

Sachbericht zum Verwendungsnachweis des vom BMFTR geförderten Vorhabens:



DAFNE: Plattform Data Fusion Generator für die Künstliche Intelligenz

FKZ 01IS21059A

Teil I: Kurzbericht

Projektleitung: Prof. Dr. Ulrike Steffens, HITeC e.V.

Laufzeit: 1.12.2021 – 29.02.2024

Autoren: Ulrike Steffens, Pamela Kunert, Tom Krause, Eric Eichholtz



Gefördert durch:



Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter dem Förderkennzeichen **01IS21059A** gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor/den Autoren.

Kurzbericht DaFne

Plattform Data Fusion Generator für die Künstliche Intelligenz

Ergebnisse des Hamburger Informatik Technologie-Centers HITeC e.V.

BMBF-gefördertes Verbundprojekt, Laufzeit: 1.12.2021 – 28.02.2024

Förderkennzeichen 01IS21059A

Ursprüngliche Aufgabenstellung

Das Projekt DaFne zielte darauf ab, die Nutzbarkeit von Methoden zur Datengenerierung zu verbessern, insbesondere in Bereichen, in denen Trainingsdaten für KI-Methoden nicht ausreichend oder in passender Form verfügbar sind. Die praktische Nutzbarkeit bisher existierender Methoden ist oft eingeschränkt, da sie auf spezifische Einsatzgebiete zugeschnitten sind und ihre Anpassung an eigene Einsatzkontexte aufwändig ist. Durch die Entwicklung flexibler Methoden und überprüfbarer Qualitätskriterien sollte die Anwendbarkeit von Datengenerierungsmethoden verbessert und erweitert werden. Dazu sollte eine flexible Plattform konzipiert werden, auf welcher die entwickelten Methoden und Metriken nutzbar gemacht werden. Außerdem war eine Evaluation der entwickelten Methoden, Metriken sowie der Plattform selbst in der Anwendungsdomäne Smart Cities geplant.

Vor Beginn des Projekts existierten bereits einige Ansätze zur Datengenerierung, die jedoch meist für spezifische Kontexte optimiert waren. Die Verwendung dieser Methoden erforderte erhebliche Expertise, und ihre Anwendung war auf spezifische Einsatzgebiete beschränkt. Das Projekt DaFne sollte diese Lücke schließen und eine generisch einsetzbare Plattform zur Datengenerierung schaffen, die sich flexibel an verschiedene Anwendungsszenarien anpassen und erweitern lässt.

Ablauf des Vorhabens

Als Konsortialführer übernahmen wir das Projektmanagement für das gesamte DaFne-Projekt. Das Vorhaben gliederte sich in mehrere Arbeitspakete, die im Folgenden kurz vorgestellt werden; detaillierte Beschreibungen sind im Abschlussbericht zu finden.

- **Arbeitspaket 1:** Grundlagen und Methoden zur synthetischen Datengenerierung – Anforderungsanalyse und Entwicklung bzw. Verfeinerung flexibler Methoden zur Datengenerierung.
- **Arbeitspaket 2:** Methoden zur Bewertung von Generierungsmethoden – Anforderungsanalyse, Auswahl, Umsetzung und Entwicklung von Verfahren zur Evaluierung der Qualität generierter Daten.
- **Arbeitspaket 3:** Plattformentwicklung – Anforderungsanalyse, Entwurf der Architektur und der GUI, sowie Implementierung eines Prototyps.
- **Arbeitspaket 4:** Evaluation der Ergebnisse im Anwendungsfeld Smart Cities – Identifikation von Use Cases, Modellierung und Evaluierung der Use Cases zur Erprobung der Methoden in realen Anwendungsszenarien.

Die Methodik umfasste umfangreiche Literaturrecherchen, Prototyping und iterative Entwicklung sowie Evaluation und Validierung durch statistische und empirische Methoden. Dies gewährleistete eine systematische und effiziente Erreichung der Projektziele.

Wesentliche Ergebnisse

Das Projekt wurde von einem Konsortium mehrerer Partner durchgeführt, darunter die HafenCity Universität Hamburg (HCU), das Forschungszentrum L3S der Leibniz Universität Hannover, das Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden, und die Industriepartner Sopra Steria und Ströer. Diese Zusammenarbeit ermöglichte eine Integration von wissenschaftlicher Expertise und praktischer Anwendung. Die folgenden Ergebnisse fassen die wesentlichen Erkenntnisse zusammen, die wir als HITEC/HAW im Rahmen des DaFne-Projekts gewonnen haben; detaillierte Beschreibungen sind im Abschlussbericht zu finden.

AP1 – Methoden zur Datengenerierung

- Es wurden verschiedene generische Algorithmen zur Erzeugung tabellarischer Daten entwickelt und getestet, darunter die regelbasierte Erzeugung, die Datenfusion, sowie die Reproduktion mithilfe von generativen Machine Learning Modellen.
- Für die Reproduktion von tabellarischen Daten wurden Modelle für die Synthese von Zeitreihen- und Mobilitätsdaten, sowie die datenschutzkonforme Synthese erprobt und (weiter-)entwickelt.

AP2 – Qualitätsbewertung

- Es wurden umfassende Evaluationsmethoden und Qualitätsmetriken für die unterschiedlichen synthetischen Daten (tabellarische und datenschutz-konforme Daten, Zeitreihen- und Mobilitätsdaten), basierend auf dem Vergleich der echten mit den synthetischen Daten, erprobt und entwickelt.
- Ein Evaluations-Service zur Bewertung der generierten Daten hinsichtlich ihrer Qualität und Ähnlichkeit zu realen Daten wurde implementiert.

AP3 – Plattformentwicklung

- Es wurde die Architektur für eine flexible und skalierbare Plattform konzipiert, welche generische und anwendungsfallspezifische Methoden zur Datengenerierung und deren Qualitätsanalyse mittels Microservices nutzbar macht und sich gut um neue Methoden erweitern lässt.
- Es wurde ein Prototyp der Plattform implementiert und deployt, welcher ein flexibles Frontend (bestehend aus Micro-Frontends) mit dem Backend verbindet und in einem Kubernetes-Cluster läuft. Die Konfiguration über ein Helm-Chart ermöglicht eine einfache Migration und erneute Bereitstellung der Plattform auf anderer Infrastruktur.

AP4 – Anwendungskontext Smart Cities

- Die generischen Methoden zur Datengenerierung wurden unter anderem erfolgreich anhand eines Predictive Maintenance Use Case für Brücken erprobt. Die Datenfusion mit öffentlich verfügbaren Wetterdaten hat bspw. gezeigt, dass die Wetterdaten anders aufbereitet werden müssen (Extremwetter), um einen höheren Informationsgehalt für den Use Case zu bieten.

Sachbericht zum Verwendungsnachweis des vom
BMFTR geförderten Vorhabens:



DaFne: Plattform Data Fusion Generator für die Künstliche Intelligenz

Projektleitung: Prof. Dr. Ulrike Steffens

Laufzeit des Vorhabens: 1.12.2021 – 29.02.2024

Autoren: Ulrike Steffens, Pamela Kunert, Tom Krause, Eric Eichholtz

Teil II: Eingehende Darstellung



HITeC e.V.
Vogt-Kölln-Str. 30
22527 Hamburg

Gefördert durch:



Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter dem Förderkennzeichen 01IS21059A gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor.

Inhaltsverzeichnis

1	Überblick	4
2	Problemdarstellung	5
3	Projektziele und Forschungsfragen	6
4	Vorgehensweise	7
4.1	Arbeitspakete	7
	Arbeitspaket 1: Grundlagen und Methoden zur synthetischen Datengenerierung	7
	Arbeitspaket 2: Methoden zur Bewertung von Generierungsmethoden	7
	Arbeitspaket 3: Plattformentwicklung	8
	Arbeitspaket 4: Evaluation der Ergebnisse im Anwendungsfeld Smart Cities	8
4.2	Methodik	8
5	Ergebnisse AP1: Grundlagen und Methoden zur synthetischen Datengenerierung	9
5.1	Regelbasiert	10
5.2	Fusion	11
5.3	Reproduktion	11
	Reproduktion von tabellarischen Daten	12
	Datenschutzkonforme Reproduktion	13
	Reproduktion von Zeitreihen-Daten	14
	Reproduktion von Mobilitätsdaten	15
6	Ergebnisse AP2: Methoden zur Bewertung von Generierungsmethoden	16
6.1	Tabellarische Daten	17
6.2	Verfahren zur Überprüfung von Privatsphäre	17
6.3	Zeitreihen-Daten	18
6.4	Mobilitätsdaten	18
7	Ergebnisse AP3: Plattformentwicklung	19
7.1	Anforderungsanalyse Plattform	19
7.2	Architekturentwurf	19
7.3	Implementierung	22
	Frontend-Implementierung	22
	Plattform-Implementierung	23
8	Ergebnisse AP4: Evaluation der Ergebnisse im Anwendungsfeld Smart Cities	24
9	Fazit	25
9.1	Forschungsfrage 1: Welche Methoden bzw. Modelle eignen sich zur flexiblen Synthese von tabellarischen Daten und lassen sich damit auf Anwendungsfälle im Bereich Smart Cities und darüber hinaus anwenden?	25

9.2	Forschungsfrage 2: Welche Qualitätskriterien sind notwendig, um die Brauchbarkeit und Integrität synthetischer Daten sicherzustellen?	27
9.3	Forschungsfrage 3: Wie kann eine Plattform zur Datengenerierung konzipiert werden, die sowohl benutzerfreundlich als auch erweiterbar und skalierbar ist?	28
9.4	Forschungsfrage 4: Können synthetische Daten zur Verbesserung spezifischer Anwendungsfälle aus dem Bereich Smart Cities beitragen?	29
10	Projektkosten: Zahlenmäßiger Nachweis	29
11	Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten ...	30
12	Ausblick: voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse	31
13	Fortschritt auf dem Gebiet bei anderen Stellen	32
14	Veröffentlichungen	33

1 Überblick

In diesem Abschlussbericht präsentiert das Hamburger Informatik Technologie-Center (HITeC) e.V. seine Ergebnisse des DaFne-Projekts (Plattform Data Fusion Generator für die Künstliche Intelligenz). Das DaFne-Projekt war ein Verbundvorhaben, das in enger Zusammenarbeit mit verschiedenen Forschungs- und Industriepartnern durchgeführt wurde, darunter die HafenCity Universität Hamburg (HCU), das L3S Forschungszentrum der Leibniz Universität Hannover, Sopra Steria, Ströer und das Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) der TU Dresden. Dieser Bericht wurde von HITeC e.V. (und der Hochschule für Angewandte Wissenschaften (HAW) Hamburg) erstellt und stellt die spezifischen Beiträge und Erkenntnisse dieser beiden Institutionen dar.

Das Projekt wurde unter dem Förderkennzeichen 01S21059A vom Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) gefördert und vom DLR Projektträger betreut. Die Laufzeit des Projekts war ursprünglich vom 1. Dezember 2021 bis zum 30. November 2023 geplant und wurde kostenneutral bis zum 29. Februar 2024 verlängert. Die kostenneutrale Verlängerung begründet sich einerseits durch einen leicht verspäteten Projektstart aufgrund der späten Projektbewilligung. Andererseits kam es Ende 2022 zu einem Cyberangriff an der HAW Hamburg, dessen Auswirkungen das Projekt ausbremsten.

Inhaltlich befasste sich das DaFne-Projekt mit der Entwicklung und Evaluierung von Methoden zur Synthese und Fusion von Daten, insbesondere im Bereich Smart Cities. Ziel war es, den Prototyp einer flexiblen und skalierbaren Plattform zu schaffen, die diese Methoden für verschiedene Anwendungsfälle nutzbar macht. Eine ausführlichere Problembeschreibung findet sich in Kapitel 2.

Das Forschungsvorhaben von HITeC wurde von vier zentralen Forschungsfragen geleitet, die in Kapitel 3 zusammen mit den Zielen des Projekts ausführlich beschrieben werden:

1. Welche Methoden bzw. Modelle eignen sich zur flexiblen Synthese von tabellarischen Daten und lassen sich damit auf Anwendungsfälle im Bereich Smart Cities und darüber hinaus anwenden?
2. Welche Qualitätskriterien sind notwendig, um die Brauchbarkeit und Integrität synthetischer Daten sicherzustellen?
3. Wie kann eine Plattform zur Datengenerierung konzipiert werden, die sowohl benutzerfreundlich als auch erweiterbar und skalierbar ist?
4. Können synthetische Daten zur Verbesserung spezifischer Anwendungsfälle aus dem Bereich Smart Cities beitragen?

Die Vorgehensweise und Methodik innerhalb des Projekts werden in Kapitel 4 detailliert erläutert. In den Kapiteln 5 bis 8 werden die Ergebnisse der einzelnen Arbeitspakete umfassend dargestellt, wobei sowohl die entwickelten Methoden als auch die erzielten Erkenntnisse dokumentiert sind. Im Fazit in Kapitel 9 werden die zentralen Forschungsfragen erneut aufgegriffen und beantwortet, um die wesentlichen Erkenntnisse des Projekts zusammenzufassen.

Ein Überblick über die Kosten des Projekts findet sich in Kapitel 10. Die Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten werden in Kapi-

tel 11 dargelegt. Anschließend bietet Kapitel 12 einen Ausblick auf die zukünftige Nutzung und Weiterentwicklung der im Zuge des DaFne-Projekts entwickelten Technologien und Methoden, einschließlich konkreter Planungen für die Verwertung der Projektergebnisse. Kapitel 13 beleuchtet darüber hinaus den während der Projektlaufzeit an anderen Stellen bekannt gewordenen Fortschritt auf dem Gebiet des Vorhabens und zeigt auf, inwiefern diese Entwicklungen in die Arbeiten im DaFne-Projekt eingeflossen sind. Der Bericht schließt mit einer Liste der entstandenen Veröffentlichungen (Kapitel 14) ab.

2 Problemdarstellung

In der heutigen datengesteuerten Welt ist der Zugang zu qualitativ hochwertigen Daten entscheidend für den Fortschritt in vielen Bereichen, einschließlich medizinischer Forschung, Finanzwesen und städtischer Planung. Oft sind diese Daten jedoch schwer zugänglich, entweder aufgrund von Datenschutzbestimmungen wie der europäischen Datenschutz-Grundverordnung (DSGVO) oder aufgrund technischer und praktischer Hindernisse. Dies stellt eine große Herausforderung dar, insbesondere für die Entwicklung und das Training von Künstlicher Intelligenz (KI) und maschinellen Lernmodellen, die auf umfangreiche und vielfältige Datensätze angewiesen sind. Ein besonderes Problem besteht in der Bereitstellung von Daten, die sowohl repräsentativ als auch datenschutzkonform sind. In Bereichen wie der vorausschauenden Wartung von Infrastrukturen (z.B. Brücken) und der Planung von Smart Cities fehlt es oft an ausreichend großen und vielfältigen Datensätzen, um robuste und genaue Modelle zu entwickeln. Dies führt zu suboptimalen Ergebnissen und begrenzt die Einsatzmöglichkeiten moderner Technologien.

Die Synthese von Daten, also die Erzeugung eines Datensatzes, welcher einem Ausgangsdatsatz in Merkmalen und (statistischen) Eigenschaften ähnelt, ohne Datenpunkte aus diesem zu kopieren, kann hier Abhilfe schaffen: Mithilfe der synthetischen Daten können Datenschutz-Bedenken ausgeräumt werden, da die generierten Daten keine direkten Rückschlüsse auf individuelle Datenpunkte zulassen. Zudem ermöglichen synthetische Datensätze eine flexible und skalierbare Bereitstellung von Daten, die für spezifische Anwendungsfälle maßgeschneidert werden können. Auch eine Erweiterung des Informationsgehalts von Daten, durch die Fusion mit weiteren öffentlich verfügbaren Daten wie bspw. Wetterdaten, stellt hier eine nützliche Anwendung dar. Dies ist besonders wichtig in Bereichen, in denen hohe Anforderungen an die Datenqualität und -vielfalt gestellt werden, um fundierte und präzise Analysen und Vorhersagen zu ermöglichen.

Jedoch ist die Generierung synthetischer Daten, die realen Daten sehr ähnlich sein sollen, aber keine sensiblen Informationen preisgeben, eine komplexe und technisch anspruchsvolle Aufgabe. Bestehende Methoden zur Datengenerierung, wie generative Adversarial Networks (GANs), stoßen oft an ihre Grenzen, wenn es darum geht, verschiedene Datentypen zu integrieren oder komplexe, nicht-lineare Beziehungen in den Daten abzubilden.

Das DaFne-Projekt (Data Fusion Generator) adressierte diese Herausforderungen, indem der Prototyp einer flexiblen und erweiterbaren Plattform zur Generierung synthetischer Daten entwickelt wurde. Ziel war es, bestehende Methoden zur Datengenerierung zu evaluieren und basierend auf den Erkenntnissen zu verbessern und so anzupassen, dass sie generisch nutzbar sind. Desweiteren sollten verifizierbare Qualitätskriterien etabliert werden, und die Nutzbarkeit von Modellen und Metriken vereinfacht werden. Besonderer Fokus bei HITeC lag dabei auf der Generierung synthetischer tabellarischer Daten, welche für Anwendungsfälle aus dem Bereich Smart Cities, wie beispielsweise Predictive Maintenance für Brücken, genutzt werden können.

3 Projektziele und Forschungsfragen

Das DaFne-Projekt verfolgte das übergeordnete Ziel, die Nutzbarkeit von Methoden zur Datengenerierung für KI-Forscher und -Entwickler zu verbessern. Dies sollte durch die Entwicklung eines flexiblen und erweiterbaren Prototypen einer Plattform zur Datengenerierung erreicht werden. Die spezifischen wissenschaftlichen und technischen Ziele des Projekts umfassten unter anderem:

1. Entwicklung flexibler Methoden zur Datengenerierung:
 - Schaffung eines Prototypen zur regelbasierten Erzeugung tabellarischer Daten.
 - Entwicklung von Methoden zur Datenfusion, insbesondere die Integration von öffentlich verfügbaren Daten wie Wetterdaten.
 - Untersuchung und Weiterentwicklung generativer Modelle zur synthetischen Datensynthese.
2. Etablierung von Qualitätskriterien für synthetische Daten:
 - Auswertung und Implementierung von Evaluationsmethoden zur Bewertung der Qualität synthetischer Daten.
 - Analyse von geeigneten Metriken zur Bewertung der unterschiedlichen tabellarischen Daten wie beispielsweise Zeitreihendaten.
3. Plattformentwicklung und -integration:
 - Entwurf und Implementierung einer Microservice-Architektur für die DaFne-Plattform.
 - Entwicklung einer benutzerfreundlichen und erweiterbaren GUI für die Plattform.
 - Bereitstellung der Plattform für lokale Nutzung und/ oder in einer Cloud-Infrastruktur.
4. Evaluation Plattform im Anwendungskontext Smart Cities:
 - Test und Validierung der tabellarischen Synthetisierungsfunktionalitäten anhand eines geeigneten Use Case.
 - Analyse und Generierung synthetischer Mobilitätsdaten für Smart Cities.

Um diese Ziele zu erreichen, haben wir im Rahmen des Projekts mehrere zentrale Forschungsfragen formuliert:

1. Welche Methoden bzw. Modelle eignen sich zur flexiblen Synthese von tabellarischen Daten und lassen sich damit auf Anwendungsfälle im Bereich Smart Cities und darüber hinaus anwenden?
2. Welche Qualitätskriterien sind notwendig, um die Brauchbarkeit und Integrität synthetischer Daten sicherzustellen?
3. Wie kann eine Plattform zur Datengenerierung konzipiert werden, die sowohl benutzerfreundlich als auch erweiterbar und skalierbar ist?
4. Können synthetische Daten zur Verbesserung spezifischer Anwendungsfälle aus dem Bereich Smart Cities beitragen?

Diese Forschungsfragen haben unsere wissenschaftlichen Untersuchungen und Entwicklungsarbeiten innerhalb des DaFne-Projekts geleitet und trugen dazu bei, die angestrebten Ziele zu erreichen.

4 Vorgehensweise

4.1 Arbeitspakete

Das DaFne-Projekt wurde in mehrere Arbeitspakete (AP) unterteilt, die jeweils spezifische Aufgaben und Ziele verfolgten. Die Arbeitspakete sind eng miteinander verzahnt und wurden durch regelmäßige Abstimmungstreffen koordiniert.

Arbeitspaket 1: Grundlagen und Methoden zur synthetischen Datengenerierung In diesem Arbeitspaket wurden die methodischen Grundlagen für die Datengenerierung gelegt.

- **Anforderungsanalyse Methoden:** Identifikation und Analyse von Generierungsmöglichkeiten.
- **Literaturübersicht:** Recherche zum aktuellen Stand der Technik im Bereich der Datensynthese.
- **Methodenentwicklung:** Untersuchung und Anpassung geeigneter bestehender generativer Modelle und damit verbunden die Erstellung von ersten Prototypen zur Datensynthese.
- **Verfeinerung der Modelle:** Weiterentwicklung der Modelle, um den Anforderungen und Erkenntnissen aus den Evaluierungen gerecht zu werden.

Arbeitspaket 2: Methoden zur Bewertung von Generierungsmethoden Dieses Arbeitspaket konzentrierte sich auf die Entwicklung von Bewertungsmethoden für die generierten Daten.

- **Analyse der Anforderungen an Qualitätskriterien und deren Bewertung:** Analyse und Benennung von konkreten Anforderungen an die Qualitätsbewertung von synthetischen Daten, mit Berücksichtigung unterschiedlicher relevanter Bereiche (z.B. Privatsphäre)

- **Auswahl der Methoden zur Qualitätsbewertung:** Untersuchung von geeigneten Methoden für eine ganzheitliche Qualitätsbetrachtung der synthetisierten Daten unter Berücksichtigung unterschiedlicher Aspekte, wie beispielsweise die Motivation zur Datengenerierung (Privatsphäreschutz vs. Datensatz-Erweiterung usw.).
- **Umsetzung der Methoden:** Implementierung, Evaluierung und ggfs. Erweiterung von geeigneten Metriken zur Bewertung der Datenqualität.
- **Darstellung der Qualitätsbewertung:** Entwicklung einer geeigneten Darstellung der Ergebnisse für den Nutzer, sodass dieser einen Überblick über die synthetisierten Daten und deren Qualität erhält.

Arbeitspaket 3: Plattformentwicklung Die Plattformentwicklung war ein zentrales Element des Projekts und umfasste mehrere Schritte.

- **Anforderungsanalyse Plattform:** Erhebung der Anforderungen an die DaFne-Plattform, sowohl aus Nutzer- als auch aus algorithmischer Sicht.
- **Architekturentwurf:** Entwurf einer, basierend auf den zuvor identifizierten Anforderungen, geeigneten Architektur für die DaFne-Plattform.
- **GUI-Entwicklung:** Entwurf und Implementierung einer benutzerfreundlichen grafischen Benutzeroberfläche (GUI) für die Plattform.
- **Implementierung:** Implementierung der zuvor entworfenen DaFne-Plattform, sowohl im Bezug auf das Backend als auch auf das Frontend.

Arbeitspaket 4: Evaluation der Ergebnisse im Anwendungsfeld Smart Cities In diesem Arbeitspaket wurden die entwickelten Methoden und die Plattform in realen Anwendungsfällen getestet und evaluiert.

- **Definition von Smart City Use Cases, Zielqualitäten und Kriterien:** Definition von Smart City Use Cases im Bereich Mobilität und Infrastruktur inklusive der Benennung von Nutzungszielen für synthetische Daten.
- **Entwicklung des Evaluationsverfahrens:** Um die Validität der im Zuge des Projekts entwickelten synthetischen Datengenerierungs-Tools überprüfen zu können, wird ein Verfahren zur Evaluierung entwickelt.
- **Modellierung (Implementierung) der Use Cases mit synthetischen Daten:** Durchführung des Use Cases zur Überprüfung der Tools sowie der synthetischen Daten selbst.
- **Validierung der synthetisierten Daten:** Simulierung/ Durchführung des Use Cases mit anschließender qualitativer und quantitativer Bewertung der Ergebnisse.
- **Quantifizierung der Ergebnisse und Rückführung von Erkenntnissen:** Analyse und kritische Betrachtung der Ergebnisse aus dem vorherigen Schritt, um hieraus ggfs. Maßnahmen bzw. Verbesserungen abzuleiten.

4.2 Methodik

Die Vorgehensweise innerhalb der Arbeitspakete basierte auf etablierten wissenschaftlichen und technischen Methoden:

- **Literaturrecherche und Anforderungsanalyse:** Umfassende Analyse des aktuellen Stands der Technik und der Anforderungen der verschiedenen Anwendungsfälle.
- **Prototyping und iterative Entwicklung:** Entwicklung und kontinuierliche Verbesserung von Prototypen basierend auf Feedback und Evaluationsergebnissen.
- **Evaluation und Validierung:** Einsatz von statistischen und empirischen Methoden zur Bewertung der Qualität und Nützlichkeit der generierten Daten sowie der Funktionalität der Plattform.
- **Laufende Forschung und Aktualisierung:** Ständige Überprüfung und Einbeziehung neuer wissenschaftlicher Erkenntnisse und technischer Entwicklungen, um sicherzustellen, dass die Projektergebnisse auf dem neuesten Stand der Forschung basieren.

Diese strukturierte Vorgehensweise gewährleistete, dass die Projektziele systematisch und effizient erreicht wurden und die entwickelten Lösungen den Anforderungen der Anwendungsfälle entsprachen.

5 Ergebnisse AP1: Grundlagen und Methoden zur synthetischen Datengenerierung

In Zusammenarbeit mit den Partnern ist eine Anforderungsanalyse für die Ausgestaltung der Datensynthese erfolgt. Dabei wurden unterschiedliche Generierungsmöglichkeiten als nützlich identifiziert, welche sich zunächst in die generische und die Anwendungsfall-spezifische Synthese unterscheiden lassen, wie in Abbildung 1 dargestellt.

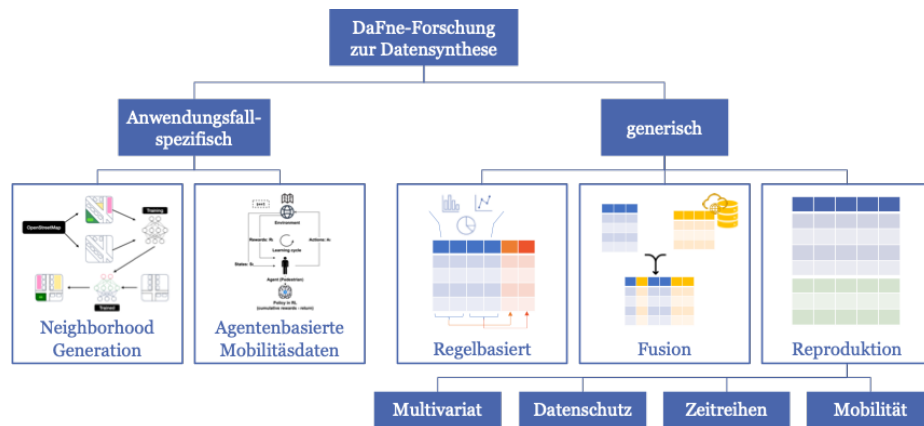


Abb. 1. Identifizierte Arten der Datensynthese im DaFne Projekt

Für die Anwendungsfall-spezifische Generierung von tabellarischen Daten wurden zwei Use Cases aus dem Smart City Bereich identifiziert: Die Neighbor-

hood Generation zielt darauf ab, synthetische Nachbarschaftsdaten für Smart Cities zu generieren. Hierbei werden graphentheoretische Methoden angewendet, um Beziehungen und Interaktionen zwischen verschiedenen Entitäten in einer Stadt zu modellieren. Der Ansatz wird von unserem Forschungspartner L3S bearbeitet und ist daher hier nicht weiter ausformuliert. Bei dem zweiten Ansatz handelt es sich um die agentenbasierte Erzeugung von Mobilitätsdaten, welche von unseren Forschungspartnern HCU und TU Dresden bearbeitet wird. Dieser Use Case fokussiert sich auf die Simulation von Mobilitätsdaten in städtischen Umgebungen. Durch den Einsatz agentenbasierter Modelle sollen Bewegungsmuster von Individuen in einer Stadt realistisch nachgebildet werden.

Zu den generischen Synthese-Algorithmen, welche im Rahmen der Anforderungsanalyse als relevant identifiziert wurden, gehören die regelbasierte Erzeugung, die Fusion, sowie die Reproduktion von tabellarischen Daten. Die einzelnen Ansätze werden im Folgenden erläutert, sowie die im Zuge des Projekts gewonnenen Erkenntnisse und Entwicklungsstände zu den Generierungsarten näher beschrieben.

5.1 Regelbasiert

Der erste der generischen Ansätze zur Erzeugung tabellarischer Daten ist der regelbasierte Ansatz, bei welcher ein tabellarischer Datensatz Spalte für Spalte erzeugt werden kann, ohne auf einen bestehenden Datensatz zurückzugreifen. Die Spalten können dabei unterschiedliche Ausprägungen haben (Zahlen, Orte, Namen usw.) und unterschiedlichen Verteilungen unterliegen. Außerdem soll es auch möglich sein, Spalten zu erzeugen, welche auf anderen Spalten bzw. Spaltenkombinationen beruhen, sodass sich innerhalb des synthetischen Datensatzes auch Beziehungen zwischen Spalten (wie Korrelationen) abbilden lassen.

Im Zuge des Projekts wurde ein Prototyp zur regelbasierten Datengenerierung entwickelt. Dieser fokussierte sich auf die Erzeugung von realitätsnahen Datensätzen im Kontext der Brückeninstandhaltung. Der Prototyp ermöglicht es, verschiedene Datentypen und Verteilungen zu definieren sowie Abhängigkeiten zwischen den Spalten abzubilden. Die Benutzeroberfläche des Prototyps wurde so gestaltet, dass Nutzer ohne tiefgehende Programmierkenntnisse in der Lage sind, realistische synthetische Datensätze zu erzeugen. Verschiedene Typen von Daten (z.B. numerische Werte, Text, Daten) können einfach durch Auswahl und Eingabe von Parametern generiert werden. Außerdem unterstützt der Prototyp die Definition komplexer Verteilungen und Abhängigkeiten zwischen den Spalten, sodass beispielsweise Prüfberichte für Brücken generiert werden könnten, bei denen die Zustandsnote von der Brückenfläche und dem Baujahr abhängt. Zukünftige Arbeiten könnten zusätzliche Regeln und Verteilungsmethoden hinzufügen, um die Generierung noch realistischer und anpassbarer zu machen. Außerdem wären Datenmodelle hilfreich, wie Namen, Adressen oder Telefonnummern, um die realistische Generierung von Daten im Smart City Kontext noch einfacher und schneller zu gestalten.

5.2 Fusion

Eine weitere generische Möglichkeit, tabellarische Daten zu erzeugen, ist die Fusion: hierbei sollen zwei (oder mehr) tabellarische Datensätze miteinander verschnitten werden. Dabei kann es sich bei dem erweiternden Datensatz potenziell um einen öffentlich verfügbaren Datensatz handeln, wie beispielsweise Wetter- oder Verkehrsdaten, sodass sich für den Nutzer ein Datensatz mit höherem Informationsgehalt ergibt. Im Rahmen des Projekts wurde ein generischer Service zur Datenfusion konzipiert und implementiert. Ziel dieses Services ist es, Daten aus verschiedenen Quellen zu integrieren und zu fusionieren, damit Nutzern basierend auf diesem neuen, erweiterten Datensatz neue Erkenntnisse gewinnen können und so die Entscheidungsfindung unterstützt wird. Der entwickelte Prototyp ermöglicht die Verarbeitung und Fusion von Daten aus mehreren Quellen, einschließlich CSV-Dateien und externen APIs, wobei speziell die Wetter-API des Deutschen Wetterservices [5] an den Service angebunden wurde, um die Anreicherung von Ort- und Zeit-basierten Datensätzen mit Wetterdaten wie Temperatur, Luftdruck u.v.m. möglich zu machen. Der Service wurde als Microservice konzipiert, um die Integration in die DaFne-Plattform so einfach wie möglich zu machen. Er besteht aus mehreren Komponenten, darunter eine API-Komponente, eine Fusionskomponente und eine Wetterkomponente. Der Service unterstützt verschiedene Join-Operationen, darunter Inner Join, Left Join, Right Join und Full Outer Join. Dadurch können Nutzer flexibel wählen, wie die Daten fusioniert werden sollen. Die Architektur des Services ermöglicht es, neue Datenquellen und Fusionsstrategien einfach hinzuzufügen. Dies gewährleistet eine hohe Flexibilität und Anpassungsfähigkeit an zukünftige Anforderungen. Durch die Entwicklung und Evaluierung des generischen Fusionservices konnte gezeigt werden, dass die Fusion von Daten aus verschiedenen Quellen nicht nur technisch machbar, sondern auch praktisch umsetzbar ist. Dies eröffnet zahlreiche Anwendungsmöglichkeiten im Smart City Kontext und darüber hinaus.

Trotzdem sind bei der Evaluation im Zuge des Predictive Maintenance Use Case für Brücken (siehe Ergebnisse von AP4 in Kapitel 8) auch Defizite deutlich geworden, wie etwa die Aggregation der Wetterdaten. Bisher war hier nur die Bereitstellung von Durchschnittswerten für die betrachteten Perioden geplant und umgesetzt worden, dabei sind für spezifische Analysen über mehrere Jahre Extremwert-Informationen von höherem Wert, sodass eine Erweiterung diesbezüglich erfolgen müsste [2].

5.3 Reproduktion

Die dritte und letzte von uns identifizierte Möglichkeit der generischen Datensynthese ist die Reproduktion. Hierbei wird, basierend auf einem bereits bestehenden tabellarischen Datensatz, ein neuer synthetischer Datensatz erzeugt, welcher mit dem Ausgangs-Datensatz in seinen Kern-Eigenschaften (wie den Merkmalsausprägungen, statistischen Eigenschaften der Spalten und ihren Zusammenhängen) übereinstimmt, ohne dass er einzelne Zeilen aus diesem kopieren würde. Hierfür wird mithilfe eines generativen Machine Learning (ML) Modells

die Datenverteilung des Ausgangsdatensatzes erlernt, sodass das Modell in der Lage ist, neue Datenpunkte aus dieser Verteilung zu sampeln.

Diese synthetischen Daten können dann entweder als eine Erweiterung des Ausgangs-Datensatzes verwendet werden, beispielsweise um mehr Trainingsdaten bzw. einen Testdatensatz für einen nachgelagerten KI-/ ML-Use Case zu liefern, oder sie können statt der Original-Daten verwendet werden, beispielsweise wenn Datenschutz-Bedenken bei der Weiterverarbeitung der Original-Daten bestehen. Die Motivation zur Datensynthese bestimmt dabei maßgeblich die Auswahl des geeigneten ML-Modell zur Generierung der Daten, auch im Bezug auf die spezifischen Eigenschaften des Datensatzes, der reproduziert werden soll. Zusätzliche Dimensionen wie Zeit und/ oder Orte in den tabellarischen Daten erhöhen dabei die Komplexität und machen besondere Algorithmen notwendig. Im weiteren Verlauf der Forschungsarbeiten wurden deshalb vier unterschiedliche Teilgebiete/ Anwendungsfälle der Reproduktion betrachtet: Die Reproduktion von „einfachen“ tabellarischen Daten (als Erweiterung des Ausgangs-Datensatzes), die datenschutz-konforme Reproduktion von schützenswerten Daten, sowie die Reproduktion von tabellarischen Daten, welche Zeitreihen oder sogar Mobilitätsdaten (also Zeit- & Ort-Dimension) enthalten.

Reproduktion von tabellarischen Daten Im Rahmen des Projekts wurden verschiedene generative Modelle zur tabellarischen Datensynthese erprobt, darunter CTGAN und TVAE [38], welche beide in der SDV Bibliothek [20] implementiert sind. Diese Modelle wurden hinsichtlich ihrer Fähigkeit, tabellarische Daten zu reproduzieren, umfassend getestet und schließlich so weiterentwickelt, dass sie auf der Plattform nutzbar sind. Dabei wurden Experimente mit unterschiedlichen Modell-Architekturen und verschiedenen tabellarischen Datensätzen durchgeführt. Zur Bewertung der generierten Daten wurden verschiedene Metriken wie bspw. der Kolmogorov-Smirnov-Test [18] verwendet. Diese Metriken halfen dabei, die Qualität und Genauigkeit der synthetischen Daten im Vergleich zu den Originaldaten zu bestimmen, mehr dazu im Kapitel 6 zu AP2. Die Ergebnisse zeigten signifikante Performance-Unterschiede der Modelle und ihrer Architekturen, die jedoch je nach Datensatz variierten, sodass eine individuelle Anpassung der Modellarchitektur für jeden Datensatz lohnenswert erscheint. Dies umfasst nicht nur die Wahl des Modells selbst, sondern auch die Auswahl der optimalen Anzahl von Schichten und Neuronen in den Modellen, sowie die Feinabstimmung der Hyperparameter.

Ein weiteres generatives ML-Modell zur Reproduktion von tabellarischen Daten ist das CTAB-GAN+[40], welches konzipiert wurde um die Limitierungen von CTGAN und TVAE zu adressieren. Dieses wurde im Rahmen dieses Forschungsprojekts ebenso hinsichtlich seiner generischen Anwendbarkeit getestet. Das Modell integriert fortgeschrittene Techniken wie mode-spezifische Normalisierung und die Gumbel-Softmax-Funktion, um die Generierung sowohl kontinuierlicher als auch kategorischer Daten zu verbessern. Die Evaluation des CTAB-GAN+ zeigte, dass das Modell in der Lage ist, die statistischen Eigenschaften und komplexen Beziehungen der Originaldatensätze sehr gut zu reproduzieren.

Insbesondere bei Datensätzen mit gemischten Datentypen (kontinuierlich und kategorisch) erzielte CTAB-GAN+ gute Ergebnisse. Trotzdem zeigt die öffentlich verfügbare Implementierung des Modells eine hohe Fehleranfälligkeit für unterschiedliche Datensätze und -typen, sodass hier umfangreiche Verbesserungen stattfinden müssen, damit dieses Modell generisch auf der Plattform einsetzbar wäre.

Datenschutzkonforme Reproduktion Ein weiterer wichtiger Aspekt der generischen Datensynthese ist die datenschutzkonforme Reproduktion von Daten, die beispielsweise auf Grund von Datenschutzbedenken schützenswert sind. Hierbei wird sichergestellt, dass die generierten Daten keine sensiblen Informationen aus den Originaldatensätzen preisgeben. Zu diesem Zweck wurde die Integration von Differential Privacy (DP) in generative Modelle untersucht. Differential Privacy fügt dem Lernprozess kontrolliertes Rauschen hinzu, um die Identifizierbarkeit einzelner Datenpunkte zu verhindern und somit den Datenschutz zu gewährleisten [6]. Im Rahmen des Projekts wurden verschiedene Modelle zur Integration von Differential Privacy in generative Adversarial Networks (GANs) untersucht. Dazu gehören unter anderem DPGAN [37], PATE-GAN [19] sowie die fortgeschrittenen Modelle DP-CGANS [24] und CTAB-GAN+ [40]. Ziel war es, ein Modell zu identifizieren, das sowohl die statistischen Eigenschaften der Originaldaten gut reproduziert als auch einen hohen Grad an Datenschutz gewährleistet. Die Untersuchung konzentrierte sich auf die Anpassung und Optimierung der Modellparameter, um eine Balance zwischen Datenqualität und Datenschutz zu erreichen. Besondere Aufmerksamkeit wurde dem CTAB-GAN+ gewidmet, welches bereits in der vorherigen Evaluationsphase durch seine Fähigkeit zur Reproduktion komplexer Datenstrukturen überzeugte. Die Integration von Differential Privacy in CTAB-GAN+ erfolgte durch den Einsatz der Differentially Private Stochastic Gradient Descent (DP-SGD) [1] Methode. Hierbei werden während des Trainings die Gradienten der Modellparameter mit Rauschen versehen, um die Privatsphäre der zugrunde liegenden Daten zu schützen. Die Evaluation der datenschutzkonformen Modelle erfolgte anhand von zwei tabellarischen Datensätzen aus dem Bereich Smart City, die unterschiedliche Komplexitätsgrade und Datentypen umfassten. Es wurde festgestellt, dass das CTAB-GAN+ in der Lage ist, die statistischen Eigenschaften der Originaldaten gut zu reproduzieren, während es gleichzeitig einen hohen Grad an Differential Privacy gewährleistet. Besonders bei Datensätzen mit gemischten Datentypen erzielte das Modell gute Ergebnisse. Neben CTAB-GAN+ wurden auch andere Modelle wie DPGAN und PATE-GAN getestet: Das DPGAN nutzt die Methode des differenziell privaten stochastischen Gradientenabstiegs, bei dem Rauschen direkt auf die Gradienten während des Trainings angewendet wird. Dies ermöglicht es, Datenschutz zu gewährleisten, indem verhindert wird, dass das Modell spezifische Datenpunkte lernt. PATE-GAN hingegen nutzt ein Ensemble von Lehrermodellen, die jeweils auf unterschiedlichen Teilen der Daten trainiert werden, um Vorhersagen für neue Datenpunkte zu machen. Diese Vorhersagen werden aggregiert und mit Rauschen versehen, um die Privatsphäre zu schützen,

bevor sie zur Trainingsdatenerstellung für ein Schülermodell verwendet werden. Die Ergebnisse zeigten, dass sowohl DPGAN als auch PATE-GAN effektive Methoden zur datenschutz-konformen Datengenerierung bieten. Allerdings zeigte das CTAB-GAN+ insgesamt die beste Performance in Bezug auf die Reproduktion der Datenqualität und den Schutz der Privatsphäre. Zusammenfassend lässt sich sagen, dass die Integration von Differential Privacy in generative Modelle ein effektiverer Ansatz ist, um synthetische Daten zu erzeugen, die sowohl qualitativ hochwertig als auch datenschutzkonform sind.

Reproduktion von Zeitreihen-Daten Im Rahmen der Evaluierung des Reproduction Service auf der DaFne-Plattform anhand eines Predictive Maintenance Use Cases für Brücken (siehe Kapitel 8 AP4: Evaluierung) wurde festgestellt, dass die bisher verwendeten Modelle, wie CTGAN und TVAE, für die Generierung von Zeitreihen-Daten weniger gut geeignet sind. Diese Modelle sind primär darauf ausgelegt, die Datenverteilungen der einzelnen Spalten und die Beziehungen zwischen den Spalten zu erfassen, berücksichtigen jedoch nicht die temporalen Abhängigkeiten, die in Zeitreihen-Daten eine wesentliche Rolle spielen. Um diese Herausforderung zu adressieren, wurden im Projekt spezielle generative Modelle für Zeitreihen-Daten erprobt und weiterentwickelt. Zu den getesteten Modellen gehören die Generative Adversarial Networks (GAN)-basierten Modelle TimeGAN [39] und TTS-GAN [14], sowie das autoregressive PAR-Modell (Probabilistic Autoregressive Model) [32]. Beim TimeGAN handelt es sich um ein Modell, welches Zeitreihen-Daten mithilfe eines rekurrenten Embedding-Moduls in einen latenten Raum abbildet, um darauf basierend sowohl zeitliche Abhängigkeiten als auch dynamische Muster in den Daten realistisch abbilden zu können. Leider war das Modell im Zuge der Evaluation mit Brückendaten nicht in der Lage, vielfältige Daten zu produzieren, sondern generierte nur sehr ähnliche, teilweise sogar sich überschneidende, gleiche Datenpunkte. Beim TTS-GAN handelt es sich um ein Transformer-basiertes GAN, welches sowohl für den Diskriminator als auch für den Generator eine Transformer Encoder Architektur nutzt, um realistische synthetische Zeitreihen-Daten unterschiedlicher Länge zu erzeugen. Leider war auch das TTS-GAN nicht in der Lage, die Brückendaten realistisch abzubilden, hier waren vor allem Datenpunkte außerhalb der realistischen Grenzen (bspw. negative Jahreswerte) der Hauptgrund für die schlechte Performance. Das PAR-Modell, das auf autoregressiven Prinzipien basiert, benötigte selbst für eine stark reduzierte Version des Brücken-Datensatzes (nur 5% der ursprünglichen Datenmenge) so lange Trainingszeiten, dass auf eine Evaluation verzichtet wurde. Aufgrund der unzureichenden Ergebnisse der bestehenden Modelle wurde ein neues Modell konzipiert und implementiert, welches die Vorteile von Variational AutoEncodern (VAE [13] & [23]) mit Long Short-Term Memory (LSTM [10]) Netzwerken kombiniert. Dieses Modell nutzt die Fähigkeit von VAEs, latente Repräsentationen der Daten zu lernen, und die Stärke von LSTMs, langfristige Abhängigkeiten in Sequenzen zu erfassen. Die Evaluation des kombinierten Modells im Rahmen des Predictive Maintenance Use Cases auf den Brückendaten zeigte vielversprechende Ergeb-

nisse bei der Generierung realistischer Zeitreihen-Daten, die sowohl die statistischen Eigenschaften als auch die temporalen Abhängigkeiten der Originaldaten besser als die zuvor getesteten Modelle TimeGAN, TSS-GAN und PAR widerspiegeln. Trotzdem sind auch hier Möglichkeiten zur Verbesserung des Modells evident geworden, wie beispielsweise die Einhaltung der statistischen Verteilung von kategorialen Merkmalen und deren Ausprägungen, besonders bei unterrepräsentierten Kategorien. Um eine breite und vor allem performante Anwendbarkeit des Modells garantieren zu können, sind daher weitere Verbesserungen, sowie Experimente mit anderen Datensätzen nötig.

Reproduktion von Mobilitätsdaten Im Smart City Bereich spielen Mobilitätsdaten eine zentrale Rolle, da sie wichtige Informationen über die Bewegungsmuster von Personen und Fahrzeugen innerhalb einer städtischen Umgebung liefern. Diese Daten enthalten sowohl zeitliche als auch räumliche Komponenten, was ihre realistische Abbildung und Synthese zu einer komplexen Herausforderung macht. Über unseren Kooperationspartner Ströer und dessen Beteiligung an der Arbeitsgemeinschaft Media-Analyse e.V. (kurz AGMA) erhielten wir einen umfassenden Datensatz zu Mobilitätsdaten in Deutschland, im Folgenden AGMA-Datensatz genannt. Dieser Datensatz enthält über mehrere Untersuchungsjahre unter anderem geografische Daten zu Bewegungsabläufen von ca. 2.500 Probanden, welche 14 Tage lang über ein GPS-Trackinggerät aufgezeichnet wurden.

Zunächst wurde der AGMA-Datensatz umfassend analysiert, um die wesentlichen Muster und Abhängigkeiten in den Daten zu identifizieren, sowie bestimmte Ausreißer aus den Daten zu entfernen. Diese Analyse bildete die Grundlage für die Auswahl und Anpassung geeigneter generativer Modelle zur Synthese von Mobilitätsdaten. Dabei wurden verschiedene Modelle hinsichtlich ihrer Fähigkeit, die komplexen räumlich-zeitlichen Muster des Datensatzes realistisch zu reproduzieren, evaluiert. Im Rahmen der Evaluierung wurde zunächst das TimeGAN-Modell auf den AGMA-Daten getestet. TimeGAN, ein Modell, das speziell für die Generierung von Zeitreihen-Daten entwickelt wurde, erwies sich als herausfordernd im Umgang mit zweidimensionalen GPS-Daten. Obwohl das Modell in der Lage war, zeitliche Muster zu erfassen, zeigte es Schwächen bei der Nachbildung von räumlichen Bewegungsmustern. Ein Hauptproblem bestand darin, dass die synthetischen Daten nicht den realen Straßennetzen folgten, sondern eher Luftlinien-Sprünge zwischen den Punkten darstellten. Dies beeinträchtigte die Nutzbarkeit der generierten Daten für Anwendungen, die auf genauen individuellen Trajektorien basieren, wie etwa Verkehrssimulationen. Die Evaluation von TimeGAN umfasste verschiedene Methoden zur Bewertung der synthetischen Datenqualität, näheres ist im Kapitel zu den Ergebnissen von AP2 (6) beschrieben. Trotz dieser umfassenden Evaluation zeigte sich, dass TimeGAN in seiner aktuellen Form nicht optimal für die Generierung von Mobilitätsdaten geeignet ist. Die Ergebnisse deuteten darauf hin, dass zukünftige Forschungen sich auf die Integration von Straßennetzen in die Modellarchitektur konzentrieren sollten, um die räumliche Genauigkeit der generierten Daten zu verbessern.

Zudem könnte die Anpassung und Optimierung von GANs, die speziell für Mobilitätsdaten entwickelt wurden, weitere Fortschritte in der Synthese realistischer Mobilitätsdaten ermöglichen. Zusammenfassend lässt sich sagen, dass die Nutzung von TimeGAN zur Synthese von Mobilitätsdaten wertvolle Erkenntnisse lieferte, jedoch auch die Notwendigkeit aufzeigte, spezialisierte Modelle weiter zu erforschen und anzupassen, um die Herausforderungen bei der realistischen Nachbildung räumlich-zeitlicher Daten zu überwinden.

Basierend auf diesen Erkenntnissen wurde eine weitere Fallstudie zur Synthetisierung von Mobilitätsdaten durchgeführt, welche vier weitere generative Modelle auf deren Tauglichkeit zur Anwendung auf den AGMA-Datensatz hin untersuchte: MoveSim [7], LSTM-TrajGAN [22], TrajGen [4] und TS-TrajGen [11]. Die Ergebnisse dieser Evaluierung zeigten, dass LSTM-TrajGAN und TS-TrajGen die vielversprechendsten Modelle für die Generierung synthetischer Mobilitätsdaten sind. LSTM-TrajGAN war in der Lage, synthetische Trajektorien zu erzeugen, die den realen Daten auf Mikroebene ähneln, jedoch wiesen alle synthetischen Trajektorien geografische Konzentrationen in der Mitte Deutschlands auf, unabhängig von den ursprünglichen Ausgangspunkten. TS-TrajGen hingegen, das eine zweistufige GAN-Architektur mit Verstärktem Lernen nutzt, konnte die Kontinuität der Trajektorien durch die Einbindung von Straßennetzen besser modellieren. Allerdings waren die Vorverarbeitungs- und Trainingsprozesse dieses Modells komplex und erforderten erhebliche Anpassungen der Datenformate und eine Reduzierung auf einen kleinen Teil des Datensatzes. Die weiteren notwendigen Schritte, um TS-TrajGen auf den gesamten AGMA-Datensatz anwenden zu können, umfassen die Verbesserung der Datenvorverarbeitung, die Durchführung des adversarialen Trainings auf dem gesamten Datensatz und die damit verbundene Generierung neuer Trajektorien.

6 Ergebnisse AP2: Methoden zur Bewertung von Generierungsmethoden

Im Zuge der oben bereits erwähnten Experimente wurden unterschiedliche Metriken zur Bewertung der synthetischen Daten erprobt, welche auf dem Vergleich der realen mit den synthetischen Daten basieren. Dabei konnten die Erkenntnisse gewonnen werden, dass eine Bewertung der synthetischen Daten grundsätzlich mit mehreren Metriken erfolgen sollte. Gerade Modell-basierte Metriken (solche bspw., die selbst ein Neuronales Netz trainieren, um die synthetischen Daten von den realen zu unterscheiden) sind nicht allgemeingültig, da schlechte Performance hier nicht allein auf die synthetischen Daten zurückzuführen ist, sondern ebenso aus einer schlechten Metrik-Modell-Performance herrühren können. Es ist also immer ein vollumfängliches Bild durch möglichst unterschiedliche Metriken nötig, um die Qualität von synthetischen tabellarischen Daten bewerten zu können. Um eine Vergleichbarkeit zu gewährleisten, ist es notwendig die jeweiligen Metrikwerte, falls sie sich nicht eh in diesen Grenzen befinden, auf Werte zwischen 0 und 1 zu beschränken, wobei 1 die perfekte Qualität bedeutet. Außerdem ist im Zuge der unterschiedlichen Varianten der Reproduktions-Algorithmen

deutlich geworden, dass die jeweiligen Aspekte (Datenschutz, Zeitreihen, Mobilität) natürlich spezialisierte Methoden zur Evaluierung der Daten benötigen. Auf diese wird im Folgenden näher eingegangen.

6.1 Tabellarische Daten

Für die Plattform wurde ein Evaluations-Service umgesetzt, welcher zunächst als Prototyp mit drei Metriken zum Vergleich von echten und synthetischen Daten entwickelt wurde. Die erste Metrik wurde „gewichtete Statistik“ genannt, da sie für die kontinuierlichen Spalten den Kolmogorov-Smirnov Testwert [28] berechnet und für die diskreten Spalten den Total Variation Distance Metrikwert [35] und dann den gewichteten Wert dieser beiden im Verhältnis der kontinuierlichen und diskreten Spalten zurückgibt. Beide Werte geben Auskunft über die Ähnlichkeit der Datenverteilung der jeweiligen Spalte, vergleichen also die echten und die synthetischen Daten auf Spaltenebene. Zusätzlich dazu wurde die „logistic Detection“ Metrik [29] implementiert, bei welcher ein logistisches Regressionsmodell darauf trainiert wird, die echten von den synthetischen Daten zu unterscheiden. Dabei ist der Metrikwert höher, umso weniger das Modell die Daten voneinander unterscheiden kann, die Daten werden hier auf Zeilenebene miteinander verglichen. Die dritte im Prototyp umgesetzte Metrik war die „ML Efficacy“ [30], bei welcher ein ML Modell auf den synthetischen Daten trainiert wird, eine Zielspalte vorherzusagen. Der echte Datensatz dient dann als Testdatensatz, auf welchem die Performance des auf synthetischen Daten trainierte Modell evaluiert wird. Diese Metrik vergleicht die Datensätze also eher indirekt, indem der Erhalt von Zusammenhängen zwischen den Spalten verglichen wird. In einem zweiten Entwicklungsschritt wurde der Evaluations-Service zum Vergleich von realen und synthetischen Daten um vier weitere Methoden von SDV erweitert. Diese bewerten die Qualität der synthetischen Daten in Aspekten wie der Einhaltung von Minimum und Maximum einzelner numerischer Spalten („Boundary Adherence“ [26]), der Übereinstimmung von Datenverteilungen der Merkmale („Value Coverage“, gewichtetes Mittel aus „Range Coverage“ [34] und „Category Coverage“ [27]), sowie Korrelationen zwischen mehreren Spalten („Column Pair Trends“ [33]). Außerdem wird berechnet, ob Zeilen kopiert wurden bzw. wie hoch der Anteil an kopierten Zeilen im synthetischen Datensatz ist („New Row Synthesis“ [31]).

6.2 Verfahren zur Überprüfung von Privatsphäre

Zur Überprüfung des Privatsphärenschutzes auf die vom privaten Reproduction-Service generierten Daten wurde das open-source Framework „Anonymeter“ [9] verwendet. Mit dem Ziel in synthetisierten tabellarischen Datensätzen verbleibende Datenschutzrisiken aufzudecken und zu bewerten, sind die Hauptindikatoren für Anonymisierung gemäß Datenschutzbestimmungen wie der Europäischen Datenschutzgrundverordnung (DSGVO) integriert. Explizit beinhaltet Anonymeter verschiedene Angriffsalgorithmen, um Risiken der Identifizierung, Verknüpfbarkeit und Inferenz zu ermitteln.

6.3 Zeitreihen-Daten

Im Zuge der Methodenentwicklung für Zeitreihen-Daten wurden unterschiedliche Metriken erprobt, welche die Erhaltung zeitlicher Zusammenhänge innerhalb der synthetischen Daten bewerten sollen. Der „discriminative score“ [39] misst, wie gut bzw. schlecht sich die Daten voneinander unterscheiden lassen. Hierfür wird ein rekurrentes Neuronales Netz auf den Zeitreihen trainiert, und soll dann die Zugehörigkeit (real/synthetisch) bestimmen – je schwieriger dies möglich ist, umso besser sind die synthetischen Daten. Eine weitere Metrik ist der „predictive score“ [39], bei welchem gemessen wird, wie gut die synthetischen Daten die „Vorhersagbarkeit“ der zeitlichen Zusammenhänge rekonstruieren. Dazu wird ein rekurrentes Neuronales Netz auf den synthetischen Zeitreihen trainiert, um die nächsten Werte vorherzusagen. Als Evaluierung dieser synthetischen Daten kann das Modell dann auf den echten Daten angewendet und die Soll- mit den vorhergesagten Werten verglichen werden. Wenn das Modell in der Lage ist, diese Werte gut vorherzusagen, erhalten die synthetischen Daten die zeitlichen Zusammenhänge gut. Ein Problem stellt aber die Negation dar: Sollte das Modell nicht in der Lage sein, gute Vorhersagen zu treffen, heißt dies nicht zwangsläufig, dass die synthetischen Daten schlecht sind, sondern auch das Modell selbst kann Schwachstellen aufweisen (falsche Architektur/ kein optimiertes Training etc.). Diese Problematik gilt auch für den „discriminative score“, und letztlich wie bereits erwähnt für jede Metrik die auf dem Training eines ML Modells basiert. Daher haben wir uns dazu entschieden, zunächst auch visuelle Metriken zu verwenden, um die Datenqualität zu überprüfen, auch wenn dies für das langfristige Ziel eines Plattform-Betriebs zur automatisierten Evaluierung von Daten nicht infrage kommt. Hier wurde eine Untermenge der echten und synthetischen Daten mithilfe von Methodiken zur Dimensionsreduktion und Visualisierung (PCA [3] und t-SNE [15]) so transformiert, dass sie gemeinsam in einem zweidimensionalen Raum dargestellt werden können. Der Grad der Überlappung dieser beiden Datenuntermengen zeigt dann, wie gut die Daten übereinstimmen.

6.4 Mobilitätsdaten

Die Evaluation der Mobilitätsdaten umfasste verschiedene Methoden zur Bewertung der synthetischen Datenqualität im Hinblick auf die Zeit- und Ortskomponenten. Dazu gehörten visuelle Vergleiche der realen und synthetischen Routen, wie die bereits erwähnte Hauptkomponentenanalyse (PCA) und t-SNE zur Darstellung der Datenverteilungen. Außerdem eignen sich Metriken, welche die echten und synthetischen Routen statistisch vergleichen, wie beispielsweise die Mean Distance Difference (MDD) und Mean Radius Difference (MRD), welche jeweils die mittlere Abweichung der Reisedistanz bzw. des geografischen Radius der Routen vergleichen. Diese Metriken halfen dabei, die Ähnlichkeit der realen und synthetischen Datensätze auf professioneller Ebene zu beurteilen. Zusätzlich wurde eine nachträgliche Klassifikation durchgeführt, um zu überprüfen, wie gut ein Klassifikator zwischen echten und synthetischen Routen unterscheiden kann, analog zum „discriminative score“ bei den Zeitreihen-Daten.

7 Ergebnisse AP3: Plattformentwicklung

Die Plattformentwicklung erfolgte iterativ, sowohl die Konzeption als auch die Implementierung, um Schwachstellen frühzeitig zu erkennen und diese zu beheben. Um die Plattform zu entwickeln, sind dabei hauptsächlich drei Schritte umgesetzt worden. Der erste Schritt bestand darin, Anforderungen, basierend auf Nutzerrollen, an die Plattform zu erheben. Basierend auf der Anforderungsanalyse wurde schrittweise ein Architekturentwurf für DaFne entwickelt und verfeinert, mit dem eine Grundlage für die anschließend durchgeführte Implementierung der Plattform geschaffen wurde. Während der Konzeption des Architekturentwurfs konnten potenzielle Optimierungen festgehalten werden, welche genauer in Kapitel 7.2 erläutert werden. Das Ergebnis der Implementierung wird in Kapitel 7.3 detailliert beschrieben.

7.1 Anforderungsanalyse Plattform

Die Anforderungsanalyse umfasste die Erhebung allgemeiner Anforderungen an die Plattform sowie Anforderungen potenzieller Plattformnutzer in unterschiedlichen Rollen. Für die Plattform wurden die in AP1 beschriebenen Möglichkeiten zur Datengenerierung als Anforderungen identifiziert, sodass die Plattform als Sammelpunkt für verschiedenartige Synthetisierungsansätze fungieren sollte. Einerseits sollte die Plattform mit der am L3S entwickelten Neighbourhood Generation und der Mobilitätsdatengenerierung zwar parametrisierbare, aber dennoch Use Case-spezifische Daten bereitstellen. Andererseits sollten mit der regelbasierten Erzeugung, der Data Fusion und der Vervielfältigung von tabellarischen Datensätzen generische Funktionalitäten für tabellarische Datensätze auf der Plattform bereitgestellt werden. Dabei wurden diverse Herausforderungen für den Entwurf der Plattform identifiziert, wie z.B. die Heterogenität der für die Generierung verwendeten Machine Learning Modelle, die Diversität der Anwendungsfälle oder Herausforderungen der Datenhaltung für unterschiedliche Akteure.

In einem spezifischen Analyseschritt wurden potenzielle Nutzertypen identifiziert: Auf Grundlage dieser Nutzertypen wurden reale Personen aus den Bereichen Urban Planning & Design, Produktmanagement für Data Analytics sowie Data-Literacy-Experten interviewt. Aus den kontextuellen Interviews konnten Nutzerprofile in Form von Personas erstellt werden, welche Informationen über Nutzerbedürfnisse in Bezug auf synthetische Daten, technische Kompetenzen sowie allgemeine Informationen zu der Arbeit mit Daten beinhalten. Die Nutzerforschung (UX/UI Design) war ein kontinuierlich andauernder Prozess, der dem GUI-Design und somit der Frontend-Entwicklung diente.

7.2 Architekturentwurf

Das Vorgehen zur Entwicklung der Architektur basierte auf mehreren Schritten. Aufgrund der unterschiedlichen Anforderungen, die sich durch die vielfältigen

Synthetisierungsmöglichkeiten auf der Plattform ergaben, wurde eine erste Architektur für die jeweiligen Funktionalitäten einzeln entworfen und nur das Frontend verbunden. Später ließen sich Gemeinsamkeiten identifizieren und der Entwurf iterativ verfeinern. Die einzelnen Funktionalitäten zur Datengenerierung wurden am Ende konzeptionell in einen Gesamtentwurf der Plattformarchitektur integriert. Die Einzelheiten der Architekturentwicklung werden im weiteren Verlauf beschrieben.

Der erste Architekturentwurf der Microservice-Anwendung wurde 2023 weiter angepasst. Die Architektur wurde neu verteilt in ein Frontend, mehrere Backends, eine Speicherverwaltung und Authentifizierungskomponenten, um eine effizientere Handhabung der verschiedenen Anforderungen zu ermöglichen. Ein Reverse-Proxy wurde eingeführt, der dazu diente, den eingehenden externen Datenverkehr effizient auf die einzelnen Plattformkomponenten zu verteilen, was die Skalierbarkeit und Sicherheit verbessert. Statt eines generischen Backends wurden spezialisierte Backends entwickelt, die auf spezifische Anwendungsfälle zugeschnitten sind, wodurch die Komplexität der Plattform reduziert und die Entwicklung beschleunigt werden sollte. Eine zentrale Speicherlösung unterstützt zudem die effiziente Verwaltung von Datensätzen, Gewichten und Parametern der trainierten Modelle.

Für das vom L3S entwickelte Verfahren der kartenbasierten Übertragung von Städte-Eigenschaften auf neue Städte (Neighbourhood Generation) wurde eine service-basierte Architektur entwickelt, in welcher die einzelnen Kernfunktionalitäten jeweils von unterschiedlichen Services umgesetzt werden. Dazu wurden zunächst mögliche Nutzerrollen sowie potenzielle Anwendungsfälle identifiziert, um anschließend Anforderungen bezüglich der Architektur zu definieren. Die notwendigen Services zur Umsetzung der vom L3S entwickelten Neighbourhood Generation wurden in die vorhandene gesamte Plattformarchitektur integriert. Diese entwickelte Architektur diente als Orientierung zur Integration weiterer Anwendungsfälle in die Plattform.

Für einen ersten Prototyp zur Datensynthese wurde die Reproduction, also die Erzeugung von zusätzlichen Zeilen basierend auf einem bereits bestehenden Datensatzes mithilfe von Machine Learning (ML) Modellen ausgewählt, und für diesen die Architektur entworfen. Hier ergab sich die Besonderheit, dass für jeden Datensatz zur Reproduction aus unterschiedlichen Modelltypen ausgewählt werden konnte, für jeden Modelltyp folglich auch individuelle Parameter gesetzt werden konnten (inkl. Netzarchitektur) und dann immer eine neue, individuelle Modellinstanz auf diesem Datensatz trainiert werden musste. Die Modelltypen wurden daher als Docker-Images mit festgelegten Schnittstellen verwaltet, sodass das Training einer Modellinstanz in einem Docker-Container stattfinden konnte.

Des Weiteren wurde die Architektur so entworfen, dass autorisierte Nutzer in der Lage sind, neue generative ML Modelle in die Plattform zu integrieren und diese so zu erweitern. Dazu werden im Backend Deep Generative Models verwendet, wie GANs, um neue Zeilen für Tabellen zu generieren, wobei besonderer Wert auf die Bereitstellung ausreichender Ressourcen gelegt wird, sodass

Modellierung und Evaluation auf derselben Hardware stattfinden können. Umgesetzt wird die Datengenerierung im Prototyp durch UserPods, welche intern aus mehreren Containern bestehen. Wenn ein Nutzer eine Generierung anfordert, startet der Orchestrator einen UserPod, der aus einem Sidecarcontainer zur Datenbereitstellung und weiteren Containern für Datenverarbeitung und -evaluation besteht. Der Orchestrator überwacht die Berechnungen im UserPod, um eine effiziente Verarbeitung zu gewährleisten. Nach Abschluss der Prozesse werden die Ergebnisse bereitgestellt, was eine schnelle und flexible Generierung von synthetischen Daten ermöglicht. Der resultierende Architekturentwurf schaffte die konzeptionelle Grundlage für die Integration weiterer Funktionalitäten, wie z.B. die Integration der regelbasierten Datensynthese sowie des in 6.1 beschriebenen Evaluations-Services für tabellarische Daten. Dem Evaluations-Service steht ein eigener Sidecarcontainer für den effizienten Datenzugriff zur Verfügung. Dieser basiert auf dem gleichen Image wie der zuvor beschriebene Sidecarcontainer der UserPods. Durch die Wiederverwendung des Codes wird eine gute Wartbarkeit bei reduzierter Fehleranfälligkeit erreicht. Durch den Entwurf des Prototypen konnten zusätzlich Anpassungen/Erweiterungen der Plattform, wie z.B. die Nutzung eines Message Broker's zur Durchführung rechenintensiver Prozesse, identifiziert werden.

Für die Kombination mehrerer Datensätze, also die Fusion von Datensätzen, ist ein Architekturentwurf konzipiert worden. Dabei handelt es sich um einen service-basierten Entwurf. Der Architektur-Entwurf ist mit einer einheitlichen Schnittstelle entwickelt worden, um effizient tabellarische Daten aus verschiedenen Datenquellen zu fusionieren. Der generische Entwurf, sowie dessen Implementierung, war in der Lage, Daten mit angemessener Genauigkeit und Robustheit zu fusionieren. Optimierungspotenziale konnten jedoch bezüglich der Speicher- und Ressourcenanforderungen festgestellt werden.

Die regelbasierte Datengenerierung, welche dazu in der Lage ist, eigene Spalten mit ausgewählten Daten, Verteilungen und Merkmalswerten basierend auf definierten Regeln zu generieren, ist konzeptionell als Architekturentwurf entwickelt worden. Dieser Entwurf schuf eine Grundlage zur Nutzung von Grundfunktionalitäten zur regelbasierten Datengenerierung über eine Nutzeroberfläche, z.B. die Erstellung/- und Erweiterung tabellarischer Daten durch Spalten.

Die Generierung synthetischer Fußgängerrouen, welche im Kern auf agentenbasierten Simulationen in Kombination mit verstärktem Lernen basiert, ist architektonisch ebenfalls berücksichtigt worden. Der hier konzipierte Architekturentwurf ermöglicht es, basierend auf Services, die tabellarische Daten durch agentenbasierte Simulation (Reinforcement Learning) innerhalb der Plattform zu generieren.

Die zuvor beschriebenen Ansätze zur Datengenerierung (Fusion, regelbasierte Datengenerierung, Generierung synthetischer Fußgängerrouen) wurden abschließend einheitlich konzeptionell in eine gesamte service-basierte Plattformarchitektur integriert. Dieser Plattformentwurf wurde, basierend auf funktionalen & nichtfunktionalen Anforderungen untersucht. Dabei konnten im Gesamtent-

wurf mögliche Verbesserungen, z.B. Anpassungen auf Service-Ebene, festgehalten werden.

7.3 Implementierung

Frontend-Implementierung Im Rahmen des Designs für das Minimum Viable Product (MVP) wurde zunächst eine Benutzeroberfläche entwickelt, die es Nutzern ermöglichen soll, den generischen Reproduction Algorithmus auszuführen. Die Ausführung lief in der frühen Projektphase lediglich mit einem Mock-Daten-Server, da das Backend aufgrund der Infrastrukturprobleme (Abschaltung der HAW-Infrastruktur aus Sicherheitsgründen aufgrund eines Cyber-Angriffs) keine sichere Ausführung bieten konnte. Auf Grundlage des Backend-Modells konnten die Nutzung des Reproduction Algorithmus simuliert und die Ergebnisse und Daten auf dem Nutzerprofil eingesehen und verwaltet werden. Beim Entwicklungsansatz wurde untersucht, wie die Plattform, die Artificial Intelligence as a Service (AIaaS) bietet, mithilfe einer Micro-Frontend-Architektur entworfen und umgesetzt werden kann. Dazu wurde für eine Persona aus der Nutzerforschung zuerst ein User-Flow zur Nutzung des Prototyps des Reproduction-Algorithmus modelliert. Nach der Entwicklung eines allgemeinen Design-System für die Plattform, konnte der modellierte User-Flow in ein UI-Design übersetzt werden. Anschließend wurde der modellierte User-Flow, gemeinsam mit unterstützenden Seiten wie Landing Page und Dashboards, ebenfalls in ein UI-Design übersetzt. Dabei wurde das initial entworfene Design im Laufe des Entwicklungsprozesses leicht an die Beschaffenheiten des Back- und Front-Ends angepasst. Des Weiteren wurde die Anwendung von Micro-Frontends zur Entwicklung der GUI untersucht. Dazu wurde die Benutzeroberfläche, welche im anfänglichen Architekturentwurf als ein Service konzipiert worden ist, in einzelne Frontend-Module aufgeteilt, welche aus verschiedenen Domänen stammen. Wie zuvor beschrieben, teilt diese Architektur die Plattform in mehrere Frontend-Anwendungen auf, ähnlich der Vorgehensweise bei Backend-Microservices. Das System umfasst: Eine Marketing-App mit einer Landing Page, die die Plattform bewirbt; Eine Authentifizierungs-App mit Nutzerregistrierung und Login; eine Theme-App, die ein Design System als Modul bereitstellt, eine DaFne-App, die primär den Reproduction Algorithmus und Dashboard-Funktionalitäten ausführt; sowie eine Neighborhood App, die in den Use Case Explorer integriert wird. Diese Struktur ist besonders wichtig für die geplante Erweiterung der Plattform durch Use-case-spezifische Features wie Neighborhood Generation oder Pedestrian Path Simulation. Diese Features greifen nicht in die Kernlogik der Plattform ein, basieren auf einem unabhängigen Datenmodell und benötigen keine aufwendige Konfiguration von Machine Learning Modellen, da sie vortrainierte Modelle bieten. Die Implementierung erfolgte hauptsächlich mit ReactJS und der Webpack 5 Library Module Federation. Besonders vorteilhaft erweist sich die Implementierung dadurch, dass die externe Neighborhood Generation als VueJS App integriert werden kann. Somit trägt sie dazu bei, dass auch Frontend-Features technologie-agnostisch erweitert werden können. Das Frontend wurde zum aktuellen Stand unabhängig von der Plattform-Infrastruktur entwickelt und auf

Diensten von Amazon AWS deployed. Der daraus resultierende Prototyp für die GUI ermöglichte Flexibilität der Systemgestaltung, allerdings konnte festgehalten werden, dass Micro-Frontends für andere Projektdimensionen aufgrund der Komplexität geeigneter wären. Nach dem Aufsetzen einer stabilen Backend-Infrastruktur wurde der Mock-Daten-Server durch den Produktiv-Server ersetzt. Somit kann mit dem Reproduction Algorithmus ein Modell trainiert werden und echte User-Daten wie Modelle, Jobs und Datensätze können auf dem Dashboard eingesehen und verwaltet werden. Der User-Flow kann vollständig abgeschlossen werden, indem der Nutzer die generierten Daten sowohl als Samples bereits auf der Plattform einsehen, aber auch auf den lokalen Rechner downloaden kann. Einige Backend-Eigenschaften haben zu minimalen Anpassungen der Ausführungslogik geführt. An dem untersuchten Micro-Frontend-Ansatz wurden keine weiteren großen Änderungen vorgenommen, da eine Architektur implementiert worden ist, die Änderungen und Erweiterungen zulässt. Trotz der Vorteile sind durch die Micro-Frontend-Architektur während des Wechsels der Deployment Strategie von AWS Diensten zu dem Kubernetes-Cluster auch Herausforderungen im Routing und Deployment entstanden.

Plattform-Implementierung Auf Grundlage des ersten Architekturentwurfs wurde die Microservice-Architektur weiter ausgeplant. Es wurde sich zunächst dazu entschieden den Prototypen auf der HAW Cloud-Infrastruktur zu betreiben. Folgende erste Prototypen der Komponenten wurden dazu implementiert: Bereitstellung der generellen Plattform, Nutzer- und Speicherverwaltung. Ein MVP (Minimum Viable Product) sollte bis Anfang des Jahres 2023 umgesetzt werden, bei dem der Reproduktions-Use Case durchführbar gewesen wäre. Durch den Cyber-Angriff auf die HAW Ende 2022 wurde das Weiterarbeiten verhindert, da die Infrastruktur der HAW aus Sicherheitsgründen komplett abgeschaltet werden musste. In der Implementierungsphase 2023 fokussierte sich das Projekt auf die Anpassung und Weiterentwicklung der Plattform, um eine lokale Nutzung und eine provisorische Cloudinfrastruktur zu ermöglichen. Aufgrund des Ausfalls der Cloudinfrastruktur der HAW mussten die Plattformanwendungen und der Infrastrukturcode so angepasst werden, dass eine lokale Nutzung mittels Minikube möglich wurde, wodurch die Entwicklung auch ohne externe Cloudinfrastruktur fortgesetzt werden konnte. Im September 2023 befasste sich ein Entwickler mit dem Aufbau einer provisorischen Cloudinfrastruktur, auf der die Plattform im MVP-Status betrieben werden konnte. Diese Infrastruktur basierte aufgrund verschiedener Gegebenheiten auf ARM-Prozessoren. Das machte eine Anpassung der Anwendung zur Unterstützung multipler Architekturen erforderlich. Das Backend für die Generierung von synthetischen tabellarischen Daten erreichte im November eine Entwicklungsstufe, die als prototypisch betrachtet werden konnte. Zudem wurde eine API-Dokumentation erstellt und Skripte entwickelt, damit neue Teammitglieder die Plattform lokal aufbauen und weiterentwickeln können. Dies förderte die Forschung und Entwicklung in diesem Bereich auch von Entwicklern, die nicht im Infrastrukturteam sind. Die Arbeiten der Entwickler an der DaFne-Plattform fokussierten sich hauptsächlich auf die Integration und

das Zusammenspiel von Frontend und Backend. Diese Phase war zeitintensiv und ermöglichte die Behebung zahlreicher Bugs in beiden Bereichen. Ein weiterer Schwerpunkt lag auf der Bereitstellung der Plattform unter der öffentlich zugänglichen URL: <https://platform.dafne-project.org/>. Im Verlauf des Projekts wurde zudem an der universellen Einsatzfähigkeit der Plattform gearbeitet. Mit Hilfe eines Helm Charts wurde die Plattform so konfiguriert, dass sie leicht ausgerollt und somit in Folgeprojekten oder auf anderer Infrastruktur betrieben und migriert werden kann. Die Dokumentation des Ausrollprozesses wurde vorangetrieben und in der täglichen Arbeit einiger Entwickler getestet sowie weiter verbessert. Diese Maßnahmen tragen dazu bei, die Plattform flexibler und benutzerfreundlicher zu gestalten und ihre Anwendungsbreite zu erweitern.

8 Ergebnisse AP4: Evaluation der Ergebnisse im Anwendungsfeld Smart Cities

Zur Evaluierung der tabellarischen Synthetisierungsfunktionalitäten auf der Plattform (regelbasierte Erzeugung, Data Fusion, Vervielfältigung) wurde ein Predictive Maintenance Use Case für Brücken identifiziert. Hier wurden die Plattformfunktionalitäten eingesetzt, um einen Ausgangsdatensatz (Zustand von Brücken) anzureichern. Die Evaluierung anhand des Predictive Maintenance Use Cases sollte dann zeigen, ob die Funktionalitäten zu einer Verbesserung der Vorhersagefähigkeit führen und so als Unterstützung für KI- und ML-Use Cases eingesetzt werden können. Dazu wurden zunächst im Zuge des Preprocessings mithilfe des regelbasierten Services bestimmte Spalten, wie beispielsweise die `unique_ID` generiert. Basierend auf dem resultierenden Datensatz wurde dann zunächst der Fusion Service angewandt, um den Wetter-Datensatz zu erzeugen: hier wurden die Brückendaten um die jeweiligen Wetter-Durchschnittswerte der betrachteten Jahre und Orte erweitert, da die Vermutung nahe lag, dass bestimmte Wetterbedingungen den Brückenzustand beeinflussen und so bei der Vorhersage helfen können. Außerdem wurde der Datensatz verwendet, um einen synthetischen Brückendatensatz zu erzeugen mithilfe des in Kapitel 5.3 (AP1) beschriebenen selbst entwickelten Modell zur Generierung von synthetischen Zeitreihen. Hier sollte der Predictive Maintenance Use Case zeigen, ob die Daten als valider Ersatz der Original-Daten dienen können und so die Synthese-Qualität und den Service an sich bewerten. Dabei ist vor allem aufgefallen, dass die Wetter-Durchschnittswerte, die der Fusion-Service hinzufügte, leider keine Verbesserung der Prognose hervorrief, da sich die Werte über die einzelnen Jahre nicht signifikant unterscheiden und somit keinen Informationszugewinn darstellen. Vielmehr würden vermutlich Extremwetter-Daten wie beispielsweise die Anzahl sehr heißer Tage pro Jahr helfen, eine solche Prognose zu verbessern, sodass hier eine Erweiterung des Fusion-Service erfolgen muss. Außerdem sind bei den synthetischen Daten Defizite im Bereich der Abbildung kategorialer Merkmale aufgefallen, da unterrepräsentierte Ausprägungen oft in den synthetischen Daten gar nicht mehr repräsentiert sind. Auch hier ist Raum für weitere Verbesserungen des generativen Modells deutlich geworden. Die Ergebnisse dieser Anwendungsfall-

Studie sind bei der ICICT 2024 in London präsentiert worden und werden dort veröffentlicht ([2]).

Die sog. AGMA-Daten, Mobilitätsdaten von Einzelpersonen aus über 10 Jahren, bereitgestellt von der AGMA (siehe auch Ergebnisse von AP 1, Kapitel 6), wurden analysiert und, basierend auf den Erkenntnissen der Analyse, vorverarbeitet. Hier wurden vor allem inkonsistente oder unwahrscheinliche Wege (bspw. bei großen Sprüngen von einem Ort zum anderen) identifiziert und entfernt. Der Datensatz wurde dann verwendet, um mithilfe von unterschiedlichen generativen ML-Modellen, welche spezialisiert sind auf die Synthese von Zeitreihen- bzw. Mobilitätsdaten, synthetische Mobilitätsdaten zu erzeugen. Leider bleiben die Ergebnisse bisher hinter den Erwartungen zurück, da die Komplexität, realistische geografische Zeitreihendaten zu erzeugen, welche sich an echten Straßen- und Wegenetze orientieren und keine zeitlichen oder geografischen Sprünge enthalten, sich als sehr hoch herausstellte. Hier ist weitere Forschung und auch Entwicklungsarbeit hinsichtlich der Modelle notwendig, um solche Daten verlässlich synthetisieren zu können.

9 Fazit

Im Laufe des DaFne-Projekts wurden verschiedene Ansätze und Methoden zur Synthese von unterschiedlichen Daten untersucht und weiterentwickelt. Die zentrale Fragestellung war zum einen, wie diese Methoden auf vielfältige Anwendungsfälle, insbesondere im Bereich Smart Cities, angewendet werden können. Zum anderen war es besonders für uns ein Kernziel, zu untersuchen, wie man diese Methoden auf einer flexiblen und benutzerfreundlichen Plattform zur Verfügung stellen kann, welche skalierbar und erweiterbar ist, um State-of-the-art-Generierungsmethoden nutzbar machen zu können. In diesem Fazit werden die zentralen Forschungsfragen, welche in Kapitel 3 formuliert wurden, erneut aufgegriffen und die erzielten Erkenntnisse zusammengefasst, um einen umfassenden Überblick über die Projektergebnisse zu geben.

9.1 Forschungsfrage 1: Welche Methoden bzw. Modelle eignen sich zur flexiblen Synthese von tabellarischen Daten und lassen sich damit auf Anwendungsfälle im Bereich Smart Cities und darüber hinaus anwenden?

Die von uns im DaFne-Projekt untersuchten Methoden zur generischen Synthese von tabellarischen Daten umfassen die regelbasierte Erzeugung, die Fusion und die Reproduktion von Daten mittels generativer Modelle. Diese Ansätze wurden gezielt entwickelt und evaluiert, um eine breite Palette von Anwendungsfällen, insbesondere im Smart City Bereich, abzudecken. Bei der regelbasierten Erzeugung handelt es sich um eine einfache Methode, um Mockup-Daten zu erstellen, wenn noch keine anderen Daten vorhanden sind. Hierbei wird keine bestehende Datengrundlage benötigt, sondern die Daten können selbst Spalte für Spalte erzeugt werden. Im Zuge des DaFne-Projekts wurde ein Prototyp entwickelt, der

dies umsetzt. Um die Anwendbarkeit und den Nutzen zu erhöhen, können in Zukunft mehr Datentypen integriert werden, insbesondere solche aus dem Smart City Bereich, sowie unterschiedliche Datenverteilungen, um die Realitätsnähe der erzeugten Daten noch zu verbessern.

Die Fusion von Datensätzen kann den Informationsgehalt eines bestehenden Datensatzes erhöhen, indem er mit einem anderen Datensatz angereichert wird. Auch hierfür wurde ein Prototyp entwickelt, der parametrisiert zwei Datensätze miteinander vereinen kann. Außerdem ist der Prototyp in der Lage, die Wetter-API des Deutschen Wetterdienstes zu integrieren, um einem Datensatz Wetterinformationen hinzuzufügen. Für einen höheren Nutzen bei der Auswertung über verschiedene Jahre hinweg ist eine zusätzliche Aufbereitung der Wetterdaten notwendig (Extremwetterdaten), um den Informationsgehalt signifikant zu erhöhen. Damit stellt die Fusion mit Wetterdaten ein nützliches Werkzeug dar, um tabellarische Daten zu erweitern.

Die Reproduktion von Daten mittels generativer Machine Learning Modelle ermöglicht es, einen synthetischen Datensatz zu erzeugen, der dem Original-Datensatz in seinen statistischen Eigenschaften und Ausprägungen der jeweiligen Merkmale ähnelt, ohne dabei einzelne Datenpunkte zu kopieren und somit potenziell sensible Informationen preiszugeben. Unsere Forschung hat gezeigt, dass je nach Ziel der Datensynthese unterschiedliche Modelle für verschiedene Datentypen besser geeignet sind. Für allgemeine tabellarische Daten haben sich die Modelle CTGAN [38] und TVAE [38] als geeignet erwiesen, wobei Anpassungen der Modellarchitektur die Performance signifikant verbessern können. Auch das CTAB-GAN+ [40] hat sich als vielversprechend gezeigt, obwohl es noch Herausforderungen bei der Fehleranfälligkeit für verschiedene Datentypen gibt, was weitere Verbesserungen und Vorverarbeitungsschritte erforderlich macht.

Für die datenschutzkonforme Reproduktion zeigten Modelle wie DPGAN [37], PATE-GAN [19] und das bereits erwähnte CTAB-GAN+ gute Ergebnisse bei der Wahrung der Privatsphäre der Daten. Besonders das CTAB-GAN+ erzielte die besten Ergebnisse bei den getesteten Datensätzen, sodass sich eine Verbesserung des Modells zur breiteren Anwendbarkeit auch hier als lohnenswert herausstellte.

Bei der Synthese von Zeitreihen-Daten hat sich die Entwicklung eines eigenen Modells als die vielversprechendste Methode herausgestellt. Dieses Modell kombiniert ein LSTM [10] mit einem VAE [13], um die zeitlichen Zusammenhänge in den Daten korrekt abzubilden. Andere untersuchte Modelle wie das TimeGAN [39], TTS-GAN [14] und das Probabilistic Autoregressive (PAR) [32] Modell erzielten auf den spezifischen Daten keine zufriedenstellenden Ergebnisse.

Die Reproduktion von Mobilitätsdaten stellte sich im Zuge unserer Forschung als besonders komplex heraus, da sowohl geografische als auch zeitliche Zusammenhänge korrekt abgebildet werden müssen. Je nach Anwendungsfall für die synthetischen Mobilitätsdaten erwiesen sich verschiedene Modelle als vielversprechend. Für aggregierte Analysen eignet sich LSTM-TrajGen [22] am besten. Sollten die synthetischen Daten jedoch echten Straßennetzen folgen, wie es beispielsweise für Verkehrsstrom-Analysen erforderlich ist, zeigt sich der zweistufige

Ansatz TS-TrajGen [11] als vielversprechend, wenngleich die Vorverarbeitungsschritte wie bspw. das Map Matching noch viel Entwicklungsarbeit erfordern.

Zusammenfassend lässt sich sagen, dass die von uns im DaFne-Projekt untersuchten Methoden und einige der Modelle eine flexible Synthese von tabellarischen Daten ermöglichen und sich auf eine Vielzahl von Anwendungsfällen im Bereich Smart Cities und darüber hinaus anwenden lassen.

9.2 Forschungsfrage 2: Welche Qualitätskriterien sind notwendig, um die Brauchbarkeit und Integrität synthetischer Daten sicherzustellen?

Im DaFne-Projekt haben wir verschiedene Methoden und Metriken zur Bewertung der Qualität synthetischer Daten entwickelt und getestet, um sicherzustellen, dass die generierten Daten brauchbar und integritätswahrend sind. Dabei handelt es sich um die synthetischen Daten, die durch Reproduktion, also mithilfe von generativen ML Modellen entstanden sind, sodass sich die Original- mit den erzeugten Daten vergleichen lassen. Bei den hierzu durchgeführten Experimenten wurde schnell deutlich, dass es unerlässlich ist, mehrere Metriken zur Bewertung heranzuziehen, um die Brauchbarkeit und Integrität synthetischer Daten zu gewährleisten. Eine alleinige Metrik kann oft nicht die Komplexität und Vielfaltigkeit der Daten vollständig abbilden. Modell-basierte Metriken, wie solche, die ein neuronales Netz trainieren, beispielsweise um die synthetischen von den realen Daten zu unterscheiden, können von der Performance des Modells selbst beeinflusst werden und sollten daher immer auch mit anderen Metriken betrachtet werden. Um ein umfassendes Bild über die Datenqualität erzeugen zu können, sollten die verwendeten Metriken außerdem auf eine einheitliche Skala von 0 bis 1 normiert werden, wobei 1 die perfekte Qualität repräsentiert.

Die Metriken zur Bewertung der „normalen“ tabellarischen Daten lassen sich dabei immer auf alle Datensätze anwenden. Sieben von diesen wurden von uns in einem Evaluationsservice implementiert und auf der Plattform bereitgestellt, damit die Daten vollumfänglich bewertet werden können. Ein Beispiel für eine solche Metrik ist die "Value Coverage" Metrik, welche die Übereinstimmung von Datenverteilungen bewertet, weitere sind in Kapitel 6 beschrieben.

Sobald es sich um spezielle Datensätze handelt bzw. spezielle Aspekte dieser Datensätze bewertet werden müssen, ist die Verwendung geeigneter, spezialisierter Metriken unerlässlich: für die Bewertung der Einhaltung der Privatsphäre sind dann andere Metriken notwendig, als für die Bewertung, ob und wie gut zeitliche oder geografische Zusammenhänge in den Daten abgebildet wurden.

Zusammenfassend lässt sich also sagen, dass die Verwendung eines umfassenden Sets an Metriken, angepasst an die jeweiligen Datentypen und Anwendungsfälle, unerlässlich ist, um die Qualität synthetischer Daten sicherzustellen. Dies ermöglicht eine differenzierte Bewertung der Daten und stellt sicher, dass die generierten Daten brauchbar und integritätswahrend sind.

9.3 Forschungsfrage 3: Wie kann eine Plattform zur Datengenerierung konzipiert werden, die sowohl benutzerfreundlich als auch erweiterbar und skalierbar ist?

Durch das DaFne-Projekt konnte das Vorgehen zur Entwicklung von Plattformen zur Datengenerierung umfangreich untersucht werden. Diese Untersuchung lieferte wertvolle Erkenntnisse. Dabei hat das im Projekt gewählte Vorgehen einen Ansatz dafür geliefert, eine Datengenerierungs-Plattform zu konzipieren, welche benutzerfreundlich, erweiterbar und skalierbar ist.

Um die DaFne-Plattform konzeptionell zu entwickeln, wurde die Plattform zunächst aus der Sicht von potenziellen Nutzern charakterisiert. Die Identifikation von Nutzerrollen lieferte Ergebnisse über allgemeine und domänenspezifische Nutzerbedürfnisse zur Generierung synthetischer Daten, um die Benutzerfreundlichkeit zu gewährleisten. Das DaFne-Projekt hat gezeigt, dass Nutzerforschung als kontinuierlicher Prozess berücksichtigt werden sollte, um eine Plattform zu Datengenerierung zu entwickeln.

Die in Kapitel 7.1 beschriebene Anforderungsanalyse belegt, dass sowohl nichtfunktionale Anforderungen als auch funktionale Anforderungen für die Plattform eine große Rolle spielen. Die Definition nichtfunktionaler Anforderungen erleichterte die Implementierung der Plattform. Die Gewährleistung von Erweiterbarkeit, Skalierbarkeit und Wartbarkeit der Plattform einerseits und das Angebot einer breiten Palette an Generierungsfunktionalität andererseits stellte im weiteren Verlauf eine Herausforderung für den Architekturentwurf dar. Für eine weitere Erforschung entsprechender Architekturmuster wird ein Bedarf gesehen. Die funktionalen Anforderungen schafften eine erste Grundlage dafür Use Case-spezifische Daten zu generieren und somit die prototypische Funktionalitäten in DaFne (Reproduction, Fusion etc.) umzusetzen.

Im DaFne-Projekt wurden erste Schritte zur Erforschung von Architekturmustern für Maschinelles Lernen unternommen. Die Untersuchung bestehender Architekturstile hat gezeigt, dass Microservices einen grundsätzlich geeigneten Ansatz darstellen, um die Architektur für eine Datengenerierungs-Plattform zu konzipieren. Dies ist der Fall, da gemeinsame Funktionalitäten zur Datengenerierung in Services gebündelt werden können. Die Services können dann einheitlich in eine Gesamtarchitektur integriert werden. Hierdurch kann die Erweiterbarkeit der Datengenerierungs-Plattform konzeptionell gewährleistet werden. Die Aufteilung in ein Frontend, mehrere Backends sowie die Einführung eines Reverse-Proxy, welcher die Verteilung des Datenverkehrs auf einzelne Plattformkomponenten verwaltet, dienen zusätzlich dazu, die Skalierbarkeit des Entwurfs zu erhöhen. Andererseits können jedoch auch bei dem vorerst finalen Microservice-Entwurf für DaFne mögliche Verbesserungen identifiziert werden. Dabei handelt es sich um Anpassungen auf Service-Ebene.

Zusammenfassend ist festzuhalten, dass das im Rahmen des DaFne-Projekts gewählte Vorgehen dazu geeignet ist, eine benutzerfreundliche, erweiterbare und skalierbare Plattform zur Datengenerierung zu konzipieren. Die im DaFne-Projekt durchgeführten Arbeiten sind ein Beleg dafür, dass der Entwurf von Datengenerierungs-Plattformen wie in DaFne aufgrund der heterogenen, teilweise recht spezi-

fischen Anforderungen und Anwendungsfälle herausfordernd ist und sich weitere Untersuchungen in diese Hinsicht lohnen, um KI-basierte Services einem großen Anwendungspublikum zugänglich machen zu können.

9.4 Forschungsfrage 4: Können synthetische Daten zur Verbesserung spezifischer Anwendungsfälle aus dem Bereich Smart Cities beitragen?

Im Rahmen des DaFne-Projekts wurde die Anwendung synthetischer Daten in spezifischen Anwendungsfällen im Bereich Smart Cities umfassend untersucht und evaluiert. Ein zentrales Beispiel hierfür war der Predictive Maintenance Use Case für Brücken. Hierbei haben wir, gemeinsam mit unseren Partnern, verschiedene Plattformfunktionalitäten wie die regelbasierte Erzeugung, die Daten-Fusion und die Vervielfältigung von Datensätzen eingesetzt, um den Ausgangsdatensatz über den Zustand von Brücken anzureichern und zu replizieren. Die Anwendung dieser Funktionalitäten sollte die Vorhersagefähigkeit verbessern und so die Entscheidungsfindung im Kontext von KI- und ML-Use Cases unterstützen.

Die Ergebnisse zeigten, dass die synthetischen Datensätze, potenziell wertvolle Zusatzinformationen liefern bzw. ein realistischer Ersatz sein könnten. Jedoch wurde deutlich, dass die verwendeten Wetter-Durchschnittswerte keinen signifikanten Informationszugewinn darstellten, sodass, basierend auf dieser wertvollen Erkenntnis, zukünftig detailliertere Wetterdaten (wie Extremwetterereignisse) die Prognosegenauigkeit verbessern und somit die Nutzenfähigkeit der synthetischen Daten erhöhen können.

Zusammenfassend lässt sich sagen, dass synthetische Daten erhebliches Potenzial zur Verbesserung spezifischer Anwendungsfälle im Bereich Smart Cities besitzen. Insbesondere die Möglichkeit, bestehende Datensätze anzureichern und realistische, datenschutzkonforme Alternativen zu echten Daten bereitzustellen, eröffnet vielfältige Anwendungsmöglichkeiten. Jedoch erfordern die Komplexität und die spezifischen Anforderungen der verschiedenen Anwendungsfälle kontinuierliche Weiterentwicklungen und Anpassungen der verwendeten Methoden und Modelle.

10 Projektkosten: Zahlenmäßiger Nachweis

In diesem Abschnitt werden die wichtigsten Positionen des zahlenmäßigen Nachweises dargestellt, diese sind in Tabelle 1 zusammengefasst.

Obwohl das DaFne-Projekt zu Dezember 2021 bewilligt wurde, erfolgte aufgrund von erschwelter Rekrutierung die Einstellung der Mitarbeiter erst im Laufe des Jahres 2022. Insgesamt waren für das Projekt bei HITeC sechs Mitarbeiter eingestellt, welche von 2022 bis 2024 für das Projekt gearbeitet haben. Davon war eine Mitarbeiterin in Vollzeit angestellt, während fünf weitere Mitarbeiter in unterschiedlichen Teilzeitmodellen (10-30 Stunden pro Woche) beschäftigt waren. Insgesamt belaufen sich die Personalkosten auf 285.957,30 €, wobei mit 278.023,56 € der Großteil auf die Beschäftigten E12-E15 entfällt.

Tabelle 1. Gesamtausgaben im DaFne-Projekt 2022-2024

Position	Beschreibung	Entstandene Ausgaben
0812	Beschäftigte E12-E15	278.023,56
0820	Studentische Hilfskraft	7.933,74
0843	Sonstige allgemeine Verwaltungsausgaben	18.625,30
0846	Dienstreisen	5.838,90
	Summe	310.421,50

Für Projekttreffen und Besuche von Konferenzen sind Kosten für Dienstreisen i.H.v. insgesamt 5.838,90 € angefallen. Außerdem sind Kosten i.H.v. 18.625,30 € für allgemeine Verwaltungsausgaben entstanden, sodass sich die Summe für das DaFne-Projekt bei HITEC auf insgesamt 310.421,50 € beläuft.

11 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Das DaFne-Projekt war angesichts fehlender, datenschutzkonformer Lösungen für die Generierung realistischer Trainingsdaten im Smart-City-Kontext von vornherein notwendig: Es schließt eine wesentliche methodische und technologische Lücke, indem es Methoden der synthetischen Datenerzeugung und -fusion in einem prototypischen Plattform-Ansatz zusammenführt und so erstmals die Bereitstellung realistischer, anpassbarer Datensets ermöglicht.

In vier eng verzahnten Arbeitspaketen wurden die methodischen Grundlagen (regelbasierte Erzeugung, Fusion, Reproduktion mittels generativer Modelle), die systematische Bewertung dieser Verfahren, die prototypische Umsetzung einer skalierbaren und erweiterbaren Microservice-Architektur sowie deren Evaluierung anhand realer Use Cases (etwa Predictive Maintenance für Brücken) strukturiert bearbeitet und iterativ verbessert. Dabei ermöglichte der gezielte und effiziente Einsatz von Ressourcen über die gesamte Projektlaufzeit (Dezember 2021–Februar 2024) sowie das zur Verfügung stehende Budgetvolumen eine termingerechte Umsetzung aller definierten Meilensteine sowie die erfolgreiche Entwicklung eines praxisnahen Prototyps zur Datensynthese und -evaluation.

Durch die enge Zusammenarbeit mit relevanten Industriepartnern und die Einbindung in Folgeprojekte wie INSIGHT (vgl. Kapitel 12) ist zudem sichergestellt, dass die entwickelten Ansätze nachhaltig weiterentwickelt und in neuen Anwendungsfeldern erprobt werden können. Die projektbegleitenden Evaluationen in konkreten Anwendungsfällen verdeutlichten zudem, dass die gewählten methodischen und technischen Lösungswege den zuvor identifizierten Anforderungen angemessen waren, zugleich aber auch Handlungspotenziale für die weitere Optimierung der Ansätze aufzeigen konnten.

Zusammengefasst belegen Umfang, Tiefe und Ausgestaltung der geleisteten Arbeiten, dass diese sowohl wissenschaftlich notwendig als auch hinsichtlich ihres Ressourceneinsatzes und ihrer prototypischen Validierung angemessen waren.

12 Ausblick: voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse

Das DaFne-Projekt hat uns als HITeC (und HAW Hamburg) wertvolle Erkenntnisse und umfangreiches Wissen im Bereich der Datensynthese sowie der Konzipierung und Umsetzung einer flexiblen Plattform-Architektur und Infrastruktur geliefert. Diese Erfahrungen sind nicht nur für unsere zukünftigen Forschungsprojekte von unschätzbarem Wert, sondern bieten auch konkrete Möglichkeiten für die Verwertung der Projektergebnisse. Durch die enge Zusammenarbeit mit den DaFne-Partnern konnten starke Partnerschaften aufgebaut werden, die über das Projekt hinaus Bestand haben und zukünftige Kooperationen erleichtern. Mit dem Projektpartner Ströer finden derzeit Gespräche über eine Weiterverwendung und die Erweiterung der erzielten Ergebnisse im wirtschaftlichen Kontext des Unternehmens statt. Eine Vernetzung mit der Hamburger 12min.me-Community hat zudem großes Potenzial, die Sichtbarkeit und Reichweite unserer Forschungsergebnisse zu erhöhen, was weitere Kollaborationen und Projekte anstoßen kann.

Ein konkretes Folgeprojekt, das direkt aus den Erkenntnissen und Erfolgen des DaFne-Projekts hervorgeht, ist das Forschungsprojekt „INtelligent Synthesis and Generation of High-quality Test data“, kurz INSIGHT genannt. Gemeinsam mit unserem Partner AKARA Solutions GmbH widmen wir uns im Zuge des Projekts der Generierung von Testdaten, um so die Softwareentwicklung nachhaltig verbessern zu können. Dieses Projekt zielt darauf ab, die entwickelten Methoden und Technologien aus DaFne zu verfeinern und auszubauen, und in neuen Anwendungsfeldern zu erproben. Der Weiterbetrieb der DaFne-Plattform, bzw. ausgewählten Teilen davon, ist außerdem Bestandteil dieses Vorhabens. Unser Forschungspartner verfügt über eine eigene Cloud-Infrastruktur, über welche die Plattform deployed werden kann, sodass hier noch offene Integrationsarbeiten stattfinden können. INSIGHT wird durch die Investitionsförderbank der Freien und Hansestadt Hamburg gefördert. Hierbei handelt es sich um Wirtschaftsförderung, was belegt, dass die in DaFne erzielten Ergebnisse bereits eine anerkannte wirtschaftliche Relevanz besitzen.

Angedacht ist derzeit auch, das in DaFne erworbene Wissen weiteren Wirtschaftspartnern in offenen Workshops zum Thema Datengenerierung zugänglich zu machen, was wiederum auf die sachkundige Verwendung sowohl von Datensynthesierung im Speziellen als auch für den Einsatz von KI ganz allgemein einzahlen würde.

Im Bereich der akademischen Forschung befindet sich HITeC (HAW) mit den Erkenntnissen des DaFne-Projekts an einer für die Umsetzung von KI-Projekten relevanten Schnittstelle zwischen Machine-Learning-basierter Datengenerierung und dem Einsatz von Machine Learning einerseits und der Entwicklung entsprechend flexibler, skalierbarer und sicherer Softwaresysteme andererseits. Die Betrachtung solcher fachlicher Schnittstellen ist relevant für eine erfolgreiche Weiterentwicklung und Umsetzung von KI-Methoden. An der HAW Hamburg formiert sich derzeit eine interdisziplinäre Gemeinschaft aus Forschenden, die sich dieser und weiteren Schnittstellen widmen, um auf dieser Grundlage allge-

meine Gelingensbedingungen für KI zu erforschen. Das DaFne-Team der HAW ist ein wichtiger Teil dieser Community, für die auch eine entsprechende Forschungsförderung angestrebt ist.

Zusammenfassend lässt sich sagen, dass die im DaFne-Projekt gewonnenen Erkenntnisse und entwickelten Technologien eine sehr gute Basis für zukünftige Forschungs- und Anwendungsprojekte bieten. Die Vernetzung mit starken Partnern, der umgesetzte Prototyp einer flexiblen Plattform, die gewonnen konzeptuellen Erkenntnisse sowie die Planung und Durchführung konkreter Folgeprojekte gewährleisten, dass die Ergebnisse des DaFne-Projekts nachhaltig genutzt und weiterentwickelt werden.

13 Fortschritt auf dem Gebiet bei anderen Stellen

Während der Laufzeit des Vorhabens wurden zahlreiche relevante Fortschritte auf dem Gebiet der synthetischen Datengenerierung sowie im Bereich der Software- und Plattformarchitektur an anderen Stellen veröffentlicht, die einen unmittelbaren Bezug zum DaFne-Projekt haben und dieses in seiner Zielsetzung bestätigt, ergänzt oder konkret beeinflusst haben.

Im Bereich der synthetischen Datenerzeugung wurde insbesondere das Modell CTAB-GAN+ [40] als Weiterentwicklung etablierter Ansätze wie CTGAN und TVAE [38] veröffentlicht. Es adressiert spezifische Schwächen dieser Vorgängermodelle, etwa im Umgang mit heterogenen tabellarischen Daten, und kombiniert moderne Techniken wie mode-spezifische Normalisierung und Gumbel-Softmax. Im DaFne-Projekt wurde das Modell exemplarisch evaluiert und als Benchmark zur vergleichenden Analyse bestehender Verfahren herangezogen.

Auch im Bereich der zeitbasierten Datensynthese wurden relevante Fortschritte erzielt, insbesondere durch die Entwicklung und Verbesserung Transformer-basierter Modelle wie TTS-GAN [14], die eine differenziertere Abbildung komplexer zeitlicher Zusammenhänge ermöglichen. Im Projektkontext wurden diese Ansätze in einer Vorstudie zur Generierung synthetischer Zeitreihen hinsichtlich ihrer Qualität und ihrer Eignung für spezifische Anwendungsfälle (Predictive Maintenance für Brücken) bewertet. Dabei zeigte sich, dass neuere Modelle zwar Fortschritte bei der Vielfalt der generierten Daten erzielen, aber nach wie vor Einschränkungen im Hinblick auf Realitätsnähe und Grenzwerttreue aufweisen.

Für die Generierung synthetischer Mobilitätsdaten konnten insbesondere die Modelle LSTM-TrajGAN [22] und TS-TrajGen [4] als relevante Entwicklungen identifiziert werden. Beide wurden kurz vor Projektbeginn veröffentlicht und in DaFne im Rahmen einer prototypischen Evaluation untersucht. In der Evaluation überzeugten die Modelle durch eine realistische Abbildung räumlich-zeitlicher Muster, offenbarten jedoch zugleich Schwächen in der semantischen Steuerung und bei der Detaillierung der Trajektorien.

Auch im Bereich der Softwarearchitektur und Plattformentwicklung wurden im Projektzeitraum relevante Beiträge veröffentlicht, die mit den Architekturentscheidungen und dem Plattformansatz von DaFne korrespondieren. So stellen

Markov et al. [16] mit ihrer Analyse skalierbarer End-to-End-ML-Plattformen ein umfassendes Rahmenwerk vor, das sowohl AutoML-Funktionalitäten als auch Self-Service-Architekturen adressiert. Die dort diskutierten Anforderungen und Strukturen wurden im DaFne-Projekt aktiv aufgegriffen, etwa bei der Modularisierung der Plattform und bei der nutzerzentrierten Gestaltung der Workflows.

Die Plattform Looper [17] dient als konkretes Beispiel für die Umsetzung produktzentrierter Entscheidungsunterstützung mittels integrierter ML-Architekturen. Ihre strukturierte Pipeline-Architektur floss in Überlegungen zur Datenflusssteuerung innerhalb der DaFne-Microservice-Architektur ein.

Tranquillin et al. [36] lieferten mit ihrem Fachbuch eine umfassende Aufarbeitung etablierter Architekturprinzipien für Data- und ML-Plattformen, insbesondere im Cloud-Kontext. Im Projekt wurden daraus unter anderem Impulse für die Orchestrierung containerisierter Dienste in Kubernetes sowie für das technische Monitoring übernommen. Weitere Arbeiten, etwa von Pérez et al. [21], Karimi und Barfroush [12] oder Gao et al. [8], thematisieren dynamische, verteilte Plattformsätze und das Ressourcenmanagement für datenintensive KI-Systeme. Diese Konzepte wurden in DaFne insbesondere im Zusammenhang mit dem Management paralleler Generierungsprozesse und der Skalierung prototypischer Komponenten reflektiert.

Auch in Bezug auf Frontend-Architekturen und deren Modularisierung wurden Fortschritte erzielt, etwa durch die Etablierung von Micro-Frontend-Ansätzen (Taibi & Mezzalana [25]), deren Prinzipien sich auch in der UI-Strategie der DaFne-Plattform widerspiegeln.

Zusammenfassend zeigen diese Entwicklungen, dass die gewählten methodischen und technischen Ansätze im DaFne-Projekt im Einklang mit dem internationalen Forschungs- und Entwicklungsstand stehen. Die parallelen Fortschritte bei anderen Stellen wurden kontinuierlich beobachtet, evaluiert und - wo sinnvoll - in die eigene Entwicklung integriert, was die Anschlussfähigkeit und Aktualität der Projektergebnisse deutlich unterstreicht.

14 Veröffentlichungen

Im Rahmen des Projekts sind die folgenden Veröffentlichungen entstanden:

1. SAML (Software Architecture and Machine Learning) Konferenz als Teil der ECSA 2022 (European Conference on Software Architecture) in Prag (09/22): **“A Platform Providing Machine Learning Algorithms for Data Generation and Fusion - An Architectural Approach”**
2. SDF (Smart Digital Futures) Konferenz in Rom (06/23): **“DaFne: Data Fusion Generator and Synthetic Data Generation for Cities”**
3. ICICT 2024 (International Congress on Information and Communication Technology) hybrid/ in London (02/24): **“Evaluating Tabular Data Generation Techniques on the DaFne Platform: Insights from a Predictive Maintenance Case Study on Bridges”**

Literatur

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. p. 308–318. CCS '16, Association for Computing Machinery, New York, NY, USA (2016). <https://doi.org/10.1145/2976749.2978318>, <https://doi.org/10.1145/2976749.2978318>
2. Baddam, P., Glass, A., Jäkel, R., Jander, J., Krause, T., Kunert, P., Noennig, J.R., Okhrin, I., Sanchez, M., , Steffens, U.: Evaluating tabular data generation techniques on the dafne platform: Insights from a predictive maintenance case study on bridges. In: Yang, X., Sherratt, S., Dey, N., Joshi, A. (eds.) Proceedings of 9th International Congress on Information and Communication Technology - ICICT 2024, London, UK. Lecture Notes in Networks and Systems (02 2024)
3. Bryant, F.B., Yarnold, P.R.: Principal-components analysis and exploratory and confirmatory factor analysis. (1995), <https://api.semanticscholar.org/CorpusID:140743544>
4. Cao, C., Li, M.: Generating mobility trajectories with retained data utility. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. p. 2610–2620. KDD '21, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3447548.3467158>, <https://doi.org/10.1145/3447548.3467158>
5. Deutscher Wetterdienst: Deutscher wetterdienst: Api (2024), <https://dwd.api.bund.dev/>
6. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wehner, I. (eds.) Automata, Languages and Programming. pp. 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
7. Feng, J., Yang, Z., Xu, F., Yu, H., Wang, M., Li, Y.: Learning to simulate human mobility. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 3426–3433. KDD '20, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3394486.3412862>, <https://doi.org/10.1145/3394486.3412862>
8. Gao, B., Xia, X., Ji, Z., Zhang, S., Yan, Z., Li, Y.: Research and application of task scheduling and resource management technology in electric power artificial intelligence platform. In: 2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). pp. 945–951. IEEE (2022)
9. Giomi, M., Boenisch, F., Wehmeyer, C., Tasnádi, B.: A unified framework for quantifying privacy risk in synthetic data (2022), <https://arxiv.org/abs/2211.10459>
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (nov 1997). <https://doi.org/10.1162/neco.1997.9.8.1735>, <https://doi.org/10.1162/neco.1997.9.8.1735>
11. Jiang, W., Zhao, W.X., Wang, J., Jiang, J.: Continuous trajectory generation based on two-stage gan. Proceedings of the AAAI Conference on Artificial Intelligence **37**(4), 4374–4382 (Jun 2023). <https://doi.org/10.1609/aaai.v37i4.25557>, <https://ojs.aaai.org/index.php/AAAI/article/view/25557>
12. Karimi, M., Barfroush, A.A.: Proposing a dynamic executive microservices architecture model for ai systems. arXiv preprint arXiv:2308.05833 (2023)
13. Kingma, D.P., Welling, M.: Stochastic gradient VB and the variational auto-encoder. In: Second International Conference on Learning Representations, ICLR. vol. 19, p. 121 (2014)

14. Li, X., Metsis, V., Wang, H., Ngu, A.H.H.: Tts-gan: A transformer-based time-series generative adversarial network. In: International conference on artificial intelligence in medicine. pp. 133–143. Springer (2022)
15. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008), <http://www.jmlr.org/papers/v9/vandermaaten08a.html>
16. Markov, I.L., Apostolopoulos, P.A., Garrard, M.R., Qie, T., Huang, Y., Gupta, T., Li, A., Cardoso, C., Han, G., Maghsoudian, R., et al.: Scalable end-to-end ml platforms: from automl to self-serve. *arXiv preprint arXiv:2302.14139* (2023)
17. Markov, I.L., Wang, H., Kasturi, N.S., Singh, S., Garrard, M.R., Huang, Y., Yuen, S.W.C., Tran, S., Wang, Z., Glotov, I., et al.: Looper: An end-to-end ml platform for product decisions. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pp. 3513–3523 (2022)
18. Massey Jr, F.J.: The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association* **46**(253), 68–78 (1951)
19. Papernot, N., Abadi, M., Úlfar Erlingsson, Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data (2017), <https://arxiv.org/abs/1610.05755>
20. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. pp. 399–410. IEEE (2016)
21. Pérez, F.M., Fonseca, I.L., Martínez, J.V.B., Maciá-Fiteni, A.: Distributed architecture for an elderly accompaniment service based on iot devices, ai, and cloud services. *IEEE multimedia* **30**(1), 17–27 (2022)
22. Rao, J., Gao, S., Kang, Y., Huang, Q.: LSTM-TrajGAN: A Deep Learning Approach to Trajectory Privacy Protection. In: Janowicz, K., Versteegen, J.A. (eds.) *11th International Conference on Geographic Information Science (GIScience 2021) - Part I. Leibniz International Proceedings in Informatics (LIPIcs)*, vol. 177, pp. 12:1–12:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2020). <https://doi.org/10.4230/LIPIcs.GIScience.2021.I.12>, <https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.GIScience.2021.I.12>
23. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. p. II–1278–II–1286. ICML’14, JMLR.org (2014)
24. Sun, C., van Soest, J., Dumontier, M.: Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of biomedical informatics* p. 104404 (2023), <https://api.semanticscholar.org/CorpusID:259029626>
25. Taibi, D., Mezzalana, L.: Micro-frontends: Principles, implementations, and pitfalls. *ACM SIGSOFT Software Engineering Notes* **47**(4), 25–29 (2022)
26. The Synthetic Data Vault: Boundary Adherence Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/boundaryadherence>, accessed: 2024-07-19
27. The Synthetic Data Vault: Category Coverage Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/categorycoverage>, accessed: 2024-07-19
28. The Synthetic Data Vault: KSComplement - Kolmogorov-Smirnov Statistic Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/kscomplement>, accessed: 2024-07-19
29. The Synthetic Data Vault: Logistic Detection Metric. <https://sdv.dev/SDV/api-reference/metrics/api/sdv.metrics.tabular.LogisticDetection.html>, accessed: 2024-07-19

30. The Synthetic Data Vault: ML Efficacy Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-in-beta/ml-efficacy-single-table>, accessed: 2024-07-19
31. The Synthetic Data Vault: New Row Synthesis Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/newrowsynthesis>, accessed: 2024-07-19
32. The Synthetic Data Vault: Probabilistic AutoRegressive model. https://sdv.dev/SDV/user_guides/timeseries/par.html, accessed: 2024-07-16
33. The Synthetic Data Vault: Quality Report: Column Pair Trends. <https://docs.sdv.dev/sdmetrics/reports/quality-report/whats-included>, accessed: 2024-07-19
34. The Synthetic Data Vault: Range Coverage Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/rangecoverage>, accessed: 2024-07-19
35. The Synthetic Data Vault: TVComplement - Total Variation Distance Metric. <https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/tvcomplement>, accessed: 2024-07-19
36. Tranquillin, M., Lakshmanan, V., Tekiner, F.: Architecting data and machine learning platforms: enable analytics and AI-driven innovation in the cloud. Ö'Reilly Media, Inc.”(2023)
37. Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J.: Differentially private generative adversarial network (2018), <https://arxiv.org/abs/1802.06739>
38. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf>
39. Yoon, J., Jarrett, D., van der Schaar, M.: Time-series generative adversarial networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf
40. Zhao, Z., Kumar, A., Birke, R., Van der Scheer, H., Chen, L.Y.: Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in big Data* **6**, 1296508 (2024)