VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Projektgruppe MARVIN

Abstract

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Dipl. oec. Univ. Michael Mattern
	Dr.-Ing. Andreas Solsbach

Abgabetermin:	31. März 2024

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

Abkürzungsverzeichnis **4**

Glossar **5**

1 Einleitung **6**
1.1 Problemstellung 6
1.2 Zielsetzung 8

2 Umsetzung **9**
2.1 Datenmodell 9
2.2 Systemarchitektur 11
2.3 Workflow 12
2.4 DAG-Editor 13
2.5 Reporting 17

3 Ausblick auf MARVIN 2.0 **18**

Literaturverzeichnis **19**

Abbildungsverzeichnis

1 Datenbank Schema 10
2 Systemarchitektur 11
3 DAG-Editor 13

Abkürzungsverzeichnis

DAG Directed Acyclic Graph

MARVIN Management, Architectures, VUCA Worlds, Causal Inference

OOWV Oldenburgisch-Ostfriesischer Wasserverband

RCT Randomized Controlled Trial

SAC SAP Analytics Cloud

UUID Universally Unique Identifier

VUCA Volatility, Uncertainty, Complexity, Ambiguity

Glossar

Accuracy Kennzahl, die in diesem Kontext zur Beschreibung der DAG-Qualität verwendet wird. 15

Directed Acyclic Graph (DAG) Ein DAG ist eine grafische Darstellung von Beziehungen zwischen verschiedenen Variablen. Die Variablen werden dabei als sogenannte Knoten dargestellt und die Abhängigkeiten durch Pfeile, auch Kanten genannt, wobei Zyklen ausgeschlossen sind. Die Berechnung und Analyse anhand dieser Struktur ermöglicht eine klare Visualisierung von kausalen Zusammenhängen zwischen den verschiedenen Variablen eines Netzwerks. 8, 9, 11, 12, 14–17

Domäne Eine Domäne beschreibt in diesem Kontext die hochgeladenen Zeitreihen in der Datenbank, die letztendlich die Knoten zum Bauen oder Generieren eines DAGs im DAG-Editor bereitstellen. 8–10, 12, 13

Exposure Der Exposure-Knoten stellt die Variable dar, die dem Outcome-Knoten ausgesetzt wird. 14–16

Kanten Kanten sind in einem DAG die Pfeile, die die gerichteten Verbindungen zwischen den Knoten repräsentieren. 14

Kantenstärken Ein Gütemaß, das Auskunft über die Aussagekraft einer Kante zwischen zwei Knoten in einem DAG gibt. 15, 16

kausale Inferenz Prozess, bei dem aus beobachteten Daten Schlussfolgerungen über kausale Zusammenhänge gezogen werden. Dabei werden statistische Methoden angewendet, um Ursache-Wirkungs-Beziehungen zwischen Variablen zu identifizieren und zu bewerten. 8

Outcome Der Outcome-Knoten stellt die zu untersuchende Zielvariable dar. 14–16

1 Einleitung

Dieses Abstract der Projektdokumentation der Projektgruppe Management, Architectures, VUCA Worlds, Causal Inference (**MARVIN**) dient der Veranschaulichung der Problemstellung, der Zielsetzung und der Umsetzung. Detaillierte Ausführungen und Informationen zur Projektarbeit befinden sich in der Projektdokumentation. Zunächst wird die Problemstellung anhand von hypothetischen Szenarien aus dem Umfeld des Projektpartners, dem Oldenburgisch-Ostfriesischer Wasserverband (**OOWV**), erläutert. Darauf folgt die Zielsetzung des Projekts **MARVIN**. Im Umsetzungskapitel werden die Ergebnisse beleuchtet und zum Schluss wird ein Ausblick auf eine Mögliche Weiterentwicklung des Tools gegeben.

1.1 Problemstellung

Informationen über Kausalzusammenhänge sind für Entscheider auf allen Managementebenen von hohem Wert. Es ist entscheidend, potenzielle Ergebnisse alternativer Handlungen zu kennen oder einschätzen zu können, um operative, taktische oder strategische Maßnahmen unter bestimmten Rahmenbedingungen zielgerichtet durchführen zu können. Die übergeordnete Frage in diesem Kontext ist die Frage nach der Kausalität. Letztlich nehmen Manager die wissenschaftliche Perspektive ein, denn sie entwickeln und nutzen Hypothesen über Ursache-Wirkungs-Zusammenhänge, wodurch Theorien entstehen, die als Leitfaden dienen, um fundierte Erkenntnisse zu gewinnen, indem sie die Erkenntnismöglichkeiten einschränken (vgl. Felin & Zenger, 2017).

Im Rahmen der Projektarbeit **MARVIN** nimmt der **OOWV** eine bedeutende Position als Projektpartner ein. Die Thematik der Deutung und Nutzung von Daten gewinnt auch für den **OOWV** zunehmend an Bedeutung, denn als regionaler Wasserversorger liegt die maßgebliche Verantwortung des **OOWV** darin, eine zuverlässige und effiziente Wasserversorgung für die Bevölkerung zu gewährleisten. Dabei spielen sowohl operative als auch strategische Entscheidungen eine entscheidende Rolle. Auf operativer Ebene erlangen präzise Analysen von Betriebsdaten fundamentale Bedeutung. Der **OOWV** muss die Leistung seiner Wasserversorgungseinrichtungen überwachen, um mögliche Engpässe, Lecks oder Qualitätsprobleme frühzeitig zu erkennen und effektive Maßnahmen zur Behebung zu ergreifen. Durch die Analyse von Betriebsdaten kann der **OOWV** operative Abläufe optimieren und sicherstellen, dass die Wasserversorgung den höchsten Qualitätsstandards entspricht [1].

[1] https://www.oowv.de/home

Um Wissen aus Unternehmensdaten zu ziehen, könnten Korrelationsbeziehungen herangezogen werden. Korrelationen beschreiben statistische Zusammenhänge zwischen zwei oder mehr Variablen. Eine positive Korrelation bedeutet, dass sich die Variablen tendenziell gemeinsam in eine bestimmte Richtung bewegen, während eine negative Korrelation darauf hinweist, dass sich die Variablen in entgegengesetzte Richtungen bewegen. Nur weil zwei Variablen miteinander korrelieren, bedeutet dies nicht zwangsläufig, dass eine Ursache-Wirkungs-Beziehung besteht. Andere Faktoren könnten im Spiel sein, die die beobachtete Korrelation erklären. Hier kommen Randomized Controlled Trial (**RCT**)s zum Einsatz. Dies sind kontrollierte Studien mit zufälliger Zuweisung von Studienteilnehmenden in Kontroll- und Interventionsgruppen. **RCT**s sind darauf ausgerichtet, Kausalität aufzudecken und die Zufälligkeit der Zuweisung minimiert mögliche Störfaktoren und sorgt für eine Vergleichbarkeit zwischen den Gruppen. Dieser Ansatz ermöglicht es, die Auswirkungen einer Intervention auf die Zielvariablen zu isolieren und die Ergebnisse auf kausale Zusammenhänge zu prüfen. Trotz der methodischen Stärken von **RCT**s stoßen sie auf ethische und praktische Grenzen. In einigen Fällen ist es nicht möglich oder ethisch vertretbar, Menschen oder Gruppen zufällig bestimmten Bedingungen auszusetzen. In anderen Situationen ist es außerdem praktisch nicht möglich, ein **RCT** durchzuführen (vgl. Deaton & Cartwright, 2018).

Die Motivation hinter der Projektgruppe **MARVIN** ist eng mit den vier zentralen Aspekten verbunden, für die der Name **MARVIN** steht: Management Control, Architectures, Volatility, Uncertainty, Complexity, Ambiguity (**VUCA**) Worlds und Causal Inference. Die Erforschung kausaler Zusammenhänge mithilfe von kausaler Inferenz ist besonders relevant für die Entscheidungsfindung innerhalb von Unternehmen, insbesondere in einer **VUCA** World - also einer durch Volatilität, Unsicherheit, Komplexität und Ambiguität geprägten Welt (vgl. Amann et al., 2019) - von großer Relevanz. Eine intensive Auseinandersetzung mit diesen Zusammenhängen ermöglicht es Unternehmen nicht nur, Verbindungen zwischen Ereignissen festzustellen, sondern auch ein tiefgreifendes Verständnis der zugrunde liegenden Mechanismen zu entwickeln. Dies ist von entscheidender Bedeutung für das strategische Management, die operative Kontrolle und die Gestaltung effizienter Architekturen zur Unterstützung geschäftlicher Abläufe (vgl. Pearl, 2010).

1.2 Zielsetzung

Die Zielsetzung der Projektgruppe **MARVIN** leitet sich aus den zuvor thematisierten Problemstellungen der Korrelationen und **RCT**s ab. Das Ziel ist es demnach, ein Tool zur Analyse von Kausaleffekten zu entwickeln. Zentraler Bestandteil dieses Tools soll der sogenannte DAG-Editor sein, ein Werkzeug, das es ermöglicht, Directed Acyclic Graph (DAG)s zu erstellen, zu modifizieren und zu analysieren. Dabei soll die Übertragbarkeit des Schemas auf verschiedene Thematiken sowie der Zugriff auf unterschiedliche Datenquellen möglich sein. Im Kontext dieses Projekts wird die kausale Inferenz als bevorzugte Methode gewählt, um tiefgreifende Einblicke in die zugrunde liegenden Ursache-Wirkungs-Beziehungen zu erhalten. Im Vergleich zu anderen Methoden bietet die kausale Inferenz den entscheidenden Vorteil, tatsächliche Ursache-Wirkungs-Beziehungen zwischen den Variablen zu identifizieren und Konfundierungsvariablen zu berücksichtigen. Des Weiteren sollen die Features von dagitty und dem causaleffects R Paket integriert werden, wobei die Funktionen des causaleffect Packages genutzt und erweitert werden sollen und mit einer von Dagitty inspirierten Benutzeroberfläche für Nutzer zugänglich gemacht werden sollen. Dies soll mit der Möglichkeit, Datenquellen zu hinterlegen und in die Analysen zu integrieren, erweitert werden. Dazu soll mit bereitgestellten Ticketdaten des **OOWV** gearbeitet werden. Trotz dessen soll das **MARVIN** Tool nicht nur spezifisch den Anforderungen des **OOWV** gerecht werden, sondern auch universell einsetzbar sein. Die Implementierung soll daher mit dem Fokus auf eine flexible, skalierbare und erweiterbare Struktur erfolgen, die Anpassungen an verschiedene Anwendungsfälle ermöglicht. Ein Aspekt der Erweiterung dieser Architektur ist die Integration der SAP Analytics Cloud (**SAC**). Die **SAC** fungiert als Plattform für vorgelagerte Analysen, indem es Möglichkeiten zur Betrachtung und Visualisierung von Daten bereitstellt, bevor die Analysen im DAG-Editor durchgeführt werden. Durch die Einbindung der **SAC** können Benutzer vorab Knoten und Variablen in verschiedenen Diagrammen betrachten, Korrelationen erkennen und somit eine präzisere Vorauswahl für die spätere Auswahl im DAG-Editor treffen. Diese Integration analytischer Funktionen soll zur Effizienz und Genauigkeit der Analysen beitragen und soll eine Vorabprüfung von potenziellen kausalen Beziehungen innerhalb des jeweiligen Netzwerks ermöglichen, um gegebenenfalls auch ohne spezifisches Domänenwissen aufschlussreiche Directed Acyclic Graph (DAG)s bauen zu können oder vorhandenes Domänenwissen vorab zu prüfen.

2 Umsetzung

In diesem Kapitel sollen zuerst das Datenmodell und die Systemarchitektur des DAG-Editors erklärt werden. Danach soll mit dem Workflow der Umgang mit allen Komponenten des **MARVIN**-Tools beleuchtet werden. Anschließend werden die Funktionen innerhalb des DAG-Editors erklärt. Des Weiteren wird ein Einblick in das Reporting gegeben.

2.1 Datenmodell

Das Datenmodell ist ein zentraler Teil der Architektur und ausschlaggebend für die Erreichung einiger Anforderungen. Bspw. ist die Anwendbarkeit oder Übertragbarkeit des Tools auf andere Anwendungsfälle und andere Daten nur möglich, wenn das Datenmodell dies auch zulässt. Es soll dabei möglichst flexibel sein, ohne dabei andere Aspekte wie die Leistungsfähigkeit, Effizienz, Sicherheit oder Konsistenz zu vernachlässigen.

Das Datenmodell besteht dabei aus zwei Arten von Daten: Daten, die im direkten Zusammenhang mit der technischen Benutzung der App stehen und Daten, die die Zeitreihen betreffen. Also die Daten, die es zu analysieren gilt. Daten, die im direkten Zusammenhang mit der technischen Benutzung der App stehen, wie beispielsweise Nutzerdaten oder Daten bezüglich gespeicherter Directed Acyclic Graph (DAG)s, werden in einem relationalem Modell gespeichert. Über Fremdschlüsselbeziehungen werden Zusammenhänge zwischen den Tabellen dargestellt. So werden bspw. Nutzerberechtigungen oder Beziehungen innerhalb eines Directed Acyclic Graph (DAG)s über Verknüpfungstabellen abgebildet. Dadurch ist es dann möglich, Benutzer zu einer Benutzergruppe hinzuzufügen, welche wiederum die freizuschaltenden Domänen beinhalten. Technisch wurde dieser Teil des Datenmodells auf einer PostgreSQL-Datenbank umgesetzt, in der Abbildung 1 sind die genannten Tabellen gelb markiert. Die in Abbildung 1 blau markierten Tabellen bilden den inhaltlichen Kern des Datenmodells und sind technisch auf einer SAP HANA-Datenbank umgesetzt. Es handelt sich hierbei um eine „holonische“ Datenstruktur. Der Begriff „Holon“ stammt dabei aus dem griechischen und bedeutet „das Teil eines Ganzen seiend“ und wurde von Arthur Koestler in „The Roots of Coincidence“ geprägt (vgl. Koestler, 1973). Abstrahiert auf Informationssysteme bedeutet „holonisch“, dass einzelne Teile des Systems oder der Daten eigenständig und in sich geschlossen sind, allerdings weitergehend auch Teil etwas Übergeordnetem sind. Im Kontext der Kausalanalyse trifft dies auf die einzelnen Zeitreihen zu, welche in sich geschlossen und vollständig sind und auch alleinstehend für Analysen nutzbar sind. Das Übergeordnete in diesem Fall ist dann die Zuordnung der Zeitreihen in Domänen und letztendlich die Modellierung in Directed Acyclic Graph (DAG)s. Domänen

beinhalten in diesem Kontext dann inhaltlich zusammengehörige Zeitreihen (bspw. würde die Domäne „Finanzdaten“ Zeitreihen bezüglich Aktienkurse oder Wechselkurse beinhalten, wohingegen die Domäne „Tickets“ Zeitreihen bezüglich Ticketlaufzeiten oder Bearbeiterwechsel beinhalteten würde).

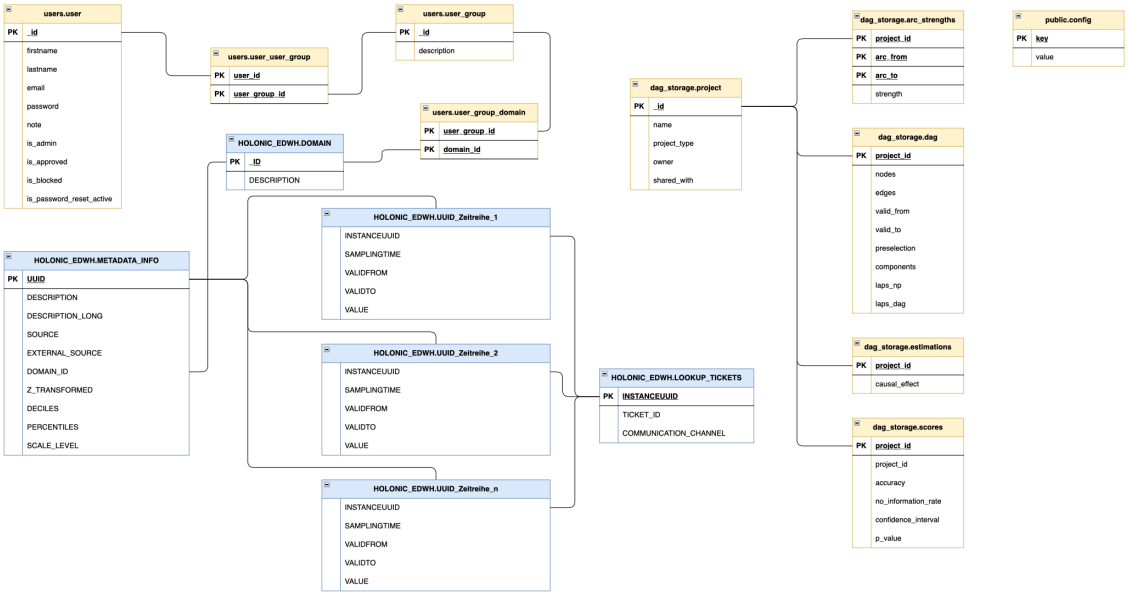

Abbildung 1: Datenbank Schema

Möchte man nun bspw. Kausalanalysen im Bereich des Finanzwesens durchführen, müsste man Aktienkurse, Wechselkurse, Anleihen, etc. als Zeitreihen in das Datenmodell laden, um darauf aufbauend Analysen durchzuführen. Die einzelnen Kurse/Zeitreihen würden dann jeweils mit Universally Unique Identifier (**UUID**)s als Tabellenname gespeichert. Die dazugehörigen Einträge in der metadata_info beinhalteten weitere Informationen, wie bspw. die Bezeichnung der Aktie und die Art des Kurses (Schluss- oder Eröffnungs-Kurs, Handelsvolumen, etc.). Über die Domänenzuordnung der Zeitreihen wäre es möglich, Benutzergruppen zu erstellen. Zu diesen Benutzergruppen könnte man Benutzer hinzufügen und diese hätten die Möglichkeit, die Finanzdaten-Zeitreihen in Form von Knoten im DAG-Editor zu nutzen. Durch die Entkopplung von Datenmodell und Berechnungslogik und die Verknüpfung von Zeitreihen und dem restlichen Datenmodell über Domänen und Benutzergruppen ist es möglich, flexible weitere oder andere Anwendungsfälle zu integrieren. Grundvoraussetzung dafür ist, dass die zu analysierenden Daten im Zeitreihenformat vorliegen.

2.2 Systemarchitektur

Die Systemarchitektur wird im Folgenden anhand einer Abbildung und eines textuell skizzierten Szenarios beschrieben.

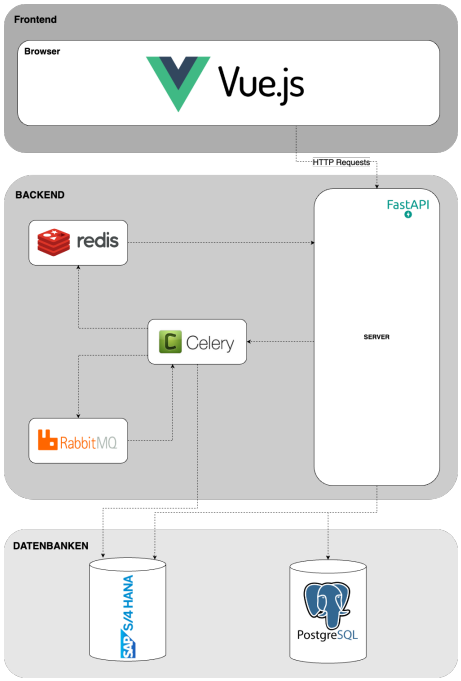

Abbildung 2: Systemarchitektur

In Abbildung 2 wurde die Systemarchitektur in Form eines Diagramms bildlich dargestellt. Interagiert ein Benutzer nun mit der Benutzeroberfläche der Anwendung, werden entsprechende Anfragen ausgelöst und Prozesse im Backend angestoßen. Die Benutzeroberfläche sendet eine HTTP-Anfrage an den FastAPI-Server, der die API-Route für die entsprechende Anfrage aufruft. Der FastAPI-Server empfängt die Anfrage und validiert sie. Sollte die vom Nutzer gesendete Anfrage eine Lesende-Abfrage (bspw. das Laden der Nutzerdaten, Projekte oder Projektinhalte) auf die Nutzerdankenbank sein, greift der Server nur auf die Nutzerdatenbank zu und liefert die Ergebnisse an die Benutzeroberfläche zurück, welche dort entsprechend dargestellt werden. Handelt es sich um eine Analyseanfrage, wie bspw. das Anstoßen einer Directed Acyclic Graph (DAG) Generierung oder einer Directed Acyclic Graph (DAG) Analyse, wird die Anfrage anschließend an die Celery-Warteschlange weitergeleitet, um die Anfrage in einer Queue einzureihen. Dem Benutzer wird das Feedback gegeben, dass die entsprechende Anfrage angestoßen wurde. Ein freier Celery-Worker wählt die in der Warteschlange liegende Task aus und beginnt mit der Ausführung. Dies

umfasst die Ausführung von R-Algorithmen, die in der Berechnung der Anfrage involviert sind. Der Celery-Worker führt die erforderlichen Berechnungen durch, um die angeforderten Ergebnisse zu generieren. Nach Abschluss der Berechnungen veröffentlicht der Celery-Worker die Ergebnisse an den Redis-Publisher. Der FastAPI-Server empfängt die Ergebnisse über den Redis-Subscriber, da er über eine Subscriber-Verbindung verfügt, die auf neue Ereignisse lauscht und speichert diese für den Nutzer bereits backend-seitig ab. Die berechneten Ergebnisse werden vom FastAPI-Server an das Frontend über die bestehende WebSocket-Verbindung gesendet. Die Nutzeroberfläche empfängt die Ergebnisse entsprechend und stellt sie dem Benutzer dar, sofern dieser noch eingeloggt ist.

2.3 Workflow

Die erfolgreiche Anwendung des **MARVIN** Tools im Rahmen von Entscheidungsprozessen erfordert einen durchdachten Workflow, der das **MARVIN** Tool selbst, bestehend aus dem DAG-Editor und der **SAC**, integriert. Die **SAC** fungiert als eine Plattform für vorbereitende Analysen, noch bevor die eigentlichen Kausalanalysen im DAG-Editor durchgeführt werden. Innerhalb der **SAC** können Nutzer die zuvor beschriebenen explorativen Analysen nutzen, um relevante Datenmuster und Korrelationen zu identifizieren. Diese Analysemöglichkeiten dienen als Grundlage für die vorläufige Auswahl von Knoten, die dann gezielt im weiteren Prozess genutzt werden können. Durch die visuelle Darstellung der Diagramme und Analysen in der **SAC** wird eine solide Basis für den nachfolgenden Einsatz im DAG-Editor geschaffen. Die Integration der **SAC** erleichtert Domänenexperten nicht nur die Validierung vorhandenen Wissens und die Entwicklung von Hypothesen, sondern ermöglicht auch eine Analyse ohne tiefgehendes Domänenwissen. Auf Grundlage der Erkenntnisse aus der **SAC** kann im Anschluss eine gezielte Auswahl von Knoten für den DAG-Editor getroffen werden.

Nach der präzisen Vorauswahl von Knoten in der **SAC** können diese entweder manuell oder automatisch in den DAG-Editor eingebaut werden. Die manuelle Integration eignet sich besonders für den Nachbau und die Analyse von in der **SAC** aufgestellten Hypothesen. Die automatische Integration ist hilfreich, wenn das Domänenwissen begrenzt ist oder wenn Nutzer Zusammenhänge analysieren wollen, an die sie zuvor noch nicht gedacht haben. Nachdem ein Directed Acyclic Graph (DAG) konstruiert wurde, ermöglicht der Workflow die Analyse und den Vergleich kausaler Zusammenhänge mithilfe der bereits erwähnten kausalen Effekten und Gütemaßen.

2.4 DAG-Editor

Der DAG-Editor ist ähnlich wie herkömmliche Webanwendungen aufgebaut und beinhaltet somit eine Benutzeranmeldung und damit verbundene Aspekte, wie die Authentifizierung über die gesamte Anwendung hinweg und das Verwalten der eigenen Benutzerdaten. Benutzern mit dem Administratorenstatus ist es zudem möglich eine Benutzer- und Benutzergruppenverwaltung über das sogenannte Adminpanel vorzunehmen. Die Rollenverteilung innerhalb der Anwendung sieht die Rollen des Administrators und des normalen Nutzers vor, wobei sich der Administrator einzig durch den Zugang zum Adminpanel und dem Verwalten von Nutzern vom normalen Nutzer unterscheidet. Die Benutzergruppen beinhalten Domänen und sind dazu da, um Nutzern über die Benutzergruppenzuweisung einen Domänenzugriff zu vergeben.

Nach dem Klicken eines Projektnamens in der Projektübersicht wird der Nutzer in den eigentlichen Editor des DAG-Editors weitergeleitet. Dieser besteht aus der Knotenübersicht und dem Bearbeitungsbereich.

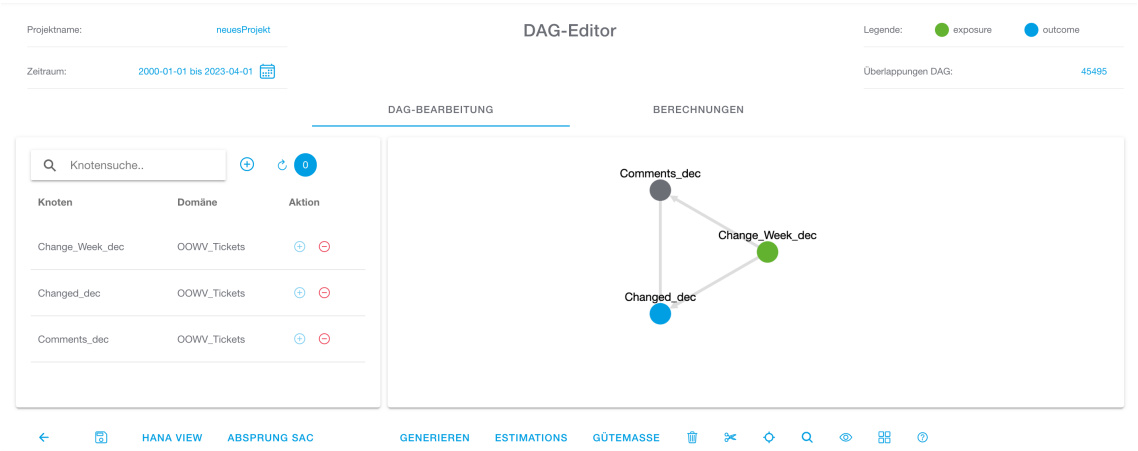

Abbildung 3: DAG-Editor

In der Knotenübersicht kann der Benutzer nun die Knoten mit denen er arbeiten will hinzufügen. Dazu wählt er zunächst die Domäne aus und dann die jeweiligen Knoten aus der entsprechenden Domäne. Es ist möglich Knoten aus mehreren Domänen zur Knotenübersicht hinzuzufügen. In der Knotenübersicht wird neben den ausgewählten Knoten auch die Anzahl der Überlappungen innerhalb der Knotenliste ausgegeben, also die Anzahl der sich überschneidenden Zeitreihen. Zusätzlich dazu kann über den Kalender Button der Gültigkeitszeitraum angepasst werden, um festzulegen in welcher Zeitspanne die Daten

analysiert werden sollen. Ab diesem Punkt kann der Benutzer auf zwei verschiedene Arten vorgehen: Er kann sich entweder einen Directed Acyclic Graph (DAG) aus den ausgewählten Knoten generieren lassen oder selber einen Directed Acyclic Graph (DAG) aus diesen Knoten bauen.

Mithilfe des „Generieren“ Buttons hat der Benutzer die Möglichkeit einen Directed Acyclic Graph (DAG) mittels verschiedener Algorithmen, basierend auf seiner ausgewählten Knotenliste generieren zu lassen. Nach Auswahl eines Algorithmus kann der Benutzer optional White- und Blacklists erstellen, um Verbindungen zwischen zwei Knoten obligatorisch zu machen (Whitelist) oder zu verbieten (Blacklist). Anschließend wird der Directed Acyclic Graph (DAG) generiert und im Bearbeitungsbereich visuell ausgegeben. Der generierte Directed Acyclic Graph (DAG) enthält noch keine Exposure und Outcome Knoten, diese können manuell markiert werden. Oberhalb des Bearbeitungsbereichs wird gegebenenfalls die Komponentenauswahl sichtbar, wenn aus der festgelegten Knotenliste mehr als ein Directed Acyclic Graph (DAG) generiert worden ist. In der Komponentenauswahl können die verschiedenen Directed Acyclic Graph (DAG)-Varianten betrachtet werden und sich über den „Behalten“ Button für eine Komponente entschieden werden, die im weiteren Verlauf analysiert und manuell weiter bearbeitet werden kann.

Für das manuelle Bauen eines Directed Acyclic Graph (DAG)s greift der Benutzer ebenfalls auf die zuvor definierte Knotenliste zu. Sobald sich mehr als ein Knoten im Directed Acyclic Graph (DAG) befindet können Kanten gezogen werden, indem ein Knoten durch einen Klick und eine hellblaue Umrandung des Knotens markiert wird. Anschließend erscheint bei gedrückter option Taste (Mac) und gedrückter Shift oder Alt Taste (Windows) ein Pfeil sobald der Cursor vom Knoten wegbewegt wird. Mit einem Klick auf den Zielknoten wird dadurch eine Kante zwischen den beiden Knoten gezogen. Durch das Klicken eines Knotens und gleichzeitiges gedrückt halten der command Taste (Mac) und der Strg Taste (Windows) kann ein Knoten mittels der sich ändernden Farbe als Exposure oder Outcome markiert werden. Die farbliche Zuordnung ist der Legende im DAG-Editor zu entnehmen.

Der generierte oder gebaute Directed Acyclic Graph (DAG) lässt sich nun manuell, durch das Ziehen eines Knotens mit gedrückter linker Maustaste, in eine gewünschte Position ziehen. Alternativ kann die Layout Funktion genutzt werden, um den Directed Acyclic Graph (DAG) automatisch in verschiedenen Layouts ausrichten zu lassen. Dies geschieht mithilfe des Layout Buttons unterhalb des Bearbeitungsbereichs.

Sobald sich ein vollständiger und korrekter Directed Acyclic Graph (DAG) mit einem Exposure und genau einem Outcome Knoten im Bearbeitungsbereich befindet ist es möglich die Gütemaße berechnen zu lassen. Um die Vollständigkeit und Korrektheit des Directed Acyclic Graph (DAG)s zu prüfen gibt es verschiedene Validierungen. Zum einen wird im Bearbeitungsbereich beim Hinzufügen jeder neuen Kante auf das Entstehen eines Zyklus geprüft und darauf hingewiesen, falls ein Zyklus vorliegt. Des weiteren wird beim Klicken des Gütemaß Buttons und des Estimation Buttons geprüft, ob der Directed Acyclic Graph (DAG) aus mehr als einer Komponente besteht und darauf hingewiesen, falls mehrere Komponenten vorliegen. Sowohl festgestellte Zyklen, als auch gefundene Komponenten werden dem Benutzer farblich markiert. Zudem wird der Directed Acyclic Graph (DAG) auf Exposure und Outcome Knoten geprüft, da mindestens ein Exposure Knoten und genau ein Outcome Knoten erforderlich sind, um Gütemaße zu berechnen. Zudem muss ein Pfad vom Exposure zum Outcome Knoten bestehen. Das Klicken des Gütemaß Buttons ist dem Benutzer erst möglich, wenn diese Kriterien erfüllt sind. Die Gütemaße beinhalten eine Berechnung der Kantenstärken für jede Kante im jeweiligen Directed Acyclic Graph (DAG), sowie die Berechnung der Accuracy, des Konfidenzintervalls, der No Information Rate und des P-Werts. Diese Gütemaße können im entstehenden „Berechnungen“ Tab eingesehen werden. Diese Gütemaße können folgender maßen interpretiert werden:

- **Accuracy (Genauigkeit)**: Die Accuracy gibt an, wie gut der erstellte Directed Acyclic Graph (DAG) die Daten abbildet. Sie wird in Dezimalform angegeben und repräsentiert den Anteil der korrekten Vorhersagen. Je näher der Wert an 1 liegt, desto besser ist die Übereinstimmung des Directed Acyclic Graph (DAG)s mit den tatsächlichen Daten (vgl. Yin et al., 2019).

- **95% Konfidenzintervall (CI)**: Das Konfidenzintervall gibt an, wie sicher die Genauigkeit des Directed Acyclic Graph (DAG)s ist. Es ist ein Intervall, innerhalb dessen mit einer Wahrscheinlichkeit von 95% erwartet werden kann, dass die wahre Genauigkeit liegt. Ein kleineres Intervall deutet auf eine präzisere Schätzung hin (vgl. Hazra, 2017).

- **No Information Rate (NIR)**: No Information Rate (NIR): Die No Information Rate gibt die Genauigkeit an, die erreicht wird, wenn der Directed Acyclic Graph (DAG) einfach die häufigste Klasse für jede Vorhersage wählt, unabhängig von den Eingabevariablen. Ein niedrigerer Wert im Vergleich zur Accuracy deutet auf eine

bessere Leistung des Directed Acyclic Graph (DAG)s hin [2]

- **P-Wert [Acc >NIR]**: Der P-Wert gibt an, ob die beobachtete Genauigkeit des Directed Acyclic Graph (DAG)s signifikant von der No Information Rate abweicht. Ein niedriger P-Wert deutet darauf hin, dass die beobachtete Genauigkeit statistisch signifikant ist. Ein P-Wert unter einem bestimmten Signifikanzniveau (häufig 0,05) deutet darauf hin, dass die Genauigkeit des Directed Acyclic Graph (DAG)s nicht zufällig ist (vgl. Hazra, 2017).

- **Kantenstärken**: Die Kantenstärken im Directed Acyclic Graph (DAG) zeigen die Stärke der Beziehung zwischen den verschiedenen Knoten an. Eine Kantenstärke nahe bei 1 deutet auf eine starke Beziehung zwischen den beteiligten Knoten hin. Ein Wert nahe 1 zeigt an, dass Änderungen an dem Ursprungsknoten einen signifikanten und direkten Einfluss auf den Zielknoten haben. Eine Kantenstärke nahe bei 0 zeigt eine schwache Beziehung zwischen den Knoten an. Dies bedeutet, dass Änderungen am Ursprungsknoten nur einen geringen oder keinen Einfluss auf den Zielknoten haben [3].

Nachdem Gütemaße für den jeweiligen Directed Acyclic Graph (DAG) berechnet wurden, ist das Durchführen von Estimations möglich. Sollte der Directed Acyclic Graph (DAG) nach der Berechnung der Gütemaße bearbeitet worden sein so müssen zuerst neue Gütemaße berechnet werden, bevor Estimations durchgeführt werden können. Um die Estimations anstoßen zu können muss zunächst, falls vorhanden, ein Adjustment Set für den Directed Acyclic Graph (DAG) ausgewählt werden, um Knoten, die Störfaktoren für den kausalen Effekt sein könnten zu behandeln. Daraufhin müssen für einige ausgewählte Knoten die gewünschten Ausprägungen ausgewählt werden. Nach der erfolgreichen Berechnung werden die Estimation Ergebnisse im „Berechnungen“ Tab sichtbar. Dazu gehört der numerische kausale Effekt für jede mögliche Ausprägung des Exposure Knotens, sowie die Summe der Prozentzahlen. Zusätzlich wird die Adjustments Formel ausgegeben. Der numerische kausale Effekt kann wie folgt interpretiert werden:

- **Numerischer Kausaler Effekt**: Der numerische kausale Effekt für jede mögliche Ausprägung des Outcome Knotens gibt an, wie stark die Veränderung der Exposure Knoten den Outcome Knoten beeinflusst. Ein hoher numerischer kausaler Effekt deutet darauf hin, dass das Exposure einen signifikanten Einfluss auf den Outcome

[2] https://statisticallearning.org/binary-classification.html
[3] https://www.bnlearn.com/documentation/man/arc.strength.html

hat, während ein niedriger Effektwert auf eine geringeren kausalen Zusammenhang hindeutet [4].

Mittels des „Absprung SAC“ Buttons kann eine HANA View des aktuellen Knotenliste gespeichert werden, um in der **SAC** abgerufen zu werden. Dort können verschiedene Analysen auf Zusammenhänge zwischen den Knoten des Directed Acyclic Graph (DAG)s durchgeführt werden, woraufhin der Directed Acyclic Graph (DAG) gegebenenfalls im DAG-Editor angepasst werden kann. Nach Klicken des Buttons wird der Nutzer zum Login Screen der **SAC** weitergeleitet, wo er z.B. die HANA View behandeln kann. Die Prozesse in der **SAC** werden im nachfolgenden Unterkapitel ausführlicher beschrieben.

2.5 Reporting

Der primäre Zweck des Reportings ist es, dem Anwender eine explorative Datenanalyse zu ermöglichen. Das bedeutet, dass der Nutzer durch die Verwendung verschiedener Diagramme Einblicke in Daten und mögliche Zusammenhänge erhält. Mit den Erkenntnissen, die aus den Visualisierungen in diesem Bereich gewonnen werden, lässt sich das Datenverständnis des Anwenders erweitern, wodurch die Nutzung des DAG-Editors erleichtert wird. Dabei sind mögliche Unterschiede in den Anwendungsfällen oder in der konkreten Datengrundlage – wie auch beim DAG-Editor – irrelevant, entscheidend ist nur, dass die Daten in derselben Struktur vorliegen. Um die Variabilität zu demonstrieren, wurden verschiedene Tools zur Visualisierung von Daten verwendet, wobei Power BI den größten Anteil umfasst. Neben dem Business Intelligence Tool aus dem Hause Microsoft wurde außerdem **SAC** genutzt. Mögliche Alternativen zu dieser Auswahl sind zum Beispiel Qlik oder Tableau. In den verwendeten Lösungen ist es zudem möglich, verschiedene Datenbanken, wie bspw. die verwendete HANA-Datenbank, als Datenquelle zu nutzen. So kann im Reporting auf dieselbe Datenquelle wie auch im DAG-Editor zugegriffen werden. Aus den geladenen Daten werden im Reporting verschiedene Diagramme erstellt, die durch unterschiedliche Analysen Einblicke in große Datenmengen ermöglichen. Zum Teil ist es möglich, über Filter mit den Diagrammen zu interagieren. Folgende Diagramme werden angeboten: Scatterplot, Korrelationsmatrix, Scatterplot Matrix, Random Forest, Boxplot, Balkendiagramm und Liniendiagramm.

[4] https://cran.r-project.org/web/packages/causaleffect/vignettes/causaleffect.pdf

3 Ausblick auf MARVIN 2.0

Im Verlauf des Projekts hat sich durch die Identifikation von weiteren Potenzialen herausgestellt, dass eine Weiterentwicklung von **MARVIN** durchaus interessant sein könnte. Vor allem eine Erweiterung des Join-Mechanismus der Zeitreihen in der HANA Datenbank würde zu einer erheblichen Aufwertung des Tools führen. Aus Zeitgründen wurde sich dazu entschieden, starr nur eine Art des Joins zu implementieren. In einer zweiten Version wäre es sicherlich sinnvoll, den Join-Mechanismus dahingehend zu erweitern, dass er flexibel auf die Art, wie eine Instanz zu einer Entität zusammengeführt werden kann, reagieren würde. So wäre es möglich, die gegebenen Ticketdaten beispielsweise mit Daten über die Anzahl verfügbarer Agenten zu erweitern und sowohl Daten bezogen auf das Ticket, als auch zeitliche Daten bezogen auf die Agentenanzahl zu analysieren. Des Weiteren könnten weitere Komfort-Funktionen in einer zweiten Version ergänzt werden, welche im Tool oder außerhalb des Tools das Arbeiten mit **MARVIN** erleichtern würden. Denkbar wären Funktionen zum Integrieren neuer Zeitreihen, eine Funktion zum Hochladen und Nutzen von Mail-Vorlagen für automatisiert generierte Mails oder eine vollständige Integration des vorgelagerten Datenanalyse-Prozesses im **SAC** oder anderweitigen BI-Lösungen.
Auch technisch gesehen könnte eine zweite **MARVIN**-Version einige Verbesserungen mit sich bringen. Eine strikte Typisierung der Variablen im Backend und Frontend, die konsequente Strukturierung von Komponenten, bspw. die DAG-Validierungen, in einzelne Services und die Verwendung von ausschließlich funktionaler Programmierung im Frontend wären sicherlich die ersten Punkte, die angegangen werden sollten, um Wartbarkeit, Skalierbarkeit und Code Qualität der Anwendung weiterführend zu optimieren. Aus Zeitgründen konnten diese Punkte nicht in der Laufzeit der Projektarbeit umgesetzt werden. Eine Weiterentwicklung und ein Aufsetzten auf die aktuelle Lösung ist also durchaus interessant und würde sowohl inhaltlich als auch technisch Fortschritte mit sich bringen.

Literaturverzeichnis

Felin, T., & Zenger, T. R. (2017). The Theory-Based View: Economic Actors as Theorists. *Strategy Science*, *2*(4), 258–271. https://doi.org/10.1287/stsc.2017.0048

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Amann, E. G., Balaban, S., Braak, J., Drath, K., Elle, K., Fahnenbruck, G., Gallenmüller, N., Gattinger, A., Golenhofen, P., Herold, A., Hofmann, B., Hofmann, O., & Huemer, B. (2019). *Resilienz für die VUCA-Welt: Individuelle und organisationale Resilienz entwickeln* (J. Heller, Ed.). Springer.

Pearl, J. Causal Inference. en. In: In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*. PMLR, 2010, February, 39–58. Retrieved March 17, 2024, from https://proceedings.mlr.press/v6/pearl10a.html

Koestler, A. (1973). *The roots of coincidence*. Vintage Books.

Yin, M., Wortman Vaughan, J., & Wallach, H. Understanding the Effect of Accuracy on Trust in Machine Learning Models. en. In: In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, 2019, May, 1–12. ISBN: 9781450359702. https://doi.org/10.1145/3290605.3300509

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*, *9*(10), 4124–4129. https://doi.org/10.21037/jtd.2017.09.14