ***horizons*** of the past and future, denoted by $\mathsf{h_p}$ and $\mathsf{h_f}$, respectively, define the level to which activity in the past and future affects the state. $\mathsf{h_p} = 0$ if the past is irrelevant to the state, and similarly $\mathsf{h_s} = 0$ if the future is irrelevant. For example, if we consider a configuration, that is, $\mathsf{h_p} = 2$ and $\mathsf{h_s} = 1$, for the state $s$ in Figure 1, then it is represented as $s =$ (b-c, d). Note that the horizons are named as *window size* in other works such as (Shershakov et al., 2017). ***Lengths***, denoted by $\mathsf{l_p}$ and $\mathsf{l_f}$, act similarly to the ***horizons***, which further restricts the length of state representation. However, this abstraction is **not** considered in this work. First, these two abstractions share very similar functionality. Second, the parameter '***lengths***' is completely artificial and is decided by the modelers, adding more subjectivity to the model. If the length parameters $\mathsf{l_p} = 1$ and $\mathsf{l_f} = 1$ are applied to the state $s$, we obtain $s =$ (c , d)

Projecting the traces onto a subset of the activities provides the opportunity to filter out unimportant activities. Given a set of important activities $\boldsymbol{X}$, the ***projection*** of a trace $\sigma$ onto $\boldsymbol{X}$ is denoted by $\sigma_{\uparrow \boldsymbol{X}}$. For example, the projection of a trace $\sigma =< a, b, a, c, d >$ on a set $\boldsymbol{X} = \{a, c\}$ is $\sigma_{\uparrow \boldsymbol{X}} =< a, a, c >$. This abstraction intends to help the modelers focus on the important activities of the process and prevent the construction of over-complicated models. However, applying the projection without *a priori* knowledge about the process can lead to over-simplification, and the resulting model can oversee some important behaviors.

The ***frequency*** defines whether the occurrences of the traces and the state-transitions are considered in the model. The frequencies can be used to filter out important behaviors of the process and to evaluate stochastic parameters. Neglecting the frequency can reduce the computational cost to a certain extent since only unique traces are considered. However, recent approaches address the value of stochastic characteristics in the models. In addition, the reduction obtained by ignoring the frequencies is far from significant compared to modern hardware improvements. Thus, frequencies are generally considered in this work.

The abstraction '***order***' determines whether the ordering of the past and/or future is important to the state representation. Consider two states based on the past $s_1 =$ (a-b-c, -) and $s_2 =$ (a-c-d, -). If the order of the past is irrelevant, then $s_1$ and $s_2$ are equivalent. If the order of the past is considered, then obviously $s_1 \neq s_2$.

***Grouping*** similar activities together is another strategy that helps to reduce the size of the model. Technically, the recent advance in Large Language Models alleviates this task tremendously (Kourani et al., 2024). There are still some issues with this abstraction. First, grouping similar semantic events requires that the event logs be properly labeled. Second, it brings confusion to the model, where there are several semantic similar events, but they lead to completely different states of the process. Thus, we leave out this abstraction in the state representation. But address the possibility of using this abstraction in the *post-processing* to get a more precise model.

We summarize the usage of the different abstractions in this work. First, the ***horizons*** and the ***order*** are affecting directly the state representation. Second, the ***frequency*** is generally considered, i.e., it is set to True by default. Thirdly, the ***projection*** and ***grouping*** are used only in post-processing to further polish the result. Last, the ***lengths***

are omitted from our approach due to their redundancy and subjectivity.

### 3.3 Metrics

Given a transition system $\mathsf{TS} = (S, E, F, s_0)$, the ***simplicity*** of the transition system is given by (Buijs et al., 2012):

$$\mathsf{Simp}(\mathsf{TS}) = \frac{|E| + 1}{|S| + |F|}, \tag{1}$$

where $|S|$, $|E|$ and $|F|$ are the cardinality of the set $S$, $E$ and $F$ respectively. The ***precision*** of the TS is:

$$\mathsf{Prec}(\mathsf{TS}) = \frac{1}{|S|} \cdot \sum_{s \in S} \mathsf{Prec}(\mathsf{s}) \tag{2}$$

with

$$\mathsf{Prec}(\mathsf{s}) = \frac{1}{\mathsf{NoV}(s)} \cdot \sum_{i=1}^{\mathsf{NoV}(s)} \frac{|s\bullet| - |\hat{s\bullet}_i|}{|s\bullet|}$$

where $\mathsf{Prec}(s)$ is the partial precision for state $s$, $\mathsf{NoV}(s)$ is a number of all visits of state $s$ during a replay of the full transition system. $\hat{s\bullet}_i$ is a set of penalized output transitions of state $s$, i.e., transitions that do not have active counterparts compared to the precise model. Details on definitions and implementations can be found in (Shershakov et al., 2017).

## 4 INCREMENTAL PROCESS MINING WITH STATE IDENTIFICATION

The goal of the incremental algorithm is to learn a suitable transition system representation from a given event log. Note that, different from the algorithm proposed in (van der Aalst et al., 2010), the completeness of the event log is not required. Thus, the central issue remains in configuring the abstractions introduced section 3. The relevant abstractions are summarized in Table 1, which are used to form a Python dictionary like the input CONFIG in the incremental algorithm.

In subsection 3.2, we have reasoned the general usage of ***frequency***, categorized ***projection*** and ***grouping*** to post-processing, and dropped the ***lengths***. Thus, these abstractions do not affect the state representation in this work. The abstractions that we need to consider in the state representation are ***horizons*** and ***order***, i.e., $\mathsf{h_p}$, $\mathsf{h_f}$ and is_order. Although we have reduced the number of relevant abstractions in the state representation, combining ***horizons*** and ***order*** can still lead to a combinatorial problem. Theoretically, there are infinite number of configurations, since the past or future can be infinitely long. Thus, a strategy for the incremental algorithm is required.

### 4.1 P/F Refinement and Incremental Algorithm

Finding a suitable or optimal abstraction configuration (CONFIG) for state representation is difficult. Typically,

Table 1: Notions and Parameters

| Key | Value | Description |
|---|---|---|
| $h_p$ | integers | horizon past |
| $h_f$ | integers | horizon future |
| is_order | boolean | order |
| $\boldsymbol{X}$ | set of activities | projection |
| is_grouping | boolean | grouping |

Table 2: Constructing TS for $\mathsf{L}^*$ with 251 traces and 247 activities using P/F-1 (Shershakov et al., 2017).

| $\mathsf{h_p}$ | #S | #ST | Simplicity | Precision |
|---|---|---|---|---|
| 1 | 248 | 1755 | 0.1240 | 0.3791 |
| 3 | 3603 | 4838 | 0.0290 | 0.8892 |
| 5 | 5601 | 6129 | 0.0210 | 0.9671 |
| 7 | 6515 | 6806 | 0.0190 | 0.9836 |
| 10 | 7228 | 7392 | 0.0170 | 0.9927 |
| 15 | 7840 | 7905 | 0.0160 | 0.9968 |
| $\infty$ | 8088 | 8087 | 0.0153 | 1.0000 |

#S: number of states, #ST: number of state-transitions.
Dataset: `https://fluxicon.com/blog/2015/05/bpi-challenge-2015/`

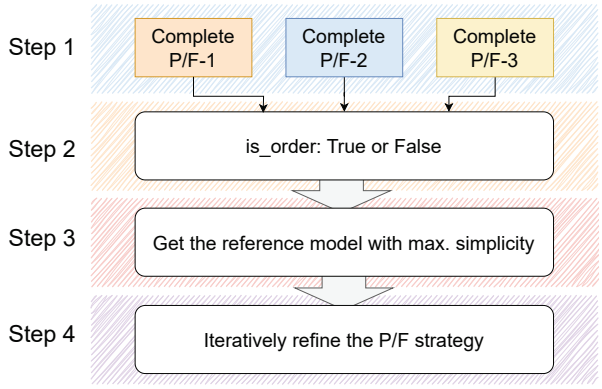



Figure 2: General Refinement Strategy.

there is no *a priori* knowledge about the characteristics of the state in an event log. Decisions are often made based on the experience of the modelers (van der Aalst et al., 2010) or by searching all combinations (Shershakov et al., 2017). Fortunately, certain state-representation features can help. Since is_order is binary, the main complexity lies in choosing the horizons $\mathsf{h_p}$ and $\mathsf{h_f}$.

Table 2 shows how the size of TS (i.e., number of states and state-transitions) learned from $\mathsf{L}^*$ changes under P/F-1 as $\mathsf{h_p}$ varies. A larger $\mathsf{h_p}$ yields a more complex but precise TS. Its precision reaches 1 when $\mathsf{h_p} = \infty$, where $\infty$ denotes the maximum length of past or future. For $\mathsf{L}^*$, the maximal $\mathsf{h_p}$ is 83. Thus, the best P/F-1 setting lies between $\mathsf{h_p} = 1$ (the simplest model) and $\mathsf{h_p} = \infty$ (the most precise). In this example, $\mathsf{h_p} = 3$ seems optimal, offering a substantial precision gain. However, simplicity drops dramatically for higher $\mathsf{h_p}$. In addition, a greater $\mathsf{h_p}$ adds complexity, underscoring the importance of transforming the TS into a PN for a more compact representation (Peng

and Szczerbicka, 2024), although that is not our focus here.

Based on these findings, we can derive a refinement strategy. Knowing that minimal values of $\mathsf{h_p}$ and/or $\mathsf{h_f}$ (i.e., 1) yield the simplest models and simplicity decreases as these parameters grow, we can start with 1 and increase them incrementally. The gain in precision and simplicity can then serve as the stopping criterion, balancing simplicity and precision.

The general refinement strategy is depicted in Figure 2. In step 1, the complete P/F strategies are considered, i.e., $\mathsf{h_p} = \infty$ and/or $\mathsf{h_f} = \infty$. This serves two purposes. First, reference models are required when calculating the precision. As mentioned in subsection 3.3, the precision is a characteristic of the current model against a reference model. Reference models are generally the models with the highest precision. By combining different configurations, we obtain six distinct settings, yielding six different $TS$s. In step 3, the TS with the highest simplicity is selected as the reference model, and its P/F strategy is used for iterative refinement in step 4. This procedure is demonstrated in Algorithm 1 as a pseudo-code. In this work, the P/F refinement algorithm is stopped when the user-defined min_gain is reached. The precision gain, $\tau$, is the difference between the precision of the current TS and that of the last TS.

The incremental algorithm in Algorithm 2 demonstrates briefly the idea of how a TS is created. First, the set of state-transitions F of the super-TS is initialized as an empty list. This allows us to evaluate the frequencies and stochastic parameters. Then the state-transitions are obtained from traces using the get_st_from_trace function, which are appended to the list $F$. The algorithm runs until all the traces in the event log $L$ are processed. Let the algorithm run until the event log is empty, which has the benefit that it will also process the traces added to the event log after the algorithm is triggered. The super-TS is created by the function create_ts_from_st with the input F and CONFIG.

**Algorithm 1** P/F refinement (ts_pf_refinement)

**Require:** $L$, $\boldsymbol{X}$, is_grouping, min_gain
**Ensure:**
$\mathsf{TS_{ref}} = \text{get\_reference\_model}(L)$
$\mathsf{CONFIG} = \text{get\_initial\_config}(\mathsf{TS_{ref}}, \boldsymbol{X}, \text{is\_grouping})$
$\text{simp\_0}, \text{prec\_0} = 0$
$\tau = 1$
**while** $\tau \geq \text{min\_gain}$ **do**
    $\mathsf{TS} = \text{get\_ts\_incremental}(L, \mathsf{CONFIG})$
    $\text{prec}_1 = \text{get\_metrics}(\mathsf{TS})$ #Equation 1Equation 2
    $\tau = \max(\text{simp\_1} - \text{simp\_0}, \text{prec\_1} - \text{prec\_0})$
    $\text{simp\_1} = \text{simp\_0}$
    $\text{prec\_1} = \text{prec\_0}$
    $\mathsf{CONFIG} = \text{update\_config}(\mathsf{CONFIG})$
**return** TS

**Algorithm 2** Incremental Algorithm (get_ts_incremental)

**Require:** L, CONFIG
**Ensure:**
$\mathsf{F} = [\,]$
**while** $\mathsf{L} \neq \emptyset$ **do**
    $\sigma = L.pop()$
    $\mathsf{F}_\sigma = \text{get\_st\_from\_trace}(\sigma, \mathsf{CONFIG})$
    $\mathsf{F}.\text{append}(\mathsf{F}_\sigma)$
$\mathsf{TS} = \text{create\_ts\_from\_st}(\mathsf{F}, \mathsf{CONFIG})$
**return** TS

### 4.2 Tools Support & Implementation

The incremental algorithms including the numerical case study presented in section 5 are implemented in Python. The code can be found in our Git repository (link: `https://git.com/peng_luh/ecms_2025`). To reduce the work on visualization of transition systems, the TransitionSystem class from the PM4PY package (Berti et al., 2023) is used as the base class.

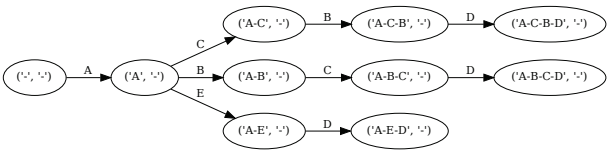
## 5 NUMERICAL CASE STUDY

To demonstrate the concepts introduced in the last sections, we use the following simple event log as an example: $\mathsf{L}_1 = \left\{[A, B, C, D]^5, [A, C, B, D]^2, [A, E, D]^2\right\}$. The event log is given as a multiset, where the power is used to count the frequency of the sequences. In order to get a view on the difference between the P/F strategies and how they are interacting with the other abstractions, three different *super-transition systems* are constructed, which are depicted in Figure 3 with their configurations, respectively.

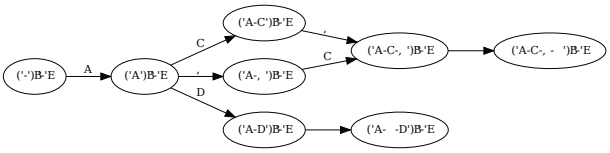
First, there are 10 states in Super-TS-1 (Figure 3a) but only 6 states in Super-TS-3 (Figure 3c), where the complete past ($\mathsf{h_p} = \infty$) and only the last event ($\mathsf{h_p} = 1$) are considered in the state representation, respectively. Indeed, the Super-TS-1 is much more rigid than Super-TS-3 regarding system behaviors. The Super-TS-1 is basically sequential. There is only one decision to be made in the state (A, -). It mimics a system behavior such as a production process, where resources are distributed at (A, -) to three production lines, where they work independently. Even though the Super-TS-3 own fewer states, it depicts more behaviors than the other two. Thus, it is a more general representation of the event log. We observe that there are concurrency, synchronization, and loops in the process.

In Super-TS-1 we also notice that there is a concurrent behavior between ⟦(A-C, -), B, (A-C-B, -) ⟧and ⟦(A-B, -), C, (A-B-C, -)⟧. In some cases, the order of the past events is irrelevant. By ignoring the order of the events, we obtain a slightly different super-transition system depicted in Figure 3b, where the state (A-C-B, -) and the state (A-B-C, -) are merged into one. Thus, the states (A-B, -) and (A-C, -) are synchronized in the state (A-B-C, -) in Super-TS-2. However, the order of the past/future could be irrelevant when the horizon is short enough. For example, in Super-TS-3, the states depend only on the last event.

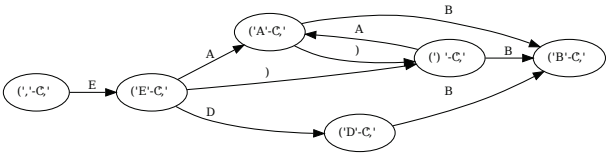
Following the strategy outlined in section 4, the reference models are constructed and their simplicities are evaluated, as shown in Table 3. In this case, the combinations $(\infty, -1, \text{False})$ and $(-1, \infty, \text{False})$, highlighted




(a) Super-TS-1 with $\mathsf{h_{past}} = \infty$ and is_order = True




(b) Super-TS-2 with $\mathsf{h_{past}} = \infty$ and is_order = False




(c) Super-TS-3 with $\mathsf{h_{past}} = 1$ and is_order = False

Figure 3: Super-TSs with $\mathsf{h_{future}} = 0$, $\boldsymbol{X} = \emptyset$, is_frequency = False, is_grouping = None

Table 3: Reference Models and Simplicities

| $\mathsf{h_p}$ | $\mathsf{h_f}$ | is_order | Simplicity |
|---|---|---|---|
| $\infty$ | $-1$ | True | 0.3158 |
| $\infty$ | $-1$ | False | 0.3750 |
| $-1$ | $\infty$ | True | 0.3158 |
| $-1$ | $\infty$ | False | 0.3750 |
| $\infty$ | $\infty$ | True | 0.2400 |
| $\infty$ | $\infty$ | False | 0.3158 |

in green, represent the *complete P/F-1 without order* and *complete P/F-2 without order*, respectively. These configurations yield super-TSs with the highest simplicity. For this example, we select the super-TS resulting from $(\infty, -1, \mathsf{False})$ as the reference model, and its configuration serves as the root for refinement. This step determines whether order should be considered. Here, is_order is set to False and carried forward to the next refinement step. The values of $\mathsf{h_p}$ and $\mathsf{h_f}$ dictate further refinement: If either is $-1$, it is excluded from refinement; if $\infty$, it undergoes refinement in the next step. Thus, only $\mathsf{h_p}$ will be refined in the next step. The numerical results of the refinement on the horizon $\mathsf{h_p}$ are presented in Table 4. We observe that the precision reaches 1 at $\mathsf{h_p} = 2$. However, the algorithm does not stop when precision reaches 1 but continues until the improvement becomes negligible compared to a user-defined threshold. Although precision cannot be further increased beyond $\mathsf{h_p} = 2$, a noticeable gain in simplicity is observed at $\mathsf{h_p} = 3$. The algorithm continues until $\mathsf{h_p} = 4$ is evaluated and a plateau is detected. At this point, the model with $\mathsf{h_p} = 3$ is selected as the optimal configuration. Details regarding the implementation of this numerical example can be found in the given GitHub repository subsection 4.2.

Table 4: Numerical results of P/F refinement on $\mathsf{L}_1$.

| $\mathsf{h_p}$ | #S | #ST | Simplicity | Precision |
|---|---|---|---|---|
| 1 | 6 | 9 | 0.4000 | 0.8333 |
| 2 | 9 | 9 | 0.3833 | 1.0000 |
| 3 | 8 | 8 | 0.3750 | 1.0000 |
| 4 | 8 | 8 | 0.3750 | 1.0000 |

#S: number of states, #ST: number of state-transitions.

## 6 CONCLUSION

This paper introduced an incremental approach that enhances the construction and refinement of transition systems (TS) from event logs. Addressing computational challenges, our method optimizes the representation of states by systematically configuring state abstractions based on past, future, and hybrid activity horizons. The proposed refinement strategy ensures that transition systems are both precise and simple, striking a balance between model accuracy and computational efficiency. Through a detailed numerical case study, we demonstrated that our approach significantly reduces the complexity of transition systems while maintaining high model quality. The results confirm that our incremental strategy enables the construction of more efficient and interpretable process models, making it a valuable contribution to the field of process mining and workflow analysis.

Future work will extend this approach to incorporate a region-based process discovery approach, allowing for adaptive changes in the resulting Petri nets model. In addition, integrating the method into large-scale enterprise process mining applications will further validate its practical effectiveness in handling complex, data-driven workflows. Lastly, comparing the proposed approach to existing approaches would help measure the computational effort and show the superiority over existing ones.

### Acknowledgments

The authors gratefully acknowledge the financial support by the DFG (German Research Foundation) for the Project "Offshore-Plan" (Porject nummer: 411010215).

### References

Adriano Augusto, Josep Carmona, and Eric Verbeek. Advanced Process Discovery Techniques. In Wil M. P. Van Der Aalst and Josep Carmona, editors, *Process Mining Handbook*, volume 448, pages 76–107. Springer International Publishing, Cham, 2022. doi: 10.1007/978-3-031-08848-3_3.

R. Bergenthum, J. Desel, R. Lorenz, and S. Mauser. Process Mining Based on Regions of Languages. In G. Alonso, P. Dadam, and M. Rosemann, editors, *Business Process Management*, volume 4714, pages 375–383. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-75183-0_27.

A. Berti, S. van Zelst, and D. Schuster. PM4Py: A process mining library for Python. *Software Impacts*, 17:1–7, September 2023. doi: 10.1016/j.simpa.2023.100556.

J.C.A.M. Buijs, B.F. van Dongen, and W.M.P. van der Aalst. On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. In *On the Move to Meaningful Internet Systems: OTM 2012*, volume 7565, pages 305–322. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-33606-5_19.

J. Cortadella, M. Kishinevsky, L. Lavagno, and A. Yakovlev. Synthesizing Petri nets from state-based models. In *Proceedings of IEEE International Conference on Computer Aided Design (ICCAD)*, pages 164–171. IEEE, Nov. 5–9 1995. doi: 10.1109/ICCAD.1995.480008.

J. Cortadella, M. Kishinevsky, L. Lavagno, and A. Yakovlev. Deriving Petri nets from finite transition systems. *IEEE Transactions on Computers*, 47(8):859–882, August 1998. doi: 10.1109/12.707587.

J. Desel and W. Reisig. The synthesis problem of Petri nets. *Acta Informatica*, 33(4):297–315, 1996. doi: 10.1007/s002360050046.

A. Ehrenfeucht and G. Rozenberg. Partial (set) 2-structures: Part II: state spaces of concurrent systems. *Acta Informatica*, 27(4):343–368, March 1990. doi: 10.1007/BF00264612.

Humam Kourani, Alessandro Berti, Daniel Schuster, and Wil Van Der Aalst. Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis. 2024. doi: 10.13140/RG.2.2.11821.70880.

S. Peng and H Szczerbicka. Improvements to the region-based petri nets synthesis algorithm for process mining. In *Proceedings of the 16th EAI International Conference on Simulation Tools and Techniques*, Bratislava, 2024. EAI.

C. A. Petri. *Kommunikation mit Automaten*. Phd thesis, Fakultät für Mathematik und Physik der Technischen Hochschule Darmstadt, 1962. Available at `https://edoc.sub.uni-hamburg.de/informatik/volltexte/2011/160/`.

S.A. Shershakov, A.A. Kalenkova, and I.A. Lomazova. Transition Systems Reduction: Balancing Between Precision and Simplicity. In M. Koutny, J. Kleijn, and W. Penczek, editors, *Transactions on Petri Nets and Other Models of Concurrency XII*, volume 10470, pages 119–139. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017. doi: 10.1007/978-3-662-55862-1_6.

M. Solé and J. Carmona. Incremental process discovery. In Kurt Jensen, Susanna Donatelli, and Jetty Kleijn, editors, *Transactions on Petri Nets and Other Models of Concurrency V*, volume 6900, pages 221–242. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-29072-5_10.

V. Teren, J. Cortadella, and T. Villa. Generation of synchronizing state machines from a transition system: A region-based approach. *International Journal of Applied Mathematics and Computer Science*, 33(1), 2023a. doi: 10.34768/amcs-2023-0011.

Viktor Teren, Jordi Cortadella, and Tiziano Villa. Seto: A Framework for the Decomposition of Petri Nets and Transition Systems. In *2023 26th Euromicro Conference on Digital System Design (DSD)*, pages 669–677, Golem, Albania, September 2023b. IEEE. doi: 10.1109/DSD60849.2023.00096.

W. M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi: 10.1007/978-3-642-19345-3.

W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler, and C. W. Günther. Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling*, 9(1):87–111, January 2010. doi: 10.1007/s10270-008-0106-z.

B. F. van Dongen and W. M. P. van der Aalst. Multi-phase Process Mining: Building Instance Graphs. In *Conceptual Modeling – ER 2004*, volume 3288, pages 362–376. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-540-30464-7_29.

B.F. van Dongen and W.M.P. van der Aalst. Multi-phase process mining : Aggregating instance graphs into epcs and petri nets. In *Proceedings of the Second International Workshop on Applications of Petri Nets to Coordination, Workflow and Business Process Management*, 2005.

B.F. van Dongen, N. Busi, G.M. Pinna, and W.M.P van der Aalst. An iterative algorithm for applying the theory of regions in process mining. In *Proceedings of the Workshop on Formal Approaches to Business Processes and Web Services*, pages 36–55, Eindhoven, 2007. Publishing house of university of Podlasie.

## AUTHOR BIOGRAPHIES

**Shengrui Peng** went to Leibniz University of Hanover, where he started with *Civil Engineering* as a Bachelor's and obtained a Master's degree in *Computational Methods in Engineering* in 2019. Immediately after graduation, he joined *L3S Research Center* as a Ph.D. candidate. He has published several papers on modeling, simulation-based optimization, and process mining. He is strongly interested in the concept of combining process mining techniques with advanced AI approaches. His e-mail address is: `peng@l3s.de` and his profile can be found at `https://www.linkedin.com/in/shengrui-peng-9884926a/`.

**Helena Szczerbicka** is Professor Emeritus of the Faculty of Electrical Engineering and Computer Science at the Leibniz University of Hanover, Germany. Her research interests focus on modeling, simulation, and optimization. She has served on the Board of Directors of the Society of Modeling and Computer Simulation International SCS for many years. She has received much recognition for her various scholarly publications, achievements, and professional activities that focus on the importance of modeling and simulation in its interaction with other disciplines. Her email address is: `hsz@sim.uni-hannover.de`. Her website is `https://www.l3s.de/people/members/helena-szczerbicka/`

# Why the Fossen model is so popular for marine vehicles' dynamic analysis?

Paweł Piskur, Piotr Szymak, Rafał Kot, Łukasz Marchel, Krzysztof Naus
Polish Naval Academy
Smidowicza 69, 81-127 Gdynia, Poland
Email: p.piskur@amw.gdynia.pl, p.szymak@amw.gdynia.pl, r.kot@amw.gdynia.pl,
l.marchel@amw.gdynia.pl, k.naus@amw.gdynia.pl

**KEYWORDS**

Dynamic models simulation; Fossen model; Marine vehicles; mas-spring-dumper.

**ABSTRACT**

Among various approaches, the Fossen model has gained significant popularity due to its comprehensive representation of nonlinear hydrodynamic forces, stability, and modular structure. This paper compares different simulation models using a one-degree-of-freedom (1-DOF) system, demonstrating the benefits of the Fossen model over conventional mass-spring-damper-based approaches. By incorporating added mass, nonlinear damping, and restoring forces, the Fossen model provides a more accurate representation of marine dynamics. Additionally, its difference equation formulation ensures efficient computation while preserving stability, making it suitable for both real-time applications and high-fidelity simulations. The study highlights the impact of different numerical integration techniques and simulation environments, providing a foundation for extending the analysis to multi-degree-of-freedom systems.

## INTRODUCTION

A system model describes the relationship between input and output variables, typically using differential equations [1], state-space representations or transfer functions. Dynamic models can be represented in both continuous and discrete time domains [2], making them essential for analyzing and designing control systems [3].

One of the most widely recognized dynamic models in marine vehicle dynamics and control is the Fossen model, which has been a standard since the 1990s. Its foundation was laid by Thor I. Fossen in his Ph.D. dissertation (1991) and later formalized with the publication of *Guidance and Control of Ocean Vehicles (1994)* [4]. This work established a systematic mathematical framework for modeling marine craft, integrating hydrodynamic forces, control strategies, and nonlinear dynamics.

Throughout the late 1990s and 2000s, the model gained widespread adoption for various marine applications, including Autonomous Underwater Vehicles (AUVs) [5], Remotely Operated Vehicles (ROVs), and ships. Significant contributions from the Marine Cybernetics Group at NTNU further enhanced its applicability. The release of Fossen's second book, *Handbook of Marine Craft Hydrodynamics and Motion Control (2011)* [6], introduced refined modeling techniques and advanced control strategies, reinforcing its status as a benchmark in the field.

The success of the Fossen model is attributed to its modularity, compatibility with modern control systems, and extensive real-world validation. Its influence extends beyond academia, as it has been integrated into widely used simulation platforms such as Matlab (Marine Systems Simulator) [7], Simulink, and Python-based simulation frameworks. Today, it remains an industry and research standard in dynamic modeling, simulation, and control of marine vehicles [8], [9], [10].

The Fossen model for six degrees of freedom (6-DOF) is formulated mathematically as:

$$\mathbf{M}(\boldsymbol{\eta})\dot{\mathbf{v}} + \mathbf{C}(\mathbf{v}, \boldsymbol{\eta})\mathbf{v} + \mathbf{D}(\mathbf{v})\mathbf{v} + \mathbf{g}(\boldsymbol{\eta}) = \boldsymbol{\tau}, \qquad (1)$$

where $\mathbf{M}(\boldsymbol{\eta})$ is the inertia and added mass matrix, $\mathbf{C}(\mathbf{v}, \boldsymbol{\eta})$ is the Coriolis and centripetal matrix, $\mathbf{D}(\mathbf{v})$ is the damping matrix, $\mathbf{g}(\boldsymbol{\eta})$ is the vector of hydrostatic forces and moments, and $\boldsymbol{\tau}$ represents the control input. Vectors $\mathbf{v}$ and $\boldsymbol{\eta}$ denote velocity and position/orientation in the body-fixed and inertial frames, respectively.

A one-degree-of-freedom (1-DOF) dynamic model, such as a mass-spring-damper system, provides a simplified way to introduce the main principles of marine system dynamics modeling. This approach makes it easier to understand how hydrodynamic forces, damping, and environmental interactions are represented in the Fossen model before moving to the full six-degree-of-freedom (6-DOF) system.

Using a 1-DOF model allows for the step-by-step examination of the effects of mass and added mass (influence of inertia), hydrodynamic forces (linear and nonlinear damping), restoring forces and stability (e.g., restoring moments), external forces (e.g., control forces, waves, and ocean currents) [11]. Modeling a 1-DOF system enables validation and testing of the equations of motion under controlled conditions [12], [13]. It gives comparison of simulation results with analytical solutions or experimental data, helping to understand both the limitations and benefits of the Fossen model. This structured approach helps clarify dynamic mechanisms before introducing more complex interactions.

Many control algorithms (e.g., PID, LQR, MPC) are first tested on 1-DOF models before being implemented in full-scale systems. This approach allows for a better understanding of vehicle dynamics and facilitates the identification of the main parameters, such as hydrodynamic damping and the influence of control forces.

Once the principles of the Fossen model are explained using a 1-DOF system, it becomes logical to extend the discussion to the full 6-DOF model, where all forces and moments act simultaneously. This hierarchical presentation prevents misunderstandings caused by overly complex mathematical dependencies in the initial learning phase.

This paper focuses on comparing different modeling techniques applied to a one-degree-of-freedom mass-spring-damper (MSD) system.

## 1-DOF SYSTEM SIMULATION

A second-order differential equation describes the dynamics of a MSD model in a 1-DOF, which can be presented as in Fig. 1.

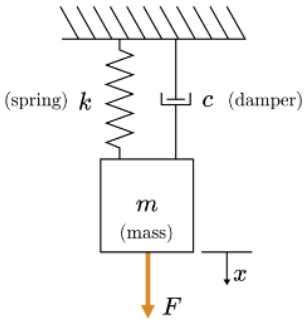



Fig. 1. Mass-spring-damper (MSD) single degree of freedom system

The equation of motion for a 1-DOF in a single direction is given by the equation (2):

$$(m + m_a)\ddot{x} + b\dot{x} + cx = F_{ext}(t) \tag{2}$$

where:
$m$ – mass of the floating body $[kg]$,
$m_a$ – added mass due to hydrodynamic effects $[kg]$,
$\ddot{x}$ – acceleration $[m/s^2]$,
$\dot{x}$ – velocity $[m/s]$,
$x$ – displacement $[m]$,
$b$ – hydrodynamic damping coefficient $[Ns/m]$,
$c$ – hydrostatic restoring coefficient (depends on buoyancy and gravity) $[N/m]$,
$F_{ext}(t)$ – external forces acting on the body, e.g., wave excitation force $[N]$.

When a marine vehicle is analysed, the mass term $(m + m_a)$ accounts for the system's inertia, including the added mass effect from the surrounding water. The damping term $b\dot{x}$ represents the hydrodynamic resistance to motion, which can be linear or nonlinear. The restoring force $cx$ arises due to buoyancy and gravity effects, bringing the system back to equilibrium. Additionally, the external force $F_{ext}(t)$ represents disturbances such as wave forces or control inputs. Depending on the chosen reference frame, the motion can occur in either the horizontal or vertical direction. Solving this equation provides the system's response in terms of displacement $x(t)$ under external forces $F(t)$.

## I. Methods of Modeling

### A. Block diagram in simulink

The equation (2) can be used to build the block diagram as presented in Fig. 2. For that reason, the equation (2) can be modified as presented in the equation (3). The initial values for displacement and velocity are defined as v0 and x0.

$$\ddot{x} = \frac{F_{ext}(t) - b\dot{x} - cx}{m + m_a} \tag{3}$$

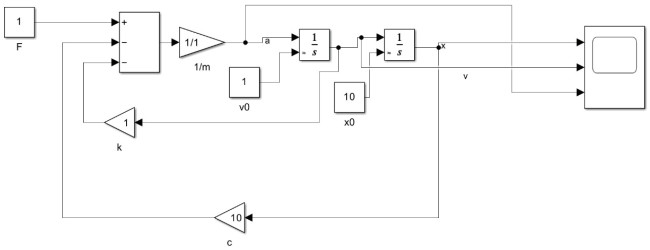



Fig. 2. Simulation model with 1-DOF designed in Matlab-Simulink as a block diagram

The equation (3) can be directly defined in the function block and added to the block diagram as presented in Fig. 3. The significant difference is the lack of initial position and speed values, while the constant values are defined in separate blocks.

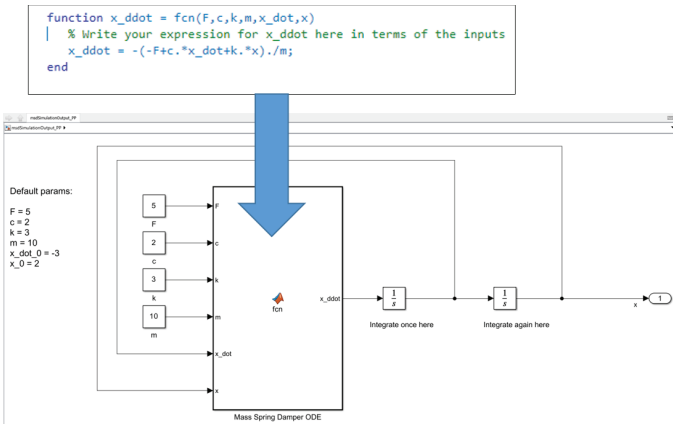



Fig. 3. Simulation model with 1-DOF designed in Matlab-Simulink as a block diagram with build in function

Also, if we assume that the initial conditions for the linear model are equal to zero, the transmittance can be used. The transmittance can be represented as equation (4):

$$G(s) = \frac{1}{ms^2 + cs + k} \tag{4}$$

It can be implemented in Simulink model with using the `Transfer Function` block (see Fig. 4). In all the above models, different ODE solvers can be used

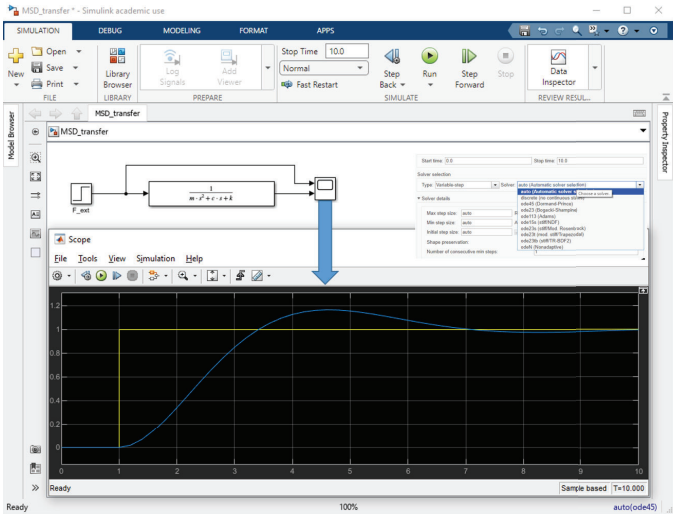



Fig. 4. Simulation model with 1-DOF designed in Matlab-Simulink with transmittance function

(see Fig. 4). Solvers available in Matlab-Simulink can be categorized into variable-step and fixed-step methods. Variable-step solvers, such as `ode45` (Dormand-Prince method), `ode15s` (stiff solver), and `ode23` (low-order method), dynamically adjust the step size to balance accuracy and computational efficiency. Fixed-step solvers, such as `ode4` (Runge-Kutta 4th order), `ode1` (Euler method), and `FixedStepDiscrete`, are useful for real-time applications and discrete-time systems. The choice of solver depends on system stiffness, accuracy requirements, and computational constraints. Another way of modelling the dynamic model is through direct definitions of differential or difference equations in Matlab script as presented in the next paragraphs.

### *B. Modeling with Ordinary Differential Equations (ODE)*

ODE-based modeling involves converting the second-order differential equation into a system of first-order equations:

$$x_1 = x, \tag{5}$$
$$x_2 = \dot{x}, \tag{6}$$
$$\dot{x}_1 = x_2, \tag{7}$$
$$\dot{x}_2 = \frac{-kx_1 - cx_2}{m}. \tag{8}$$

This system can be solved numerically using `ode45` function, with using MATLAB code for ODE-based modeling:

```
% System parameters
m = 1.0; k = 10; b = 1;
% Initial conditions
x0 = 10; v0 = 1;
% Time span
t0 = 0; tf = 10;
% ODE function
odefun = @(t, x) [x(2); (-k * x(1) - b * x(2)) / m];
% Solve the system
[T, X] = ode45(odefun, [t0, tf], [x0, v0]);
```

This method provides an accurate continuous-time solution but requires computational resources for numerical integration.

### *C. Modeling with Difference Equations*

A discrete-time approximation using difference equations is effective for numerical simulations. The system is discretized using:

$$a_k = \frac{-kx_k - bv_k}{m}, \tag{9}$$
$$v_{k+1} = v_k + a_k \cdot \Delta t, \tag{10}$$
$$x_{k+1} = x_k + v_{k+1} \cdot \Delta t. \tag{11}$$

The MATLAB code for difference equation-based modeling implementation is:

```
% Time discretization
dt = 0.05; t = t0:dt:tf;
% Initialize state variables
x = zeros(size(t)); v = zeros(size(t)); a = zeros(size(t));
% Initial conditions
x(1) = x0; v(1) = v0;
% Discrete-time simulation
for i = 1:length(t)-1
    a(i) = (-k * x(i) - b * v(i)) / m;
    v(i+1) = v(i) + a(i) * dt;
    x(i+1) = x(i) + v(i+1) * dt;
end
```

The discrete method is computationally efficient but introduces numerical errors due to discretization and requires careful parameter tuning. The choice of $\Delta t$ directly affects the simulation's numerical stability and resolution. A significant time step may cause numerical instability and loss of detail in the system's dynamic response. Small time step improves accuracy but increases computation time and may be infeasible for real-time applications. As a practical guideline, one often starts with a medium time step (e.g., 0.01 s) and performs sensitivity analyses to examine the trade-off between accuracy and computational cost.

The Fossen model describes both low-frequency ship motions (e.g., surge, sway, yaw) and potentially higher-frequency phenomena (e.g., wave-induced vibrations). To capture all relevant frequencies:

$$\Delta t \leq \frac{1}{10\, f_{\max}}, \tag{12}$$

where $f_{\max}$ is the highest frequency of interest. Longer time steps can disregard certain fast dynamics, whereas overly short time steps can be computationally expensive.

## RESULTS OF SIMULATIONS

The results of the simulations using different modeling methods provide insights into the accuracy, computational efficiency, and applicability of each approach. The ODE-based method, implemented with MATLAB's `ode45`, offers high accuracy and smooth solutions but requires greater computational resources due to its adaptive time-stepping mechanism. In Fig. 5 the results of acceleration, velocity and displacement are

presented, while in Fig. 6 the comparison of solvers (ode45, ode23, ode115, ode23s) are presented for the maximal value of each acceleration, velocity and displacement. It can be shown that value are similar, and adequate enough for dynamic simulation and analysis. Simulation provided in all presented methods gives almost the same results, but the time calculation and complexity of formulation varies. The discrete-time difference equation method is computationally efficient and straightforward to implement, making it suitable for real-time applications; however, it introduces discretization errors that must be carefully managed to maintain numerical stability. Simulations in MATLAB Simulink allow for a graphical block-based approach, offering flexibility in integrating various solver settings, control strategies, and real-time execution scenarios. The choice of method depends on the application: high-precision simulations favor ODE solvers, whereas real-time implementations benefit from discrete-time formulations.

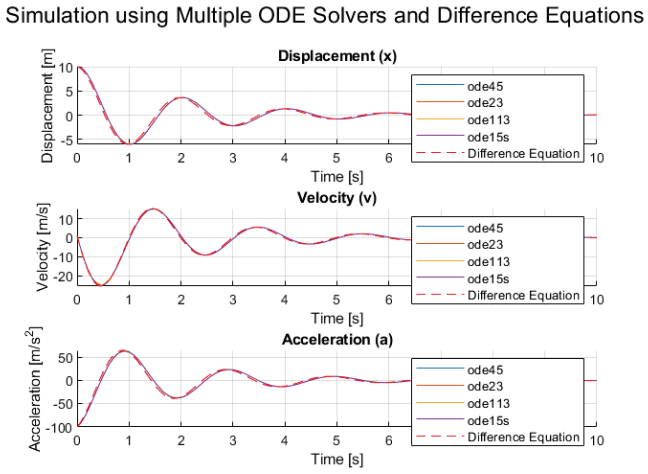



Fig. 5. Acceleration, velocity and displacement as a function in time

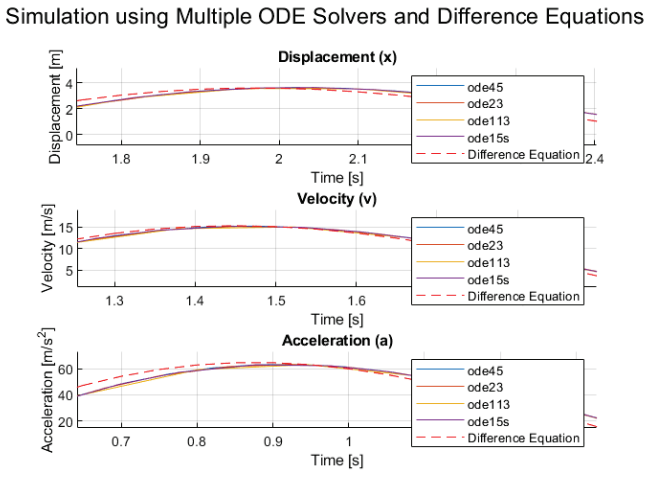



Fig. 6. Comparison of different method and solver used during the simulation

If high-frequency phenomena (e.g., wave-induced vibrations) are critical, smaller steps on the order of 0.001 s or even less may be required.

## CONCLUSIONS

In practice, models are often linearized around specific conditions for control design tasks. However, a fully nonlinear system is often necessary to accurately capture large-amplitude motions or wave interactions. With its ability to account for nonlinear effects and hydrodynamic forces, the Fossen model remains the most robust approach for marine vehicle dynamics, ensuring reliable performance across various operating conditions. The Fossen model is inherently nonlinear, and its difference equation formulation enables fast computation while preserving the stability of solutions, with matrix notation further enhancing the model's readability.

In the next step, the analysis will be extended to systems with three degrees of freedom.

## Appendices

### I. ODE (ordinary differential equation)

Below is the MATLAB code that solves the differential equations numerically using the ODE solvers and plots the results.

```
% Parameters
m = 1.0;          % mass (kg)
k = 10;           % spring stiffness coefficient
    (N/m)
b = 1;            % damping coefficient (Ns/m)

% Initial conditions
x0 = 10;          % initial displacement (m)
v0 = 1;           % initial velocity (m/s)

\% Simulation time and time step
t0 = 0;           % initial time
tf = 10;          % final time
dt = 0.05;        % time step
t = t0:dt:tf;  % time vector

% ODE function definition
odefun = @(t, y) [y(2); (-k * y(1) - b * y(2))
     / m];

% Solvers for differential equations
solvers = {'ode45', 'ode23', 'ode113', 'ode15s
    '};
ode_results = struct();

for s = 1:length(solvers)
    solver = solvers{s};
    [T, X] = feval(solver, odefun, [t0, tf], [
        x0, v0]);
    X(:,3) = (1/m) * (-k * X(:,1) - b * X(:,2)
        );
    ode_results.(solver).T = T;
    ode_results.(solver).X = X;
end

% Difference equation solution
x = zeros(size(t)); % displacement
v = zeros(size(t)); % velocity
a = zeros(size(t)); % acceleration

% Setting initial conditions
x(1) = x0;
v(1) = v0;

% Simulation of mass-spring-damper system
    using difference equations
```

```
for i = 1:length(t)-1
    % Calculate acceleration
    a(i) = (-k * x(i) - b * v(i)) / m;
    % Calculate velocity
    v(i+1) = v(i) + a(i) * dt;
    % Calculate displacement
    x(i+1) = x(i) + v(i+1) * dt;
end

% Plotting results
figure;

% Displacement comparison
subplot(3,1,1);
hold on;
for s = 1:length(solvers)
    solver = solvers{s};
    plot(ode_results.(solver).T, ode_results.(
        solver).X(:,1), 'DisplayName', solver)
        ;
end
plot(t, x, 'r--', 'DisplayName', 'Difference
    Equation');
title('Displacement (x)');
xlabel('Time [s]');
ylabel('Displacement [m]');
legend;
grid on;
hold off;

% Velocity comparison
subplot(3,1,2);
hold on;
for s = 1:length(solvers)
    solver = solvers{s};
    plot(ode_results.(solver).T, ode_results.(
        solver).X(:,2), 'DisplayName', solver)
        ;
end
plot(t, v, 'r--', 'DisplayName', 'Difference
    Equation');
title('Velocity (v)');
xlabel('Time [s]');
ylabel('Velocity [m/s]');
legend;
grid on;
hold off;

% Acceleration comparison
subplot(3,1,3);
hold on;
for s = 1:length(solvers)
    solver = solvers{s};
    plot(ode_results.(solver).T, ode_results.(
        solver).X(:,3), 'DisplayName', solver)
        ;
end
plot(t, a, 'r--', 'DisplayName', 'Difference
    Equation');
title('Acceleration (a)');
xlabel('Time [s]');
ylabel('Acceleration [m/s^2]');
legend;
grid on;
hold off;

% Display results
sgtitle('Simulation using Multiple ODE Solvers
    and Difference Equations');
```

## References


[1] A. Grzadziela and M. Kluczyk, "Shock absorbers damping characteristics by lightweight drop hammer test for naval machines," *Materials*, vol. 14, no. 4, p. 772, 2021.

[2] B. Szturomski and R. Kiciński, "Simulation of the impact resistance of kilo type submarine loaded with non-contact mine explosion," *Zeszyty Naukowe Akademii Marynarki Wojennej*, vol. 58, 2017.

[3] M. Przybylski, "Selection of the depth controller for the biomimetic underwater vehicle," *Electronics*, vol. 12, no. 6, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/6/1469

[4] T. I. Fossen, *Guidance and Control of Ocean Vehicles*. John Wiley & Sons, 1994.

[5] J. Zalewski and S. Hożyń, "Computer vision-based position estimation for an autonomous underwater vehicle," *Remote Sensing*, vol. 16, no. 5, p. 741, 2024.

[6] T. I. Fossen, *Handbook of Marine Craft Hydrodynamics and Motion Control*. John Wiley & Sons, 2011.

[7] ——, "Marine systems simulator (mss)," 2020, available at: https://www.fossen.biz/mss.

[8] P. Szymak, T. Praczyk, K. Naus, B. Szturomski, M. Malec, and M. Morawski, "Research on biomimetic underwater vehicles for underwater isr," in *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VII*, vol. 9831. SPIE, 2016, pp. 126–139.

[9] P. Szymak and M. Przybylski, "Thrust measurement of biomimetic underwater vehicle with undulating propulsion," *Maritime Technical Journal*, vol. 213, no. 2, pp. 69–82, 2018.

[10] M. Akdağ, T. A. Pedersen, T. I. Fossen, and T. A. Johansen, "A decision support system for autonomous ship trajectory planning," *Ocean Engineering*, vol. 292, p. 116562, 2024.

[11] K. Jaskolski, "Two-dimensional coordinate estimation for missing automatic identification system (ais) signals based on the discrete kalman filter algorithm and universal transverse mercator (utm) projection," *Zeszyty Naukowe Akademii Morskiej w Szczecinie*, no. 52, pp. 82–89, 2017.

[12] M. Morawski, T. Talarczyk, and M. Malec, "Depth control for biomimetic and hybrid unmanned underwater vehicles," *Technical Transactions*, vol. 118, no. 1, 2021.

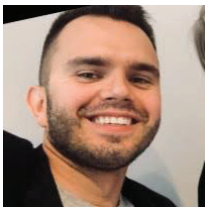
[13] T. Talarczyk, "A dynamic submerging motion model of the hybrid-propelled unmanned underwater vehicle: simulation and experimental verification," *International Journal of Applied Mathematics and Computer Science*, vol. 33, no. 2, pp. 207–218, 2023.


**Paweł Piskur.** Graduated from the Military University of Technology in Warsaw with a degree in Airplane Studies in 2004. He worked for 13 years in the Marine Aviation Base in Polish Army. In 2010 he received PhD degree from Koszalin University of Technology in mechanical engineering. Since 2017 he is working at Polish Naval Academy. In 2024 he obtained the academic degree of habilitated doctor at Air Force Institute of Technology in Warsaw. His research area is strictly connected with underwater unmanned vehicles, especially biomimetic propulsion systems.

**Piotr Szymak.** He is a scientist and navy officer. He has been graduated from the Military University of Technology in Warsaw with an MSc degree in electronics and telecommunication in 1999. He served for 2 years in the 6th Radioelectronic Centre in Polish Navy. Since 2001 he has been serving in Polish Naval Academy (PNA) in the following positions: assistant, lecturer, adjunct, head of department, assistant professor, and associate professor. He was a director of the Institute of Electrical Engineering and Automatics in PNA in the years 2015-2019. In the years 2019 - 2022, he was the Polish naval Captech National Coordinator

at European Defence Agency. In 2004 he received a PhD degree from Polish Naval Academy, and in 2016 he obtained the academic degree of habilitated doctor at Cracow University of Technology. His research area is strictly connected with the modelling and control of marine objects, especially the unmanned surface and underwater vehicles, including recently biomimetic underwater vehicles.

**Rafał Kot.** Graduated from the Military University of Technology in Warsaw with a degree in Electronics and Telecommunication Studies in 2016. He worked for almost 6 years in the Marine Aviation Base in the Polish Army where he gained experience in servicing and maintenance of aircraft in the avionics area. From the beginning of 2022, he is working at Polish Naval Academy. His main research includes underwater obstacle detection systems, path planning algorithms, and collision avoidance systems for Autonomous Underwater Vehicles.



**Lukasz Marchel.**
He is a researcher and lecturer at the Polish Naval Academy in Gdynia. He obtained his Ph.D. in engineering sciences in 2020. His research interests focus on the modeling of physical phenomena, extreme optimization, parallel computing, and their applications in marine engineering. He has authored and co-authored numerous scientific publications in these fields.

**Krzysztof Naus**
He is a professor and researcher at the Polish Naval Academy in Gdynia. He obtained his Ph.D. in 2003 and his habilitation at the AGH University of Science and Technology in Kraków. His research focuses on signal processing, especially in radar and sonar systems. He has published extensively in the areas of software design for swarming underwater vehicles and path-planning algorithms for autonomous marine units.

# On the Optimisation of Machine Learning Models for Predicting the Photosynthetically Available Radiation in the Water Column

Frederic Stahl[1], Lars Nolle[1,2], Martin Maximilian Kumm[2] and Christoph Tholen[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|martin.kumm}@jade-hs.de

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

**ABSTRACT**

Photosynthetically Available Radiation (PAR) is a crucial parameter in oceanography. This study explores the optimization of machine learning models to predict PAR in the water column using selected wavelengths of downwelling irradiance. By leveraging Genetic Algorithms (GA), optimal wavelength combinations were identified for two machine learning models: Linear Regression (LR) and Regression Trees (RT). The models were trained on data from the HE533 expedition and validated using datasets from multiple ship expeditions across different geolocations. Experimental results indicate that the LR model, with an optimal wavelength combination of $E_d(469)$, $E_d(501)$, and $E_d(600)$, achieved the highest prediction accuracy ($R^2$ = 0.9992, MAE = 5.78). The RT model, using $E_d(433)$, $E_d(586)$, and $E_d(687)$, performed slightly worse ($R^2$ = 0.9954, MAE = 16.37). While both models generalised well to unseen datasets, significant prediction errors were observed for small PAR values at lower water depths.

## INTRODUCTION

Photosynthetically Available Radiation (PAR) is an important parameter in modern oceanography. PAR is defined as the integrated radiation between 400-700 nm. One potential application is modelling vegetation growth, because the radiation in this wavelength is a requirement for the photosynthesis process (Holinde and Zielinski, 2016; Wang et al., 2013).

Modelling vegetation growth in the water column is crucial for understanding ecosystem dynamics (Krause-Jensen and Duarte, 2016), predicting climate change impacts (Dutkiewicz et al., 2019), managing water resources (Glibert, 2020) or modelling the oxygen production (Field et al., 1998). Therefore, providing reliable PAR values for different water bodies is an important task.

Figure 1 – Freefall profiler

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; Tholen et al., 2024). 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. Tholen et al. (2024) used data from Freefall Profilers (Figure 1) to

improve the prediction accuracy of PAR by incorporating incoming surface irradiance ($E_s$) (Wollschläger et al., 2020d). Most recently Pitarch et al. (2025) investigated the accurate estimation of PAR based on the different radiance wavelength available on different Argo Float configurations. The long term goal of this series of research is to make the PAR sensor, mounted on the Argo Floats obsolete, to allow the replacement with another sensor (Stahl et al., 2022). However, even if the models show a good performance on the data gathered by Argo Floats (Kumm et al., 2022; Stahl et al., 2022). However, most recent research has shown a unsatisfactory accuracy of the models relaying on the three wavelengths $E_d(400)$, $E_d(412)$, $E_d(490)$ for the Freefall Profiler dataset, which covers more geolocations and is therefore more complex (Tholen et al., 2024).

In Figure 2 two spectral profiles for a water depth of o1 m and 195 m are shown. It can be observed that the radiation profile depends on the water depth. Furthermore, one can observe that the wavelength measured by the Argo Floats might best reflect the characteristics of the full spectra. Therefore, in this paper Genetic Algorithm (GA) (Holland, 1975) will be used to optimise the selection of suitable wavelength for PAR predictions on Argo Floats.

## DATASETS

In this research, data from different ship expeditions are used. The different models are trained on data from the HE533 Expedition (Voß et al., 2020e). As shown in Figure 3, this expedition was undertaken near the northern coast of Norway. The HE533 dataset contains originally 9858 tuples of which 37.77 % had to be discarded because of missing values.

To validate the generalisability of the models developed, data from ten other cruises was used in the validation process (Friedrichs et al., 2020; Mascarenhas et al., 2020; Voß et al., 2020f, 2020a, 2020b, 2020c, 2020d; Wollschläger et al., 2020a, 2020b, 2020c). As shown in Figure 3 these cruises cover different geolocations all over the world. The combined dataset for validation contains 64,060 tuples. However, only 10,954 tuples can be used due to missing values. All datasets used in this research are publicly available on the data portal Pangaea[1].

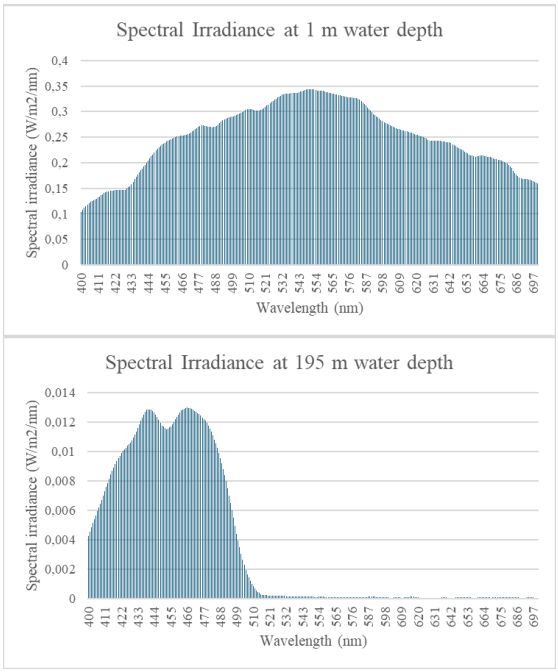



Figure 2 Example Spectral Irradiance for 1m and 195 water depth

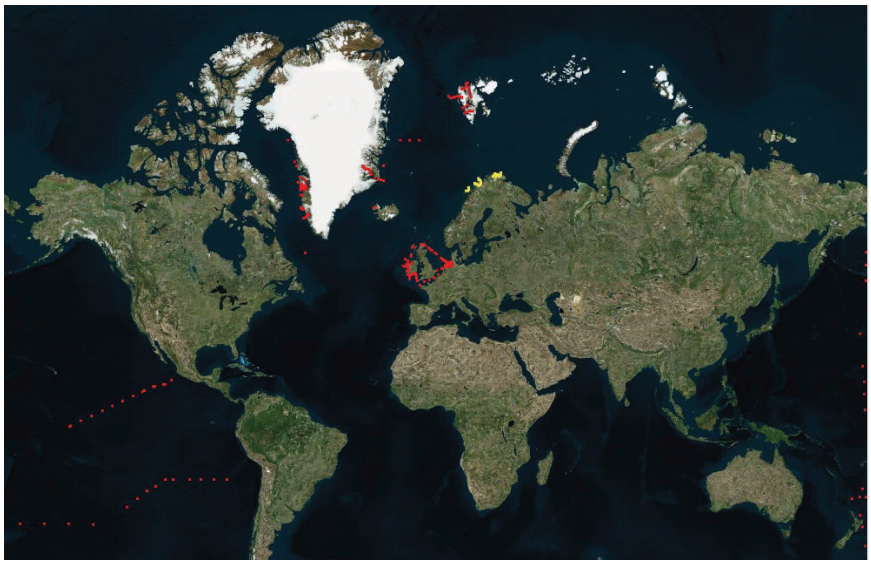


Figure 3 Locations of stations from cruise HE 533 (yellow) used for training and all other cruises (red) used for validation

[1] www.pangea.de

## MODELLING

For modelling purposes, data from the HE533 dataset was used after pre-processing, i.e. normalisation and removal of data records with missing values. Random sampling without replacement was applied, to split this data into a training set (70 %) and a test set (30 %). The training set was used to find optimal wavelength for PAR prediction for two different AI based models, i.e. a Linear Regression model (LR), and a Regression Tree model (RT) utilising GA.

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 were 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 4.

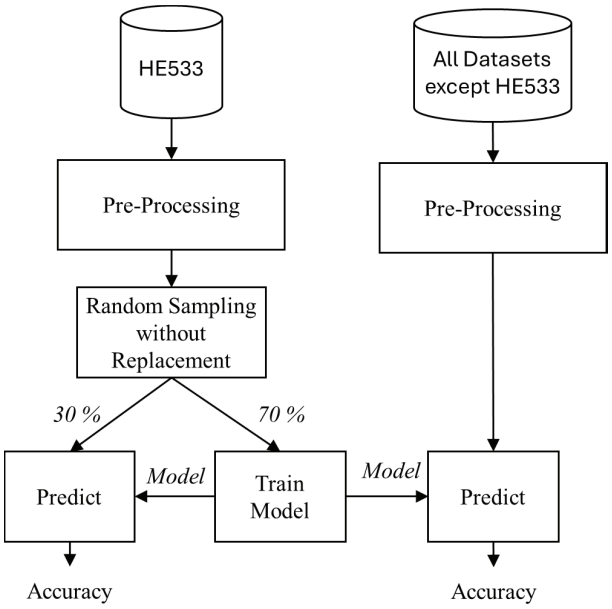



Figure 4: Modelling approach used

## EXPERIMENTAL SETUP

All experiments were conducted using the KNIME workbench (Berthold et al., 2009). For training 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).

During the experiments, GA is used to find an optimal subset of wavelengths to model the PAR value. Due to the limitations of the Argo Float platforms, a maximum number of three wavelengths is chosen during the optimization. The optimization was done utilizing the KNIME *Feature Selection Node* applying Genetic Algorithm as feature selection strategy. The population size was set to 20, while the maximum number of iterations was set to 10. The optimal parametrisation of the GA is not discussed in this research.

## RESULTS

For the LR the highest accuracy was achieved by the combination of $E_d(469)$, $E_d(501)$, and $E_d(600)$, achieving an $R^2$ value of 0.9992. For this configuration, the accuracy on the validation dataset was 0.9982. A scatter plot showing the relation between predictions based on the LR-model and the true PAR values is given in Figure 5. It can be observed that the model tends to underestimate the PAR values, especially for higher values.


Figure 5 Scatter Plot for Linear Regression Predictions of PAR for the validation Datasets

For RT the highest accuracy was achieved by the combination of $E_d(433)$, $E_d(586)$, and $E_d(687)$, achieving an $R^2$ value of 0.9954. For this configuration, the accuracy on the validation dataset was 0.9603. Figure 6 shows the relation between PAR predictions and true PAR values utilising the RT model as scatter plot. It can be observed that the model performance decreases for higher PAR values.

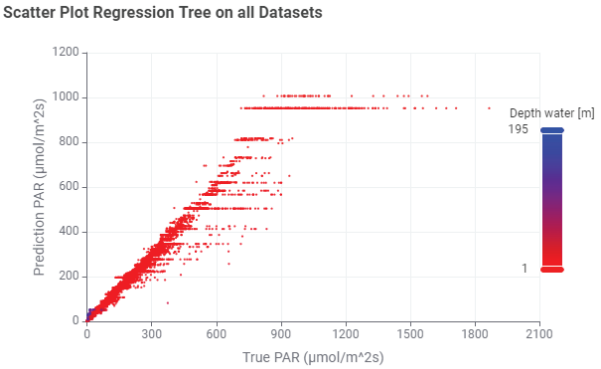



Figure 6 Scatter Plot for Regression Tree Predictions of PAR for the validation dataset

Statistical results of both best found ML models are given in Table 1. The mean absolute error (MAE) of the

LR is approximately three times smaller than the MAE of the RT model.

Table 1 Statistical values for both ML models on the validation set

| Statistical Value | LR | RT |
|---|---|---|
| $R^2$ | 0.9969 | 0.9564 |
| Mean Absolute Error | 5.780 | 16.37 |
| Mean Squared Error | 141.8 | 1979 |
| Mean Absolute Percentage Error | 0.1160 | 0.2208 |

## DISCUSSION

Models trained in this research, i.e. based on the optimised subset of wavelength, clearly outperform the models trained in Tholen et al. (2024). The best performing LR model found in previous work achieved an $R^2$ of 0.884, trained on $E_d(400)$, $E_d(412)$, $E_d(490)$. The best performing RT model achieved an $R^2$ of 0.822 on six wavelengths, incorporating three wavelengths of the surface radiation.

In Figure 6, depicting the results of the regression tree, it can be seen that groups the plotted data points are aligned horizontally. This is because a regression tree partitions the feature space into distinct regions based on decision rules. This is also known as local discretisation (Bramer, 2020). Within each region, the model assigns a constant predicted value to that region. If predicted values (i.e. the predicted PAR values), are plotted on the vertical axis against an independent variable on the horizontal axis (i.e. the true PAR values), all data points within the same leaf node will have the same predicted value. This results in horizontally aligned points because multiple input values share the same output value.

Wollschläger et al (2020d) introduced a trimodal approach to model the underwater light field, splitting the spectral information into three different bands. Within the optimisation GA automated chooses one wavelength from each of the bands for the LR model, while for the RT-model two wavelengths from the third band are chosen. For the RT model, two wavelength from the third band are chosen, while none of the wavelength from the second band is chosen. The trimodal approach was motivated by the absorption characteristics of the dominant substances affecting the underwater light field (Wollschläger et al., 2020d). Thus, the wavelength chosen for the RT might not be optimal for generalising PAR predictions.

As summarised in Table 1, the LR model outperforms the RT-model. Therefore, further discussions will focus on the LR model. To validate the suitability of the ML model to replace the PAR sensor, the relative error of predictions compared to true values is calculated as follows:

$$E_{relative} = \frac{PAR_T - PAR_P}{PAR_T} \cdot 100 \qquad (1)$$

Where $PAR_T$ denotes the true PAR value, while $PAR_P$ is the predicted PAR value. In Figure 7 the relative error is shown in dependence of the water depth for the LR model. One can observe measurements with high relative errors for low water depths.

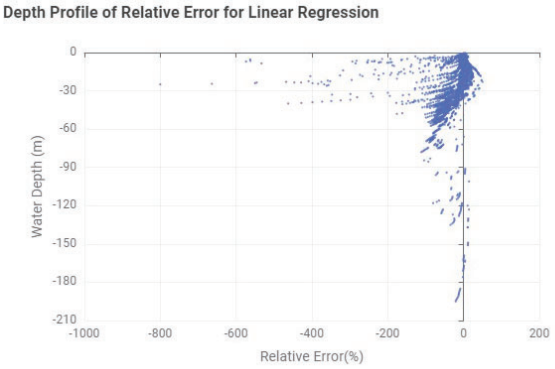



Figure 7 Depth profile of relative prediction error for the linear regression model

In Table 2 the ten highest relative error values are given together with the true PAR values and the PAR predictions. It can be observed that in all cases the true PAR value was smaller than 0.2197.

Table 2 Summary of data tuples with the highest relative error values of the validation set for linear regression

| Relative Error (%) | True PAR | Prediction LR |
|---|---|---|
| -800.08 | 0.1328 | 1.195 |
| -663.74 | 0.1591 | 1.215 |
| -573.67 | 0.1867 | 1.258 |
| -562.72 | 0.1917 | 1.270 |
| -562.05 | 0.1888 | 1.250 |
| -549.95 | 0.1903 | 1.237 |
| -545.66 | 0.1895 | 1.224 |
| -533.03 | 0.1974 | 1.250 |
| -468.04 | 0.2197 | 1.248 |
| -462.45 | 0.2175 | 1.223 |

As shown in Table 2 in all cases of high relative errors the true PAR values are small. For further investigation, Table 3 summarises statistics of the true PAR values of the subset of the validation set with higher relative errors than the given threshold, while the relative error over the true PAR values is shown in Figure 8. It can be observed that high relative errors occur for small true PAR values. If the prediction is limited to PAR values greater than one the maximum relative error is reduced to 50.04 %. Higher thresholds for the maximal relative errors result in smaller mean values for true PAR.

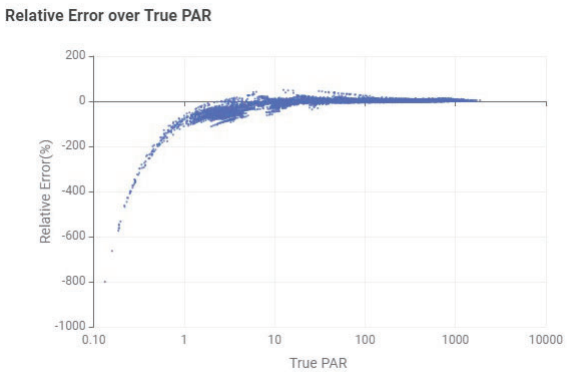



Figure 8 Relative error (%) of the linear regression model over the true PAR value

Table 3 Summary of statistics of true PAR values for tuples of the validation set with specific max. relative error values higher than the threshold

| | Max. relative error (threshold) | | | |
|---|---|---|---|---|
| **Statistics true PAR** | **10 %** | **50 %** | **100 %** | **200 %** |
| Samples | 1930 | 705 | 134 | 57 |
| Mean | 31.745 | 2.092 | 0.617 | 0.322 |
| Std. dev. | 110.7 | 1.258 | 0.42 | 0.096 |
| Minimum | 0.133 | 0.133 | 0.133 | 0.133 |
| Maximum | 900.44 | 12.538 | 2.353 | 0.508 |

## CONCLUSION AND FUTURE WORK

The paper presented a modelling approach for predicting PAR in the water column, which uses downwelling irradiance in selected wavelengths. Two different AI-based modelling approaches, i.e. linear regression and regression tree, were used. The wavelength selection was optimised utilising a Genetic Algorithm. All experiments were conducted using the KNIME workbench.

It was shown that the linear regression model outperforms the regression tree model in terms of $R^2$ and mean absolute error. It was also shown that the models generalise well on data recorded in other geolocations without additional modification or re-training.

However, further analysis revealed that for small true PAR values high relative prediction errors occur, especially in lower water depths. It was already shown that incorporating additional environmental parameters such as e.g. pressure or salinity enhance the accuracy of the regression (Kumm et al., 2022). Therefore, the incorporation of additional parameters can also make the regression more stable and prevent high relative errors. This will be investigated in future research. In addition, domain experts will be incorporated to decide whether the performance of the ML models is high enough to replace the PAR sensor on the Argo Floats for future missions.

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 zukunft.niedersachsen (ZN3480).

## REFERENCES

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

Bramer, M., 2020. Principles of Data Mining, Undergraduate Topics in Computer Science. Springer London, London. https://doi.org/10.1007/978-1-4471-7493-6

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

Dutkiewicz, S., Hickman, A.E., Jahn, O., Henson, S., Beaulieu, C., Monier, E., 2019. Ocean colour signature of climate change. Nat Commun 10, 578. https://doi.org/10.1038/s41467-019-08457-x

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science 281, 237–240. https://doi.org/10.1126/science.281.5374.237

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

Glibert, P.M., 2020. Harmful algae at the complex nexus of eutrophication and climate change. Harmful Algae, Climate change and harmful algal blooms 91, 101583. https://doi.org/10.1016/j.hal.2019.03.001

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

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

Krause-Jensen, D., Duarte, C.M., 2016. Substantial role of macroalgae in marine carbon sequestration. Nature Geosci 9, 737–742. https://doi.org/10.1038/ngeo2790

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

Pitarch, J., Leymarie, E., Vellucci, V., Massi, L., Claustre, H., Poteau, A., Antoine, D., Organelli, E., 2025. Accurate estimation of photosynthetic available radiation from multispectral downwelling irradiance profiles. Limnology & Ocean Methods lom3.10673. https://doi.org/10.1002/lom3.10673

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

Tholen, C., Nolle, L., Wollschlaeger, J., Stahl, F., 2024. Model generalisation for predicting the amount of photosynthetically available radiation in the water column from freefall profiler observations, in: ECMS 2024 Proceedings Edited by Daniel Grzonka, Natalia Rylko, Grazyna Suchacka, Vladimir Mityushev. Presented at the 38th ECMS International Conference on Modelling and Simulation, ECMS, pp. 381–386. https://doi.org/10.7148/2024-0381

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

**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.

**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.

**MARTIN MAXIMILIAN KUMM** graduated from Jade University of Applied Sciences in Wilhelmshaven, Germany, with a Master degree in Mechanical Engineering in 2022. Since 2020 he is a research fellow at the Jade University of Applied Sciences responsible for the research aircraft. He also works in a joint project between Jade University of Applied Sciences, the German Research Center for Artificial Intelligence (DFKI) and marinom GmbH for the development of explainable artificial intelligence decision support systems for nautical officers.

**CHRISTOPH THOLEN** is a Senior Researcher at the German Research Center for Artificial Intelligence (DFKI), where he is Deputy Head of 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.

# KI-gestützte Inhaltsanalyse von wissenschaftlichen Arbeiten mittels Fine-Tunings von Large Language Models

Samir M'Sallem
Technische Hochschule Mittelhessen
Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
samir.m.sallem@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**

Masterarbeit

**Schlüsselwörter**

Künstliche Intelligenz, Natural Language Processing, Large Language Models, Fine-Tuning, German BERT, Named Entity Recognition, Part-of-Speech-Tagging, Text Classification, Topic Modelling, Semantic Textual Similarity, Explainable Artificial Intelligence, Kohärenz, Kohäsion, Argument Mining

**Zusammenfassung**

Diese Masterarbeit untersucht, inwiefern Verfahren der künstlichen Intelligenz zur automatisierten Analyse von Qualitätsmerkmalen in wissenschaftlichen Texten eingesetzt werden können. Das Ziel ist die Entwicklung eines prototypischen Systems, das deutschsprachige Arbeiten aus der Wirtschaftsinformatik hinsichtlich definitorischer Vollständigkeit, argumentativer Bezüge im Schlussteil und semantischer Kohärenz prüft. Zur methodischen Umsetzung werden vortrainierte Transformer-Sprachmodelle herangezogen und durch Transfer Learning auf die Aufgaben feinabgestimmt. Zusätzlich wird eine Themenmodellierung zur Extraktion von Themen in Kapiteln und Argumenten implementiert. Ergänzend erfolgt die Berechnung der Kosinus-Ähnlichkeit dieser Themen, um semantische Nähe von Sätzen und Kapiteln festzustellen. Die Ergebnisse zeigen, dass die anvisierten Charakteristika von wissenschaftlichen Arbeiten zuverlässig erkannt und analysiert werden können. Zukünftige Arbeiten sollten den Fokus auf eine effiziente Erklärbarkeit der Ergebnisse und einer erweiterten Datenbasis für das Modelltraining legen.

**Problemstellung**

In den Bildungsbereichen hängt die Qualität wissenschaftlicher Arbeiten maßgeblich von der Klarheit zentraler Begriffe, der Argumentationsstruktur, sowie der inhaltlichen Kohärenz ab. Künstliche Intelligenz besitzt die Fähigkeit große Mengen an Informationen schnell und präzise zu analysieren. Diese Masterarbeit untersucht daher, inwiefern Verfahren der künstlichen Intelligenz zur automatisierten Analyse dieser Qualitätsmerkmale in wissenschaftlichen Texten eingesetzt werden können, um Lehrende und Studierende bei der Erstellung und Evaluation von wissenschaftlichen Arbeiten zu unterstützen.

**Zielsetzung**

Das Ziel ist die Entwicklung eines prototypischen Systems, das deutschsprachige Arbeiten aus der Wirtschaftsinformatik hinsichtlich definitorischer Vollständigkeit, argumentativer Bezüge im Schlussteil und semantischer Kohärenz prüft. Der Arbeit liegen dabei folgende Forschungsfragen zugrunde:

1. Wie kann automatisch überprüft werden, ob in wissenschaftlichen Arbeiten verwendete Fachbegriffe definiert wurden?
2. Wie kann überprüft werden, ob Autoren im Schlussteil einer wissenschaftlichen Arbeit auf vorhergehende Theorien und Konzepte Bezug nehmen und diese argumentativ stützen?
3. Wie kann analysiert werden, ob Textteile in wissenschaftlichen Arbeiten semantisch kohärent sind?

**Vorgehensweise**

Diese Arbeit ist in acht Kapitel gegliedert. Das erste Kapitel leitet das Thema und die zugrunde liegenden Forschungsfragen ein. Im zweiten Kapitel werden die notwendigen sprachwissenschaftlichen Grundlagen erklärt. Anschließend werden die technischen Grundlagen dargelegt. Ausgehend von den Forschungsfragen erfolgt eine ausführliche Darlegung des Forschungsstands unter Einbeziehung thematisch verwandter Arbeiten. Die anschließende Beantwortung der Forschungsfragen erfolgt in einem mehrteiligen Verfahren. Zunächst werden die Forschungsfragen unter Berücksichtigung erarbeiteter Grundlagen und verwandter Arbeiten in softwaretechnische Systemanforderungen überführt. Weiterhin werden geeignete Technologien aufgestellt und ein vollständiges Konzept mit Architekturdiagrammen aufgestellt. Im nächsten Schritt wird die Implementierung der

notwendigen Softwareartefakte erläutert. Nachdem die Implementierung abgeschlossen ist, werden die entwickelten Sprachmodelle anhand von aufgestellten Metriken evaluiert. Zusätzlich werden die hervorgebrachten Ergebnisse entsprechend der Systemanforderungen validiert. Schlussendlich mündet die Arbeit in ein Schlusskapitel, in dem die Beantwortung der Forschungsfragen sowie ein Fazit und Ausblick im Vordergrund stehen.

**Umsetzung**

Zur methodischen Umsetzung werden vortrainierte Transformer-Sprachmodelle herangezogen und durch Transfer Learning auf die Aufgaben feinabgestimmt. Die Transformer-Modelle basieren auf dem Self-Attention-Mechanismus, der es ihnen erlaubt, semantische Beziehungen zwischen Wörtern auch über größere Textkontexte hinweg zu erfassen. Im Hinblick auf die Forschungsfragen ist eine kontextübergreifende Verarbeitung erforderlich, die durch Transformer sichergestellt werden können. Die bisherigen Ansätze verwandter Arbeiten zeigen auf, dass Transformer für Klassifikationsaufgaben auf Wort- und Satzebene herangezogen werden sollten. Die Arbeit umfasst die Modellierung und Erstellung von Datensätzen. Als Basis dienen drei zentrale Datenstrategien, die aus der Nutzung eines Wikipedia-Datensatzes, der Erstellung eines Datensatzes mit Hilfe von existierenden wissenschaftlichen Arbeiten und der Erstellung eines Datensatzes mittels generativer KI bestehen. Die Annotation der Datensätze erfolgte im Falle der ersten Datenstrategie manuell, indem Merkmale von Definitionen innerhalb der Texte markiert werden. Schließlich folgt die Auswahl von Sprachmodellen, die für die Klassifikation von wissenschaftlichen Texten in der Wirtschaftsinformatik geeignet sein könnten. Hierzu fällt die Auswahl unter anderem auf die Modelle deepset/gbert-base und deepset/gbert-large, da diese auf der originalen BERT-Architektur aufbauen und darüber hinaus auf großen deutschsprachigen Textkorpora trainiert wurden, weshalb diese für die Analyse von Qualitätsmerkmalen in wissenschaftlichen Texten geeignet sein könnten. Es folgt ein Fine-Tuning der selektierten Sprachmodelle auf die Zielaufgaben unter Einsatz der zuvor entwickelten Datensätze. Zusätzlich wird eine komplexe Systemarchitektur bestehend aus Frontend und Backend implementiert, die den explorativen Einsatz der entwickelten Sprachmodelle ermöglicht. Zu diesem Zweck werden die Python-Bibliotheken Flask und Streamlit herangezogen. Zusätzlich wird die Themenmodellierung zur Extraktion von Themen in Kapiteln und Argumenten implementiert. Ergänzend erfolgt die Berechnung der Kosinus-Ähnlichkeit dieser Themen, um semantische Nähe von Sätzen und Kapiteln festzustellen. Schlussendlich werden die Resultate in die Anwendung eingearbeitet, sodass Nutzer eine Analyse von PDF-Dateien und einzelnen Textpassagen vornehmen können und passende Ergebnisse erhalten. Zusätzlich existiert eine Anzeige von Explainable Artificial Intelligence (XAI) Ausgaben, um eine Nachvollziehbarkeit der durch die Modelle vorgegebenen Ergebnisse zu ermöglichen.

**Ergebnisse und Fazit**

Die Ergebnisse zeigen, dass die zugrundeliegenden Charakteristika von wissenschaftlichen Arbeiten zuverlässig erkannt und analysiert werden können. Eine Erkennung von Definitionen wird mit einer Genauigkeit von 96,3% durch das feinabgestimmte gbert-base-Modell erfüllt. Die Klassifikation von Textteilen erfolgt unter einer Genauigkeit von 78,5% sowie die Klassifikation von Argumenten, Gegenargumenten und Konklusionen im Fazit unter einer Genauigkeit von 96,6%. Zuletzt bringt das Fine-Tuning von gbert-large eine Erkennung von Kohärenz und Kohäsion in Texten mit einer Genauigkeit von 95,3% hervor. Damit wird deutlich, dass die in der vorliegenden Arbeit angewendete Klassifikation zur Beantwortung der Forschungsfragen zielführend ist. Gleichzeitig wird erkennbar, dass die Qualität der Resultate maßgeblich von der Beschaffenheit der Datensätze und Modelle abhängt. Eine konsequente Anwendung eines „Multi-Annotator-Ansatzes“ erscheint sinnvoll, um die Konsistenz und Ausgewogenheit der Trainingsdaten sicherzustellen. Ebenfalls stellen die erweiterte Anwendung von Data Augmentation, die Berücksichtigung von weiteren Mustern von Kohärenz und Kohäsion sowie die Einbettung von Definitionen über mehrere Sätze eine potentielle Verbesserung der Erkennungsleistung dar. Auch der Einsatz von Explainable Artificial Intelligence zur Klärung der prognostizierten Resultate durch die Sprachmodelle erweist sich als vielversprechend, jedoch zeitaufwändig. Zukünftige Arbeiten sollten daher den Fokus auf eine effiziente Erklärbarkeit der Ergebnisse legen. Perspektivisch kann der vorgestellte Ansatz dazu beitragen, ein intelligentes Tutorensystem zu entwickeln, das Studierende bei der Erstellung wissenschaftlicher Texte unterstützt und die Bewertung durch Lehrende ergänzt.

**Literatur**

L. J. Heinrich, R. Riedl, und A. Heinzl, „Wissenschaftscharakter der Wirtschaftsinformatik“, in Wirtschaftsinformatik: Einführung und Grundlegung, L. J. Heinrich, A. Heinzl, und R. Riedl, Hrsgg. Berlin, Heidelberg: Springer,2011, S. 47–60. Online verfügbar: https://doi.org/10.1007/978-3-642-15426-3_4 (abgerufen: 24.04.2025).

L. J. Heinrich, R. Riedl, und A. Heinzl, „Fachsprache der Wirtschaftsinformatik“, in Wirtschaftsinformatik: Einführung und Grundlegung, L. J. Heinrich, A. Heinzl, und R. Riedl, Hrsgg. Berlin, Heidelberg: Springer, 2011, S. 61–72. Online verfügbar: https://doi.org/10.1007/978-3-642-15426-3_5 (abgerufen: 24.04.2025).

R. Musan, „Kohäsion und Kohärenz“, in Linguistik im Sprachvergleich: Germanistik– Romanistik – Anglistik,

R. Klabunde, W. Mihatsch, und S. Dipper, Hrsgg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2022, S. 577–593.

D. Jurafsky und J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3. Aufl., 2025. Online verfügbar: https://web.stanford.edu/~jurafsky/slp3/ (abgerufen: 01.02.2025).

A. Vaswani et al., „Attention Is All You Need“, Aug. 2023. Online verfügbar: http://arxiv.org/abs/1706.03762 (abgerufen: 03.03.2025). ArXiv:1706.03762 [cs].

J. Devlin, M.-W. Chang, K. Lee, und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, Mai 2019. Online verfügbar: http://arxiv.org/abs/1810.04805 (abgerufen: 16.12.2024). ArXiv:1810.04805 [cs].

A. Storrer und S. Wellinghoff, „Automated detection and annotation of term definitions in German text corpora“, in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC‘06). Genoa, Italy: European Language Resources Association (ELRA), Mai 2006. Online verfügbar: https://aclanthology.org/L06-1066/ (abgerufen: 23.03.2025).

A.-M. Avram, D.-C. Cercel, und C. Chiru, „UPB at SemEval-2020 Task 6: Pretrained Language Models for Definition Extraction“, in Proceedings of the Fourteenth Workshop on Semantic Evaluation, A. Herbelot, X. Zhu, A. Palmer,N. Schneider, J. May, und E. Shutova, Hrsgg. Barcelona (online): International Committee for Computational Linguistics, Dez. 2020, S. 737–745. Online verfügbar: https://aclanthology.org/2020.semeval-1.97/ (abgerufen: 13.01.2025).

S. Teufel, A. Siddharthan, und C. Batchelor, „Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics“, in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, P. Koehn und R. Mihalcea, Hrsgg. Singapore: Association for Computational Linguistics, Aug. 2009, S. 1493–1502. Online verfügbar: https://aclanthology.org/D09-1155/ (abgerufen: 25.01.2025).

O. Kashefi, S. Chan, und S. Somasundaran, „Argument Detection in Student Essays under Resource Constraints“, in Proceedings of the 10th Workshop on Argument Mining, M. Alshomary, C.-C. Chen, S. Muresan, J. Park, und J. Romberg, Hrsgg. Singapore: Association for Computational Linguistics, Dez.2023, S. 64–75. Online verfügbar: https://aclanthology.org/2023.argmining-1.7/ (abgerufen: 09.01.2025).

A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, und J. Qi, „Evaluating Document Coherence Modeling“, Transactions of the Association for Computational Linguistics, Bd. 9, S. 621–640, Juli 2021. Online verfügbar: https://doi.org/10.1162/tacl_a_00388 (abgerufen: 26.11.2024).

P. Törnberg, „Best Practices for Text Annotation with Large Language Models“, Febr. 2024. Online verfügbar: http://arxiv.org/abs/2402.05129 (abgerufen:23.02.2025).

B. Chan, S. Schweter, und T. Möller, „German’s Next Language Model“, Dez. 2020. Online verfügbar: http://arxiv.org/abs/2010.10906 (abgerufen: 26.03.2025).

# Evaluation KI-basierter Lösungen zur Extraktion von Informationen aus Diagrammen

Mirko Sprcic
Technische Hochschule Mittelhessen
Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
mirko.sprcic@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**

Masterarbeit

**Schlüsselwörter**

Informationsextraktion, Künstliche Intelligenz, Deep-Learning, Computer Vision, Convolutional Neural Networks, Vision Transformer, Objekterkennung, Linienextraktion, Optical Character Recognition, YOLO, PyTorch, HuggingFace

**Zusammenfassung**

Neben der Informationsextraktion (IE) aus unstrukturierten Daten im Bereich natürlicher Sprache werden zunehmend Inhalte aus Bildern verarbeitet. Diese werden unter anderem zur Auswertung von Unternehmenskennzahlen oder zur Verbesserung der Zugänglichkeit von Inhalten in Bildern für Menschen mit Seheinschränkungen verwendet.

Ein weiterer Anwendungsfall existiert im Hochschulkontext. Abgaben von Studierenden zu Fallstudien oder Übungen bestehen häufig aus Bildern, die anschließend von Dozenten oder wissenschaftlichen Mitarbeitenden manuell kontrolliert und bewertet werden. Sind große Mengen an Bildern zu analysieren, so ist dieser Prozess für einzelne Personen sehr zeitaufwendig, weshalb Systeme zu entwickeln sind, die Inhalte in Bildern automatisiert extrahieren und weiterführenden Systemen zur Verfügung stellen.

Hierfür wird im Rahmen dieser Arbeit zunächst eine Pipeline zur IE aus Bildern konzipiert, die auf bisherigen Ansätzen und spezifischen Anforderungen des betrachteten Anwendungsfalls basiert. Im Rahmen der Pipeline werden Deep-Learning (DL)-Modelle eingesetzt, die für Aufgaben wie Bildklassifizierung oder Objekterkennung optimiert sind. Die Auswahl des verwendeten Modells für den jeweiligen Prozessschritt erfolgt durch eine Evaluation bestehender Modelle. Abschließend wird die erarbeitete Pipeline mit den evaluierten Modellen als Prototyp implementiert.

Der betrachtete Anwendungsfall besteht aus Abgaben zu Fallstudien des Wirtschaftsinformatik-Moduls Business Intelligence der Technischen Hochschule Mittelhessen. Die Abgaben umfassen eine Vielzahl an Bildschirmaufnahmen der SAP Business Warehouse Umgebung, darunter Tabellen, Transformationstabellen, Datenflussdiagramme, Eingabemasken und Bilder mit Mischformen der genannten Typen. Innerhalb der Bilder sind sowohl grafische als auch textliche Inhalte zu extrahieren und in Beziehung zueinander zu setzen.

Die entwickelte Pipeline besteht aus fünf Prozessschritten. Bilder werden zunächst klassifiziert, da sich die nachfolgenden Schritte in Abhängigkeit vom jeweiligen Typ unterscheiden. Handelt es sich bei dem zu verarbeitenden Bild um eine Mischform mehrerer Typen, wird das Bild zunächst in seine einzelnen Bestandteile zerlegt. Anschließend folgt die diagrammtypspezifische Extraktion zur Lokalisierung grafischer Elemente, wie Zeilen und Spalten in Tabellen oder Objekten in Datenflussdiagrammen. Hierfür werden überwiegend DL-Modelle verwendet, ergänzt durch regelbasierte Methoden zur Linienextraktion. Abschließend erfolgt die Erkennung von Texten mittels Optical Character Recognition (OCR) sowie die Zusammenführung der extrahierten Inhalte in einer einheitlichen JSON-Struktur.

Für die Klassifikation und Objekterkennung werden Modelle der Anbieter PyTorch, HuggingFace und Ultralytics evaluiert. Bei den Modellen handelt es sich um vortrainierte Modelle, die bereits mit großen Datenmengen trainiert wurden und dadurch in der Lage sind, relevante Merkmale aus Bildern zu extrahieren. Die Modelle werden mit Daten des Anwendungsfalls abgestimmt, um sie für die jeweilige Aufgabe nutzbar zu machen. Im Rahmen der Evaluation werden Modelle unterschiedlicher Architekturen und Größen betrachtet.

Neben den etablierten Convolutional Neural Networks (CNNs) werden Transformer-basierte Architekturen eingesetzt, die bereits im Bereich der Verarbeitung natürlicher Sprache erfolgreich verwendet werden.

Im Rahmen der Klassifikation erzielt der Großteil der Modelle einen gemittelten F1-Score von über 95 %. Unterschiede zeigen sich insbesondere in der Modellgröße und der erforderlichen Trainingszeit. Hier überzeugen besonders die Modelle RegNet, GoogLeNet und SqueezeNet.

Bei der Objekterkennung erzielen die von Ultralytics bereitgestellten YOLO-Modelle über sämtliche Aufgaben hinweg sehr gute Ergebnisse, mit Mean Average Precision (mAP)@50-Werten von über 98 % und mAP@50–95-Werten von über 80 %. Transformer-basierte DETR-Modelle liefern bei klar voneinander abgegrenzten Inhalten vergleichbare Ergebnisse, schneiden jedoch bei der Erkennung kleiner oder eng beieinanderliegender Objekte schwächer ab. Hier überzeugen hingegen CNN-Modelle durch ihre Fähigkeit, lokale Merkmale präziser zu erkennen.

Im Bereich OCR werden die Tools Tesseract, EasyOCR und PaddleOCR miteinander verglichen. Unterschiede zeigen sich bei der Verwendung mehrerer Sprachen. Während Tesseract und EasyOCR die Konfiguration mehrerer Sprachen gleichzeitig erlauben, unterstützt PaddleOCR ausschließlich die Nutzung einer Sprache pro Ausführung. Bei der Erkennungsqualität liefern EasyOCR und PaddleOCR mit einer Character Error Rate von 8,5 % bzw. 10,5 % ähnliche Ergebnisse, während Tesseract mit 30 % deutlich mehr Fehler erzeugt. Häufige Fehlerquellen treten bei der Erkennung ähnlich aussehender Zeichen auf, insbesondere bei Zahlen und Buchstaben.

Für die abschließende Zusammenführung der extrahierten Informationen werden regelbasierte Methoden eingesetzt. Dabei werden Abstände zwischen erkannten Objekten berechnet und mit definierten Toleranzen abgeglichen. Für die betrachteten Diagrammtypen funktioniert dieser Ansatz zuverlässig, allerdings ist bei der Einbindung weiterer Diagrammtypen zusätzlicher Entwicklungsaufwand erforderlich, da diese jeweils über eigene Regeln zur Zusammenführung verfügen.

Insgesamt zeigt die Arbeit, dass eine automatisierte Extraktion von Inhalten aus unterschiedlichen Diagrammtypen durch die Kombination vortrainierter DL-Modelle und regelbasierter Methoden erfolgreich umsetzbar ist. Der Einsatz regelbasierter Methoden zur Zusammenführung der Daten erfordert jedoch Codeanpassungen. Die Integration von Large Language Models (LLMs) ermöglicht eine Reduzierung dieser Methoden und der damit verbundenen Anpassungen bei Erweiterungen des Systems. LLMs könnten auf Basis der extrahierten Inhalte sowie eines diagramm-typspezifischen Prompts, der die gewünschte Zusammenführung beschreibt, die finalen Strukturen erzeugen. Der Aufwand zur Erweiterung um neue Diagrammtypen würde sich dadurch auf das Training geeigneter DL-Modelle zur Objekterkennung und die Erstellung eines Prompts, der das Verhalten des Sprachmodells steuert, reduzieren. Voraussetzung hierfür ist jedoch der Einsatz leistungsfähiger LLMs mit entsprechender Rechenkapazität, da kleinere Modelle aktuell noch nicht über die erforderlichen Fähigkeiten im Bereich der Sprachverarbeitung verfügen.

**Literatur**


Charles, P. K. et al. (2012): A Review on the Various Techniques used for Optical Character Recognition. In: International Journal of Engineering Research and Applications 2 (1), S. 659–662.

Choi, J. et al. (2019): Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. In: Computer Graphics Forum 38 (3), S. 249–260. DOI: 10.1111/cgf.13686.

Cowie, J.; Lehnert, W. (1996): Information extraction. In: Commun. ACM 39 (1), S. 80–91. DOI: 10.1145/234173.234209.

Davila, K. et al. (2021): ICPR 2020 - Competition on Harvesting Raw Tables from Infographics. In: Pattern Recognition. ICPR International Workshops and Challenges 12668, S. 361–380. DOI: 10.1007/978-3-030-68793-9_27.

Dosovitskiy, A. et al. (2020): An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Online verfügbar unter http://arxiv.org/pdf/2010.11929v2.

Liu, F. et al. (2023): DePlot: One-shot visual language reasoning by plot-to-table translation. In: Findings of the Association for Computational Linguistics: ACL 2023, S. 10381–10399. DOI: 10.18653/v1/2023.findings-acl.660.

Luo, J. et al. (2021): ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), S. 1916–1924. DOI: 10.1109/WACV48630.2021.00196.

Maheshwari, K. et al. (2020): Convolutional Neural Networks: A Bottom-Up Approach. In: Deep Learning: De Gruyter, S. 21–50. DOI: 10.1515/9783110670905-002.

Masry, A. et al. (2023): UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. In: Proceedings of the 2023 Conference on

Empirical Methods in Natural Language Processing, S. 14662–14684. DOI: 10.18653/v1/2023.emnlp-main.906.

Mustafa, O. et al. (2023): ChartEye: A Deep Learning Framework for Chart Information Extraction. In: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA), S. 554– 561. DOI: 10.1109/DICTA60407.2023.00082.

Wang, H. et al. (2023): Pre-Trained Language Models and Their Applications. In: Engineering 25, S. 51–65. DOI: 10.1016/j.eng.2022.04.024.

Zhang, H. et al. (2013): Text extraction from natural scene image: A survey. In: Neurocomputing 122, S. 310–323. DOI: 10.1016/j.neucom.2013.05.037.

# SAP-Joule im Unternehmenseinsatz: Potenziale und Mehrwerte generativer KI-Plattformen

Miles Trierweiler

Hochschule Pforzheim
Tiefenbronner Straße 65
75175 Pforzheim
Milestrierweiler1995@gmail.com

Frank Morelli

Hochschule Pforzheim
Tiefenbronner Straße 65
75175 Pforzheim
frank.morelli@hs-pforzheim.de

**ABSTRACT**

Der Beitrag verdichtet zentrale Ergebnisse einer Bachelorarbeit zu SAP-Joule. Ziel ist es, SAP-Joule nicht als isolierten Copiloten, sondern als Architekturbestandteil im SAP-Ökosystem zu betrachten und zugleich das ökonomische Wirkpotenzial anhand veröffentlichter Use Cases systematisch einzuordnen. Methodisch kombiniert die Arbeit eine Literatur- und Dokumentenanalyse mit einer Sekundäranalyse von 39 publizierten SAP-Claims aus sechs Funktionsbereichen. Der Beitrag leitet daraus Handlungsempfehlungen für eine skalierbare und wertorientierte Einführung von SAP-Joule ab.

**SCHLÜSSELWÖRTER**

SAP-Joule, generative KI, SAP BTP, KI-Agenten, Business Case, Wirtschaftlichkeitsanalyse, Governance

## 1 EINLEITUNG UND PROBLEMSTELLUNG

Generative KI verschiebt Prioritäten in Technologie-, Prozess- und Investitionsentscheidungen. Im SAP-Kontext adressiert SAP-Joule den Bedarf nach einer prozessnahen Einbettung generativer KI in Unternehmensanwendungen. Im Unterschied zu horizontalen Assistenten ist Joule in das SAP-Ökosystem und die SAP Business Technology Platform (BTP) eingebettet und greift auf Geschäftsprozesse, Datenkontexte und Anwendungslogiken der Suite zu. Ziel des Beitrags ist es, zentrale Architektur- und Einführungsaspekte von SAP-Joule zusammenzufassen und den empirischen Teil der zugrunde liegenden Thesis in verdichteter Form darzustellen. Im Mittelpunkt steht die Frage, welche Potenziale SAP-Joule in mittelständischen Unternehmenskontexten entfalten kann und welche Bedingungen für eine wertstiftende Einführung erfüllt sein müssen.

## 2 SAP-JOULE ALS PLATFORMBAUSTEIN IM SAP-ÖKOSYSTEM

Joule ist nicht nur Chat-Oberfläche, sondern Vermittlungs- und Orchestrierungsschicht zwischen Nutzern, SAP-Anwendungen und KI-Diensten. Relevant sind insbesondere die Einbettung in die BTP, der Zugriff auf Kontexte über Knowledge Graph und Business Data Cloud, die Nutzung von Document Grounding zur Reduktion nicht belastbarer Antworten sowie die Erweiterbarkeit über Joule Studio. Damit entsteht ein Architekturmodell, in dem Standard-Fähigkeiten und kundenspezifische Agenten kombiniert werden können. Für Unternehmen ist das deshalb relevant, weil der Nutzen nicht allein aus dem Modell selbst entsteht, sondern aus der kontrollierten Verbindung von Fachkontext, Daten, Rollen, Schnittstellen und Ausführungslogik. Gleichzeitig verschiebt sich der Einführungsfokus weg vom isolierten Pilot mit einzelnen Prompts hin zu Themen wie Identitätsmanagement, Berechtigung, Governance, Monitoring, Kostenkontrolle und Rolloutfähigkeit.

### 2.1 ARCHITEKTUR, KONTEXT UND ERWEITBARKEIT

Architektonisch verortet die Arbeit SAP-Joule nicht als isolierte Assistenzfunktion, sondern als Vermittlungs- und Orchestrierungsschicht auf der SAP Business Technology Platform. Im Zentrum stehen der Zugriff auf domänenspezifischen Geschäftskontext über Knowledge Graph und Business Data Cloud, die Anbindung an KI-Dienste über die BTP sowie die kontrollierte Einbindung unternehmensinterner Dokumente über Document Grounding. Gerade dieser Punkt ist für den Unternehmenseinsatz wesentlich: Antworten sollen nicht nur aus allgemeinem Modellwissen entstehen, sondern aus freigegebenen Datenquellen, Rollenrechten und Prozesszusammenhängen. Ergänzend zeigt die Arbeit, dass Joule Studio den Standardumfang um Skills und Agents erweitert. Während Skills eher punktuelle Aktionen ausführen, adressieren Agents mehrstufige und bereichsübergreifende Abläufe. Damit verschiebt sich der Blick von der einzelnen Chat-Interaktion hin zu orchestrierten Prozessketten.

### 2.2 INTEGRATION, IDENTITÄT UND GOVERNANCE

Die technische Einbettung ist eng mit Integrations- und Governance-Fragen verbunden. Die Thesis unterscheidet konsolidierte Single-Identity-Authentication-Service-Szenarien, getrennte Multi-Identity-Authentication-Service-Domänen und hybride Konstellationen aus Public und Private Edition. Daraus ergeben sich unterschiedliche Anforderungen an Identitätsdienste, Rollenmodelle, AppRouter, Destinations, Connectivity und Berechtigungsprüfungen. Für Unternehmen ist dies deshalb relevant, weil der Nutzen von Joule nicht allein an der Mo-

dellqualität hängt, sondern an der kontrollierten Verbindung von Datenzugriff, Prozesskontext und Compliance. Hinzu kommt ein konsumptionsbasiertes Betriebsmodell: AI-Units, Grounding-Zugriffe und Agentenaufrufe müssen transparent budgetiert, überwacht und in Kosten-Nutzen-Logiken überführt werden. Die Einführung ist daher nicht nur ein Technologieprojekt, sondern zugleich ein Architektur-, Sicherheits- und Financial Operations (FinOps)-Vorhaben.

## 3 METHODISCHES VORGEHEN UND EMPIRISCHER AUFBAU

### 3.1 FORSCHUNGSDESIGN UND DATENGRUNDLAGE

Die zugrunde liegende Arbeit kombiniert eine systematische Literatur- und Dokumentenanalyse mit einer Sekundäranalyse veröffentlichter Joule-Use-Cases. Für die empirische Verdichtung wurden 39 publizierte SAP-Claims aus sechs Bereichen - Finance, Procurement, Supply Chain, Customer Experience, Technology Platform (BTP/IT) und Consulting - in einer strukturierten Excel-Anlage erfasst, klassifiziert und weiterverarbeitet. Die Analyse zielt nicht auf Kausalitätsnachweise, sondern auf eine transparente und nachvollziehbare Einordnung potenzieller Nutzenbeiträge unter expliziten Annahmen.

| Area | Process/Use Case | Statement | Value | Effect |
|---|---|---|---|---|
| Finance | Debitorenbuchhaltung (AR) – Zahlungsklärung | Reduzierung des Aufwands für den Abgleich in der Debitorenbuchhaltung um 71 %* | 71% | Reduktion (Aufwand) |
| Procurement | Sourcing – RfX Erstellung | Ausschreibungen 70 % schneller erstellen | 70% | Beschleunigung |
| Supply Chain | Produktentwicklung – Ideen/Innovation | 50 % niedrigere Erstellungskosten und 1 % Umsatzsteigerung durch neue Produkte* | 50% | Kostensenkung |
| Customer Experience (CX) | Sales – Alltagsaufgaben | Complete everyday sales tasks 80% faster* | 80% | Beschleunigung |
| Technology Platform (BTP/IT) | Build Code – App-Entwicklungskosten | 30 % geringere App-Entwicklungskosten* | 30% | Kostensenkung |
| Consulting | Code-Interpretation | Interpretationsaufwand um bis zu 40 % senken | 40% | Reduktion (Aufwand) |

*Abbildung 1: SAP-Claims, eine Zeile je Area*

Als Datengrundlage dienen die von SAP veröffentlichten Joule-Use-Cases und Wirkungsangaben (Prozentwerte bzw. Zeitersparnisse). Im vorliegenden Kontext wurden die entsprechenden Einträge in eine strukturierte Tabelle (vgl. Abb. 1) transformiert und nach Effekttypen klassifiziert.

### 3.2 OPERATIONALISIERUNG UND MONETARISIERUNG

Die Claims wurden zunächst in homogene Wirkungspfade überführt (time, cost, revenue, revenue_other, other). Diese Trennung ist erforderlich, weil Aussagen wie "80 % schneller" und "2 % mehr Umsatz" ohne Referenzwert nicht direkt vergleichbar sind. Die Monetarisierung erfolgte über ein Mittelstandsprofil mit expliziten Annahmen, darunter ein Vollkostensatz von 60 EUR pro Stunde, 1.760 Stunden pro FTE (Full Time Equivalent) und Jahr, eine Bruttomarge von 30 Prozent, ein Diskontsatz von 8 Prozent sowie eine stufenweise Adoption von 50, 75 und 100 Prozent über drei Jahre. Unsichere Claims wurden konservativ behandelt: Aussagen ohne belastbare Quellen erhielten einen Evidence-Multiplikator von 0,7, "bis zu"-Aussagen wurden mit einem UpTo-Multiplikator von 0,85 gedämpft. Auf dieser Basis wurden die jährlichen Nutzenwerte je Use Case und Szenario berechnet und zu einem 3-Jahres-Net Present Value verdichtet. Dabei wurde folgende Formel zu Grunde gelegt:

$$NPV_{3y} = \sum_{j=1}^{3} \frac{Y_j \times Nutzen_j}{(1+r)^j}, \quad r = 0{,}08, (Y_1, Y_2, Y_3) = (0{,}50, 0{,}75, 1{,}00)$$

Für die Instrumentenentwicklung erfolgte die Konzeption einer Nutzwertanalyse. Die dafür notwendigen sechzehn Bewertungskriterien, die in der ersten Phase aus den Experteninterviews abgeleitet wurden, mussten objektiv gewichtet werden. Dies erfolgte durch einen systematischen Paarvergleich nach der AHP-Methode. Der Expertenkreis musste hierbei für jedes Kriterienpaar entscheiden, welches als wichtiger erachtet wird.

### 3.3 SZENARIOANALYSE UND ROBUSTHEIT

Ein wichtiger Teil des empirischen Designs ist die systematische Behandlung von Unsicherheit. Die Excel-Anlage bildet nicht nur die Rohdaten ab, sondern auch die gesamte Transformationskette von der Klassifikation der Claims über Referenzwerte und Jahresnutzen bis zur Aggregation auf Bereichsebene. Für die Verdichtung werden vier Szenarien verwendet: Raw, Realistic, Conservative und Hard. Das realistische (Realistic)-Szenario reduziert die publizierten Wirkungen systematisch, zusätzliche Dämpfungen greifen bei unbelegten Claims und bei Formulierungen wie „bis zu“. Damit wird bewusst keine direkte Wirkungsgarantie unterstellt, sondern eine nachvollziehbare Mittelposition modelliert. Gleichzeitig bleibt das Modell offen für Was-wäre-wenn-Analysen, da Stundensätze, Margen, Adoptionsverläufe und Diskontsätze variabel parametriert werden können. Diese Transparenz erhöht die Nachvollziehbarkeit und macht die Ergebnisse für unternehmensspezifische Validierungen anschlussfähig.

## 4 ERGEBNISSE

### 4.1 DESKRIPTIVE MUSTER

Die deskriptive Auswertung zeigt (vgl. Abb. 2), dass Zeiteffekte die Datenlage dominieren. Besonders hohe durchschnittliche Zeiteffekte finden sich in BTP/IT (86,7 %), Procurement (81,5 %) und Customer Experience (80,0 %). Kosteneffekte sind vor allem in Finance und Supply Chain relevant, während Umsatz-orientierte Claims fast ausschließlich in Supply Chain und Customer Experience auftreten. Schon auf dieser Ebene zeigt sich, dass die Bereiche nicht nur unterschiedlich viele Claims aufweisen, sondern strukturell unterschiedliche Nutzenprofile besitzen.

| Area | Claims_Count_time | Avg_Value_Percent_time | Metric |
|---|---|---|---|
| Supply Chain | 5 | 53,0% | % |
| Finance | 3 | 60,0% | % |
| Technology Platform (BTP/IT) | 3 | 86,7% | % |
| Consulting | 3 | 20,9% | % |
| Customer Experience (CX) | 1 | 80,0% | % |
| Procurement | 4 | 81,5% | % |

*Abbildung 2: Deskriptive Ergebnisse. Verteilung nach time*

### 4.2 MONETÄRE VERDICHTUNG UND PRIORISIERUNG

Im realistischen Szenario ergeben sich auf Bereichsebene die höchsten absoluten 3-Jahres-Net-Present-Value (NPV)-Werte in Supply Chain (ca. 2,14 Mio. EUR) und Customer Experience (ca. 1,65 Mio. EUR). Dahinter folgen BTP/IT (ca. 0,77 Mio. EUR), Finance (ca. 0,59 Mio. EUR), Procurement (ca. 0,55 Mio. EUR) und Consulting (ca. 0,22 Mio. EUR). Die Top-Use-Cases stammen unter anderem aus Customer Experience, Supply Chain und BTP/IT; besonders stark fallen vertriebsnahe Umsatz- und Produktivitätsfälle, Außendienst-Produktivität, Kostenreduktionen im Wareneingang und die Joule-gestützte Informationssuche ins Gewicht.

Für eine faire Priorisierung reichen absolute Summen jedoch nicht aus, da der Supply Chain Bereich deutlich mehr Use Cases umfasst als andere. Nach durchschnittlichem NPV je Use Case liegt Customer Experience mit rund 413 TEUR pro Fall klar vorne, gefolgt von BTP/IT mit rund 153 TEUR. Procurement, Supply Chain und Finance liegen enger beieinander; Consulting bildet das Schlusslicht. Wird zusätzlich auf die Effizienz pro FTE normalisiert und nur der dominierende Zeitpfad betrachtet, verschiebt sich die Rangfolge erneut: BTP/IT erreicht rund 113 TEUR NPV pro FTE, Customer Experience rund 112 TEUR, Finance und Procurement jeweils knapp 100 TEUR. Supply Chain fällt trotz hoher Gesamtsumme zurück, da die zugrunde liegenden Ausgangswerte dort ressourcenintensiver sind.

### 4.3 FAIR METRICS UND PRIORISIERUNGSFÄHIGE VERGLEICHSACHSEN

Die Arbeit zeigt zudem, dass eine belastbare Rollout-Priorisierung nur gelingt, wenn mehrere Vergleichsachsen parallel betrachtet werden. Absolute NPV-Summen bevorzugen Bereiche mit vielen Claims oder sehr großen Ausgangswerten; das spricht für Supply Chain als größten Wertpool, sagt aber noch wenig über die Eignung als Startfeld aus. Deshalb werden ergänzend der durchschnittliche Kapitalwert je Use Case (vgl. Abb. 3) und die Effizienz pro FTE herangezogen (vgl. Abb. 4). Im realistischen Szenario liegt Customer Experience mit rund 413 Tsd. EUR je Use Case deutlich vorn, gefolgt von BTP/IT mit rund 153 Tsd. EUR, Procurement mit rund 137 Tsd. EUR, Supply Chain mit rund 134 Tsd. EUR und Finance mit rund 119 Tsd. EUR. Wird zusätzlich die FTE-basierte Effizienz betrachtet, rücken BTP/IT und Customer Experience erneut an die Spitze, während Finance die höchste Effizienz pro 100 Tsd. EUR Ausgangsbasis aufweist. Für mittelständische Kontexte ist diese Mehrfachsicht besonders relevant, weil sie zwischen großem Wertpool, starkem Einzelfall und ressourceneffizientem Startfeld unterscheidet.

| Area | Sum NPV (EUR) | Distinct Use Cases (#) | Avg NPV per Use Case (EUR) | Sum Baseline (EUR) | NPV per 100k Baseline (EUR) | Fair Score (Avg NPV per UC) | Rank (by Fair Score) |
|---|---|---|---|---|---|---|---|
| Finance | 592.979,04 € | 5 | 118.595,81 € | 586.080,00 € | 101.177,15 € | 118.595,81 € | 5 |
| Procurement | 546.987,18 € | 4 | 136.746,80 € | 580.800,00 € | 94.178,23 € | 136.746,80 € | 3 |
| Supply Chain | 2.142.020,69 € | 16 | 133.876,29 € | 6.614.320,00 € | 32.384,59 € | 133.876,29 € | 4 |
| Customer Experience (CX) | 1.650.309,12 € | 4 | 412.577,28 € | 4.821.800,00 € | 34.226,00 € | 412.577,28 € | 1 |
| Technology Platform (BTP/IT) | 765.424,59 € | 5 | 153.084,92 € | 1.228.000,00 € | 62.330,99 € | 153.084,92 € | 2 |
| Consulting | 217.317,93 € | 5 | 43.463,59 € | 886.800,00 € | 24.505,86 € | 43.463,59 € | 6 |

*Abbildung 3: Normalisierung und faire Bewertung pro Use Case*

| Area | Sum NPV mode time(EUR) | Sum Baseline Mode = time(EUR) | AVG NPV per time Use Case | FTE Basis | Fair Score 2_NPV je FTE | Distinct Use Cases (#) | Rank (by Fair Score) |
|---|---|---|---|---|---|---|---|
| Technology Platform (BTP/IT) | 565.945,66 € | 528.000,00 € | 188.648,55 € | 5,00 | 113.189,13 € | 3 | 1 |
| Customer Experience (CX) | 337.039,60 € | 316.800,00 € | 84.259,90 € | 3,00 | 112.346,53 € | 4 | 2 |
| Finance | 203.979,17 € | 216.480,00 € | 50.994,79 € | 2,05 | 99.502,04 € | 4 | 3 |
| Procurement | 546.987,18 € | 580.800,00 € | 546.987,18 € | 5,50 | 99.452,21 € | 1 | 4 |
| Supply Chain | 751.317,44 € | 1.024.320,00 € | 250.439,15 € | 9,70 | 77.455,41 € | 3 | 5 |
| Consulting | 119.992,16 € | 616.800,00 € | 39.997,39 € | 5,84 | 20.543,40 € | 3 | 6 |

*Abbildung 4: Normalisierung und faire Bewertung pro Use Case = time FTE*

### 4.4 EINORDNUNG UND GRENZEN

Die Ergebnisse sprechen für ein differenziertes Priorisierungsmodell: Supply Chain adressiert den größten Wertpool, Customer Experience und BTP/IT liefern die stärksten Einzelhebel, und Finance zeigt eine hohe Effizienz bezogen auf die eingesetzte Ausgangsbasis. Gleichzeitig ist zu berücksichtigen, dass die empirische Grundlage aus Herstellerkommunikation und publizierten Kundenbeispielen stammt. Die Studie begegnet dieser Anbieterbefangenheit mit Evidenzgewichtung, Sze-

narien, transparenten Annahmen und einer offen dokumentierten Excel-Logik, ersetzt damit aber keine unternehmensspezifische Validierung. Hinzu kommt, dass heterogene Metriken – Zeit, Kosten, Umsatz und Quoten – erst über getrennte Monetarisierungspfade vergleichbar gemacht werden. Auch mögliche Überlappungen zwischen ähnlichen Use Cases sind bei einer praktischen Umsetzung nicht additiv zu interpretieren. Die Ergebnisse sind daher als belastbare Orientierungsgrößen für die Priorisierung zu verstehen, nicht als direkte Erfolgsgarantie.

## 5 HANDLUNGSEMPFEHLUNGEN FÜR UNTERNEHMEN

Aus den Befunden lassen sich fünf praktische Empfehlungen ableiten. Erstens sollte die Einführung von SAP-Joule nicht als isoliertes IT-Projekt, sondern als Programm mit Zielbild, Nutzenhypothesen und gestuftem Rollout verstanden werden. Zweitens sind Customer Experience und BTP/IT geeignete Startfelder, wenn schnelle und vergleichsweise effiziente Wirkung gesucht wird. Drittens wäre Supply Chain gezielt selektiv zu erschließen, weil der absolute Hebel hoch ist, die effiziente Skalierung aber stärker von Datenqualität und Prozessreife abhängt. Viertens braucht der Aufbau eigener Agents in Joule Studio ein sauberes Zusammenspiel aus Identitätsdiensten, Berechtigungen, Destinations, Grounding-Quellen und Test- bzw. Versionslogik. Fünftens sollte die wirtschaftliche Steuerung früh über Stückkosten erfolgen, also über Budgets, Quoten und Kosten-Nutzen-Kontrollen auf Use-Case- oder Agent-Ebene statt allein über Lizenzgrößen. Flankierend sind Qualifizierung, Pilotierung und eine kommunikativ saubere Veränderungsbegleitung notwendig, damit aus technologischer Verfügbarkeit auch tatsächliche Prozessadoption entsteht.

### 5.1 VOM PILOT ZUM BETRIEBSMODELL

Für mittelständische Unternehmen spricht die Untersuchung gegen einen flächendeckenden Rollout ohne vorgelagerte Nutzenbelege. Ein praktikabler Einstieg erfolgt vielmehr über ein Stufenmodell. In einer ersten Phase sollten Use Cases gewählt werden, bei denen Referenzwerte gut beobachtbar sind und Effekte schnell sichtbar werden, etwa Informationssuche, vertriebsnahe Assistenzfunktionen, administrative Produktivität oder klar abgrenzbare Finance-Fälle. In einer zweiten Phase sind die gemeinsamen Betriebsartefakte zu standardisieren: Rollen- und Berechtigungsmatrix, Katalog zulässiger Grounding-Quellen, Freigabe- und Testlogik für Prompts und Agents, Monitoring für AI-Units sowie Regeln für Budgetgrenzen und Eskalationen. Erst in einer dritten Phase sollte man komplexere, agentische und bereichsübergreifende Szenarien skalieren, etwa im Supply Chain Bereich oder in funktionsübergreifenden Serviceketten. Gerade dort sind zwar die größten absoluten Wertpools zu erwarten, gleichzeitig steigen aber Integrationsgrad, Abstimmungsaufwand und Risiko von Überlappungen. Der zentrale Beitrag der Arbeit liegt daher nicht in einer pauschalen Empfehlung „mehr KI“, sondern in einer priorisierungsfähigen Einführungslogik mit transparenten Nutzenannahmen und klaren Governance-Voraussetzungen.

## 6 FAZIT

SAP-Joule ist im Unternehmenseinsatz vor allem dann interessant, wenn generative KI nah an SAP-Prozessen, Daten und Anwendungen wirken soll. Die Architektur erlaubt eine tiefe Suite-Integration und die Erweiterung durch eigene Agents, erhöht aber zugleich die Anforderungen an Governance, Sicherheit und Kostensteuerung. Die empirische Verdichtung der publizierten Use Cases zeigt ein relevantes wirtschaftliches Potenzial, macht aber ebenso deutlich, dass die Einführung differenziert priorisiert werden sollte: nicht nach maximaler Aufmerksamkeit, sondern nach Wertpool, Einzelwirkung und Ressourceneffizienz. Für mittelständische Kontexte spricht daher vieles für einen gestuften Einstieg über Customer Experience und BTP/IT, während die Supply Chain als großer Wertpool gezielt und nachweisorientiert erschlossen werden sollte. Damit positioniert die Arbeit SAP-Joule weder als Selbstläufer noch als bloßes Marketingnarrativ, sondern als architektonisch und ökonomisch relevantes Gestaltungsfeld, dessen Nutzen von sauberer Integration, belastbarer Datenbasis und konsequenter Steuerung abhängt.

## LITERATUR


**Ameling, Michael** (2024): How SAP BTP Becomes the Trusted Platform for Business AI. SAP News. URL: https://news.sap.com/2024/05/sap-btp-strategy-paper-trusted-platform-business-ai/ (abgerufen am 25.07.2025)

**Balasubramanian, Arun Chinnannan** (2024): SAP-Joule with RAG and Document Grounding: Next-Gen Document-Centric Processes. International Journal of Innovative Research and Creative Technology (IJIRCT), 10(6), S. 1–6, 28. Dezember 2024. DOI: 10.5281/zenodo.14566302. URL: https://www.ijirct.org/viewPaper.php?paperId=2412126 (abgerufen am 21.08.2025).

**Banh, Leonardo; Strobel, Gero** (2023): Generative artificial intelligence. Electronic Markets, 33, Art. 63. DOI: 10.1007/s12525-023-00680-1. URL: https://hdl.handle.net/10419/306304 (abgerufen am 04.10.2025).

**Deloitte** (2025): State of Generative AI in the Enterprise Survey Q4 (July/Sept. 2024). Deloitte Insights. URL: https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html (abgerufen am 25.07.2025)

**Glaubitz, Alina; Busse, Christian; Whiteford, Elizabeth; VanDodick, Jamie; Kirkwood, Jen; Soriano, Julie; Muller, Michael; Weinberg, Nikita; Walls, Richard; Greulich, Sophia** (2025): Accelerating Business Value through AI-Focused Change Management. IBM Office of Privacy and Responsible Technology, April 2025. URL:

https://www.ibm.com/downloads/documents/us-en/12fc84a11ed9542e (abgerufen am 19.08.2025).

**Herzig, Philipp** (2024): Joule to Integrate with Microsoft Copilot for a Unified Work Experience. SAP News Center, 4. Juni 2024. URL: https://news.sap.com/2024/06/joule-microsoft-copilot-unified-work-experience/ (abgerufen am 21.08.2025).

**McKinsey Global Institute** (2023): McKinsey Global Institute (2023): The economic potential of generative AI: The next productivity frontier. Juni 2023 URL: https://www.mckinsey.de/~/media/mckinsey/locations/europe%20and%20middle%20east/deutschland/news/presse/2023/2023-06-14%20mgi%20genai%20report%2023/the-economic-potential-of-generative-ai-the-next-productivity-frontier-vf.pdf (abgerufen am 21.09.2025).

**MHP** (2025): SAP Business Data Cloud: 8 Fragen, 8 Antworten. Newsroom, 27. März 2025. URL: https://www.mhp.com/de/insights/newsroom/news-detail/view/sap-business-data-cloud (abgerufen am 20.08.2025).

**Roland Berger** (2024): Change management and AI. 8. September 2024. URL: https://www.rolandberger.com/en/Insights/Publications/Change-management-and-AI.html (abgerufen am 04.10.2025).

**SAP SE** (2024a): SAP Business Technology Platform – The Innovation Engine for Your SAP Applications. Strategy White Paper 2024/25. URL: https://www.sap.com/docs/download/2021/06/d29a141c-e47d-0010-bca6-c68f7e60039b.pdf (abgerufen am 21.09.2025).

**SAP SE** (2024b): SAP BTP Security and Compliance – June 2024 (External Version). PDF, Juni 2024. URL: https://d.dam.sap.com/m/4va-Wes7/SAP%20BTP%20Security%20and%20Compliance%20June%202024%20-%20External%20Version.pdf?inline=true (abgerufen am 08.09.2025).

**SAP SE** (2025a): SAP Business AI: Revolutionary technology. Real-World results. White Paper | Public. URL: https://www.sap.com/docs/download/2023/09/bac24f28-a57e-0010-bca6-c68f7e60039b.pdf (abgerufen am 20.08.2025).

**SAP SE** (2025c): Joule. SAP Help Portal, 03. Oktober 2025. URL: https://help.sap.com/doc/75f5965fbb514fbcb89f1e47c094ea3f/CLOUD/en-US/9d8c0ef2443b49f993a27de8f38f83af.pdf (abgerufen am 04.10.2025).

**SAP SE** (2025d): SAP AI Services List (enGLOBAL v.8-2025). Cloud Service Specifications. URL: https://www.sap.com/germany/about/agreements/policies/cloud-service-specifications.html?sort=latest_desc&pdf-asset=7c6dcf43-1f7f-0010-bca6-c68f7e60039b&page=1 (abgerufen am 29.08.2025).

**SAP SE** (2025e): Integrating Joule with SAP Solutions. Walldorf: SAP SE. URL: https://help.sap.com/doc/de3af3c0f81642dbaa4d36172ed57a72/CLOUD/en-US/79bfc83ab386450c8cd9c7937ce26a3a.pdf (abgerufen am 22.09.2025).

**SAP SE** (URL1): Joule – Unified Setup: Bridging Simplicity and Performance. SAP Community – Technology Blog Post, 12. November 2024. URL: https://community.sap.com/t5/technology-blog-posts-by-sap/joule-unified-setup-bridging-simplicity-and-performance/ba-p/13924542 (abgerufen am 30.08.2025).

**SAP SE** (URL2): Extend Joule with Joule Studio. SAP Architecture Center, zuletzt aktualisiert am 25. Juni 2025. URL: https://architecture.learning.sap.com/docs/ref-arch/06ff6062dc (abgerufen am 22.08.2025).

**SAP SE** (URL3): Harness retrieval-augmented generation in Joule and generative AI hub. 16. Oktober 2024. URL: https://community.sap.com/t5/technology-blog-posts-by-sap/harness-retrieval-augmented-generation-in-joule-and-generative-ai-hub/ba-p/13901774 (abgerufen am 04.10.2025).