TAL TECH

DOCTORAL THESIS

Measures of Impact and Confounding – An Analysis and Experimental Comparison of Novel and Established Measures

Sijo Arakkal Peious

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2025

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
24/2025

Measures of Impact and Confounding – An Analysis and Experimental Comparison of Novel and Established Measures

SIJO ARAKKAL PEIOUS

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science

The dissertation was accepted for the defence of the degree of Doctor of Philosophy (Computer Science) on 07 April 2025

Supervisor: Prof. Dr.rer.nat.habil. Dirk Draheim,
Information Systems Group,
Department of Software Science
School of Information Technologies
Tallinn University of Technology
Tallinn, Estonia

Opponents: Dr. Divesh Srivastava,
Head of Database Research,
AT&T Labs,
New Jersey, USA

Professor Arun Kumar Sangaiah, PhD
Distinguished Professor and Yushan Young Scholar,
International Graduate School of Artificial of Intelligence,
National Yunlin University of Science and Technology,
Yunlin, Taiwan

Defence of the thesis: 09 May 2025, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Sijo Arakkal Peious

signature

ISSN 2585–6898 (publication)
ISBN 978–9916–80–289–2 (publication)
ISSN 2585–6901 (PDF)
ISBN 978–9916–80–290–8 (PDF)
DOI https://doi.org/10.23658/taltech.24/2025

Arakkal Peious, S. (2025). *Measures of Impact and Confounding - An Analysis and Experimental Comparison of Novel and Established Measures* [TalTech Press]. https://doi.org/10.23658/taltech.24/2025

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
24/2025

Mõju ja segajate mõõdikud – uute ja väljakujunenud meetmete analüüs ja eksperimentaalne võrdlus

SIJO ARAKKAL PEIOUS

TAL TECH
KIRJASTUS

Contents

List of Publications

This Ph.D. thesis is written as a dissertation to which the following publications are appended and referred to in the text by Roman numbers.

I S. Arakkal Peious, M. Kaushik, M. Shahin, R. Sharma, and D. Draheim. On choosing the columnar in-memory database Hyrise as high-performant implementation platform for the GrandReport tool. In *Proceedings of NGISE'2025 – the 1st International Conference in Next Generation Information Systems Engineering*, pages 1–7. IEEE, 2025. to appear, preprint available at IEEE TechRxiv, doi: 10.36227/techrxiv.174015861.15410981/v1

II S. Arakkal Peious, M. Kaushik, S. A. Shah, R. Sharma, S. Suran, and D. Draheim. On measuring confounding bias in mixed multidimensional data. In J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, editors, *Proceedings of ICTIS'2024 – the 8th International Conference on ICT for Intelligent Systems, Volume 6*, volume 1112 of *Lecture Notes in Network and Systems*, pages 329–342, Singapore, 2024. Springer Nature Singapore. doi:10.1007/978-981-97-6684-0_27

III S. Arakkal Peious, R. Sharma, M. Kaushik, S. Mahtab, and D. Draheim. On observing patterns of correlations during drill-down. In P. D. Haghighi, E. Pardede, G. Dobbie, V. Yogarajan, N. A. Sanjaya, G. Kotsis, and I. Khalil, editors, *Proceedings of iiWAS'2023 – the 25th International Conference on Information Integration and Web-based Applications and Services*, volume 14416 of *Lecture Notes in Computer Science*, pages 134–143, Cham, 2023. Springer. doi:10.1007/978-3-031-48316-5_16

IV S. Arakkal Peious, S. Suran, V. Pattanaik, and D. Draheim. Enabling sensemaking and trust in communities: An organizational perspective. In E. Pardede, M. Indrawan-Santiago, P. D. Haghighi, M. Steinbauer, I. Khalil, and G. Kotsis, editors, *Proceedings of iiWAS'2021 – the 23rd International Conference on Information Integration and Web Intelligence*, pages 95–103. Association for Computing Machinery, 2021. doi:10.1145/3487664.3487678

V S. Arakkal Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia. Grand reports: A tool for generalizing association rule mining to numeric target values. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery*, volume 12393 of *Lecture Notes in Computer Science*, pages 28–37, Cham, 2020. Springer. doi:10.1007/978-3-030-59065-9_3

Author's Contributions to the Publications

I I was the lead author, main author and corresponding author of this publication. I was responsible for the article content, writing the manuscript constructing the discussion, formulating the future plan and proofreading.

II I was the lead author, main author and corresponding author of this publication. I was responsible for the article content, writing the manuscript, writing the simulation program, collecting and analyzing data, interpreting the results, constructing the discussion, formulating the future plan and proofreading.

III I was the lead author, main author and corresponding author of this publication. I was responsible for the article content, writing the manuscript, writing the simulation program, collecting and analyzing data, interpreting the results, constructing the discussion, formulating the future plan and proofreading.

IV I was the lead author, main author and corresponding author of this publication. I was responsible for the article content, writing the manuscript, writing the simulation program, collecting and analyzing data, interpreting the results, constructing the discussion, formulating the future plan and proofreading.

V I was the lead author, main author and corresponding author of this publication. I was responsible for the article content, writing the manuscript, writing the simulation program, collecting and analyzing data, interpreting the results, constructing the discussion, formulating the future plan and proofreading.

Abbreviations

ANCOVA	analysis of covariance
ARM	association rule mining
CBPR	community-based participatory research
CEQ	Commitment to Equity
CI	collective intelligence
C-IA	coupled impact assessment
CMDC	China Meteorological Data Service Center
DAG	directed acyclic graph
DOI	Digital Object Identifier
ERP	Enterprise Resource Planning
GIS	geographic information systems
GPU	graphical processing unit
HLM	hierarchical linear modeling
i.i.d.	independent , identically distributed
IoT	Internet of Things
LDL	low-density lipoprotein
LPG	Liquefied petroleum gas
MCC	Matthews correlation coefficient
NARM	numerical association rule mining
NHANES	National Health and Nutrition Examination Survey
OB	Oaxaca-Blinder
OLAP	online analytical processing
PAH	polycyclic aromatic hydrocarbons
RCT	randomized controlled trial
SEM	structural equation modeling
TPU	tensor processing unit

Terms

Clinical study	"A type of research study that tests how well new medical approaches work in people. These studies test new methods of screening, prevention, diagnosis, or treatment of a disease."[1]
Confounder	"something that affects the result of a scientific experiment in a way that makes it less clear that one thing causes another, because it has an effect on one of the things that is being measured [...]"[2]
Exposure	variable that has been intentionally included in a study as influencing factors
Extraneous variable	any variable other than those intentionally included in a study as influencing factor "[...] an extraneous variable whose presence affects the variables being studied so that the results do not reflect the actual relationship between the variables under study."[3]
Latent variable	unobserved extraneous variable; or (used in a more narrow sense): unobserved confounder

[1] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/clinical-study

[2] https://dictionary.cambridge.org/dictionary/english/confounder

[3] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi. How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from Bed to Bench*, 5(2):79–83, 2012. doi:10.22037/ghfbb.v5i2.246

Symbols

$\mathsf{E}(Y)$	expectation of a random variable Y
$\mathsf{E}_X(Y)$	conditional expectation of a random variable Y, conditional on an event X
$\mathsf{E}(Y\|X)$	conditional expectation of a random variable Y, conditional on an event X
$\iota(X \Rightarrow Y)$	C-IA measure, Def. 15, coupled impact assessment of a categorical random variable X onto a categorical random variable Y
$\mathsf{P}(X)$	probability of an event X
ρ_{XY}	Pearson correlation coefficient between random variables X and Y
σ_X	Standard deviation of random variable X

1 Introduction

1.1 Motivation

Understanding the relationships between exposures and outcomes is a cornerstone of data analysis. Unfortunately, these relationships may be misleading whenever there are further latent factors that influence the result. In such cases, these may introduce a form of distortion, which is occasioned by the presence of a third variable known as a confounder [175, 12]. A confounding variable influences both the exposure and the outcome [13]. This dual influence can complicate the observed correlation as it leads to imprecise, biased estimations, which then could be critically misleading regarding causality. For instance, a confounder may produce the appearance of a causal relationship that does actually not exist, or on the other hand, obscure the existence of a real causal link.

Essentially, confounding is about the issue that the direct effect of exposures is mixed with the potential effect of confounders. The issue of confounding makes it necessary to systematically assess both the exposure effects and the potential effects of confounders. Neglecting confounding may affect the conclusions and weaken the reliability of inferences made on the basis of data.

This thesis explores and evaluates established measures and a novel measure of impact and confounding. Through a combination of theoretical analysis and experimental comparisons, we aim to assess the strengths, limitations, and practical applications of these measures. The findings contribute to a deeper understanding of the methodological frameworks available for causal inference and guide researchers in selecting appropriate tools for their studies.

1.2 Study Relevance and Design

See Figure 1 for an overview on how this study evolved. In Publication IV, we have examined how organizations can facilitate sensemaking, i.e., the process by which individuals interpret and make sense of complex information and build trust within communities. The study in Publication IV explores the roles of business intelligence [43, 30, 23], collective intelligence (CI) [212, 64, 139, 131], and crowdsourcing [34, 33, 11, 38, 31] in enhancing these processes. We implemented a minimal CI platform for basic problem solving enabling a novel reputation model. The tool was evaluated by 50 users, who used the tool over a period of two weeks and answered questionnaires on its usefulness and usability [51]. During the course of this study, we have identified a significant gap in the availability of a comprehensive reporting platform that is equipped with data mining capabilities to support decision-makers effectively. To address this need, we envisioned the development of a platform that integrates multiple data analysis methods [65, 62, 193, 197]. As an initial step, we focused on implementing association rule mining (ARM) [4, 204, 5, 117] as a foundational analytical technique leading to a tool for so so-called *grand reports*[4][65, 62], see also Publication V.

Standard ARM techniques often necessitate discretizing numeric target variables [116, 120, 80, 152, 74], a process that can result in information loss and yield less precise insights. To overcome this limitation, we introduced a novel tool

[4] called *grand pivot reports* in [65]

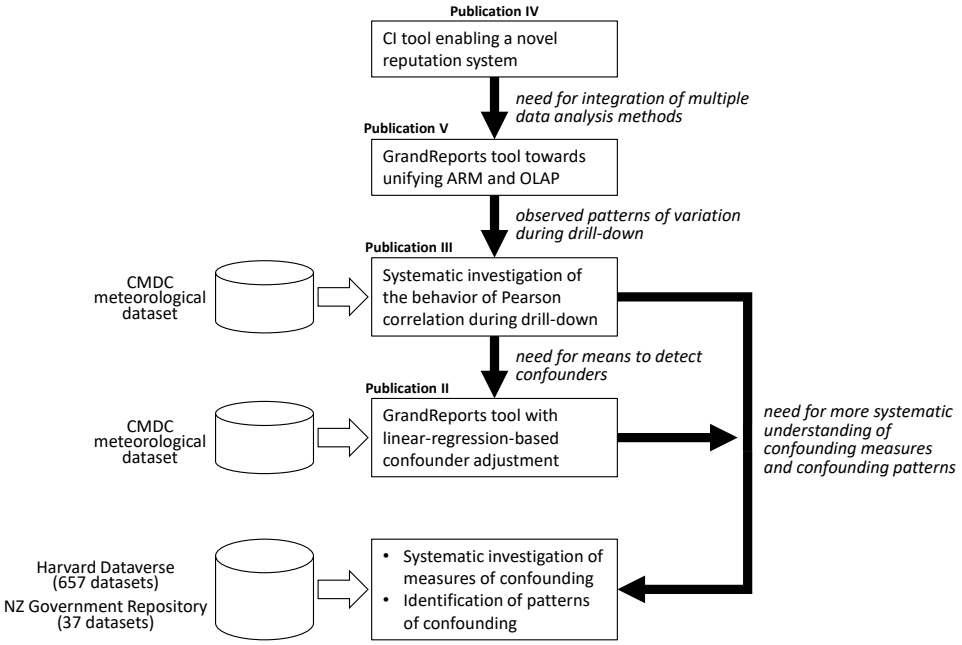


Figure 1: Evolvement of this study.

named *GrandReport*[5,6], see Publication V, which enhances ARM by calculating and reporting the mean values of a selected numeric target column across all possible combinations of influencing factors, see Publication V. This tool enables decision-makers to interpret associations based on aggregate values as in *online analytical processing* (OLAP) [47, 49, 3], however, automatically as in ARM instead of interactively as in OLAP and at the scale of ARM, providing results that align more closely with real-world analytical practices.

While conducting numerous analyses using the *GrandReport* platform, see Publication V, we observed significant patterns of variation in results between marginal values and drill-down analyses. Drill-down analyses enables users to transition from high-level summary data to more detailed levels, thereby facilitating a deeper understanding of data patterns [151, 186]. In Publication III, we have specifically investigated the behaviour of Pearson correlation coefficients between variables across different levels of data granularity during drill-down analysis. For our experiments, we used the well-known meteorological CMDC (China Meteorological Data Service Center) data set[7] [45] from China. The findings revealed that correlations observed at the marginal level do not always persist during drill-down, highlighting the risk in high-level analyses to overlook nuanced relationships that emerge at more detailed levels. This result underscores the importance of conducting drill-down analyses to capture the full spectrum of data relationships, ultimately enhancing decision-making processes.

The observed behavioural differences in correlation patterns between marginal and drill-down analyses suggest the presence of statistical paradoxes [192, 194,

[5] http://grandreport.me
[6] https://github.com/istaltech/grandreport
[7] http://data.cma.cn/en

196, 195] and other data-related fallacies [221] such as selecting suboptimal measures [198, 84, 106], and, last but not least, confounding effects [109, 170, 182, 183, 150]. To address these challenges, we have expanded the capabilities of the *GrandReport* platform[5,6] by incorporating additional data analytical methods, see Publication II. This enhancement aims to generate diverse perspectives and identify potential confounding effects within datasets. Specifically, we integrated multiple linear regression adjustment into the platform to mitigate confounding effects in mixed multidimensional data, see Publication II. This enhancement allows for more accurate analyses by adjusting for confounders without requiring the segregation of numerical and categorical data. In order to provide evidence for the usefulness and usability of the suggested approach, we have utilized the extended features of the *GrandReport* tool to investigate the correlation between standardized air quality indices[8,9] [54] and CO2 levels[10] before and after confounder adjustment in regard of additional city data with the CMDC (China Meteorological Data Service Center) data set[7] [45].

Given the results and the experience of the previous research stages, we aimed at contributing to a more systematic understanding of confounding measures and confounding patterns. We decided to compare various methods of confounding adjustments using the same datasets on a large scale, in order to gain deeper insights into the confounding effect. For this comparison, we have selected three established approaches, i.e., the familiar Ad-Hoc method, the Oaxaca-Blinder decomposition method, and the linear-regression-based method for confounding adjustment. Additionally, we introduce a novel approach for confounding adjustment for categorical factors *in their entirety*. This comparative analysis aims to evaluate the effectiveness and scalability of these methods in addressing confounding effects within multidimensional data.

For the purpose of our investigation, we have conducted experiments at a large scale (combinatorial along all possible drill-downs, i.e., in the style of grand reports, see Publication V and [65]) with 657 datasets from the Harvard Dataverse[11] repository, plus, 37 datasets from the New Zealand (NZ) government repository[12]. During the course of the study, we have been able to detect four interesting patterns of multiple confounder behaviour, that we report in a series of dataset case studies.

In mature scientific disciplines, studies are well aware of potential confounding effects; and their treatment and adjustment is standard in the vast literature of, e.g., clinical and epidemiological studies. In these studies, factors are intentionally selected to be included in the study to provide evidence for or against a hypothesis in terms of these factors. All other variables than the selected factors are considered as extraneous variables, may they be observed or not. The situation changes when it comes to the field of data mining. The data mining paradigm is more exploratory, rather serving the discovery of hypotheses in regards of potentially interesting influencing factors. Now, although confounding effects are ubiquitous in data with many factors, confounding is rather neglected in data mining tools, i.e., no systematic support for confounders (for detecting them, for

[8] https://www.transportpolicy.net/topic/air-quality-standards/

[9] https://www.legislation.gov.au/Details/F2016C00215

[10] http://www.cityghg.com/toArticleDetail?id=203#

[11] https://dataverse.harvard.edu/

[12] https://data.govt.nz/

taking them into account or even for adjusting them) is offered in the common data mining tools, be it from Association rule mining (ARM) or Online analytical processing (OLAP). Here is, where this study aims at envisioning a paradigm shift, i.e., to design an integrate systematic support for treatment of confounding in data mining tools.

1.3 Contributions

The thesis contributes to the state of art as follows:

(i.) We introduce a novel measure for the impact of a categorical variable as a whole, i.e., in its entirety, called Coupled Impact Assessment (C-IA). Furthermore, we introduce a novel method to detect confounders utilizing the C-IA measure.

(ii.) We have conducted combinatorially designed experiments with 694 datasets from the Harvard Dataverse and the NZ Government Repository to investigate three well-established methods for detecting confounders (the Ad-Hoc method, Oaxaca-Blinder decomposition, linear-regression-based) and our own novel C-IA-based method.

(iii.) Based on our experiment results, we have discovered, that the four investigated methods for detecting confounders do not show any relevant agreement or disagreement beyond chance (in terms of both Cohen's κ and Yule's ϕ). We argue, that this fact is surprising and highly relevant.

(iv.) Based on our experiments results, we have identified four patterns of confounding effects during drill-down into potential confounders, that we showcase in eight data case studies – two data case studies for each of the investigated method.

(v.) Furthermore, we elaborate a systematic interpretation of the linear regression model utilizing so-called multiplicative edges diagrams. We utilize this interpretation to reflect on linear-regression-based confounding, including a critical discussion off cutoff rules for confounding adjustments.

1.4 Thesis Outline

We proceed as follows. In Chapter 2, we detail out some of the notation that we use throughout the thesis. In Chapter 3, we provide detailed background material and discussion relevant to the study. We explain the concept of confounding and discuss its relevance from various perspectives. Furthermore, we explain the straightforward approach of confounder adjustment based on the stratification of data and coercion of probabilities to outer margins. Furthermore, we explain the theoretical foundations of Oaxaca-Blinder decomposition and its limitations as known from the literature. In Chapter 4, we provide a systematic discussion of the utilization of multiple linear regression for detecting and adjusting confounders. In Chapter 5, we introduce C-IA (Coupled Impact Assessment), a novel measure to assess the impact of a categorical factor in its entirety. In Chapter 6, we explain, how the C-IA measure of Chapter 5 can be utilized to detect and adjust confounders. In Chapter 7, we provide a detailed explanation of the datasets used in our study and the design of our experiments. In Chapter 8, we present

and discuss the results of our conducted experiments. In Chapter 9, we discuss limitations of our approach and potential future research directions. We finish the thesis with a conclusion in Chapter 10, providing a summary of the main insights and a future outlook of the work.

2 Basic Notation and Terminology

2.1 Modeling of Factors of a Data Set

Throughout the thesis, we use random variables to present the factors of a dataset. In computer science, various terminological and notational frameworks have been elaborated to present and analyze data sets, each with its specific intentions and flavors, among which are, most importantly, relational algebra [46] for the field of relational databases and OLAP [47], and the itemset apparatus [4] of association rule mining (ARM) [4, 204, 5, 117]. As important as fields such as OLAP and ARM are for the topic of this thesis, we stay with notation and terminology of statistics [75, 154, 155, 110] and probability theory [126, 154] in this thesis, as it is the widespread notation in medical research, biometrics and other data-intensive fields. Specific concepts from the computer science literature, such as drill-down or roll-up from OLAP, or confidence and lift from ARM show natural in our notation. For a more formal account on translating between the various formal frameworks, see [193, 197].

In this chapter we detail out some of the notation and terminology that we use throughout the thesis.

Definition 1 (Tuple Projection)**.** Given a tuple $\langle x_1, \ldots, x_n \rangle \in V_1 \times \cdots \times V_n$ we define the *i-th tuple projection* $\pi_i : V_1 \times \cdots \times V_n \longrightarrow V_i$ for any $1 \leqslant i \leqslant n$ as usual as follows:

$$\pi_i\Big(\langle x_1, \ldots, x_i, \ldots, x_n \rangle\Big) = x_i \tag{1}$$

Definition 2 (Dataset, Factors, Data points)**.** We model a finite *dataset (data table)* Ω of $n = |\Omega|$ *data points (rows)* as an indexed set

$$\Omega : \underbrace{\{1, \ldots, n\}}_{\substack{\text{data point} \\ \text{indices}}} \longrightarrow \underbrace{V_1 \times \cdots \times V_m}_{\text{data points}}$$

as belonging to a probability space $(\Omega, \Sigma, \mathsf{P})$ together with m random variables called *factors* (also called *variables*, also called *columns*)

$$\begin{aligned} X_1 &: \Omega \longrightarrow V_1 \\ &\vdots \quad \vdots \\ X_m &: \Omega \longrightarrow V_m \end{aligned}$$

so that

$$X_j(\langle i, v_1, \ldots, v_j, \ldots v_m \rangle) = v_j$$

for all factor indices $1 \leqslant j \leqslant m$, all data point indices $i \in \{1, \ldots, n\}$ and all factor values $v_1 \in V_1, \ldots, v_n \in V_n$, and, furthermore,

$$\mathsf{P}(A \subseteq \Omega) = \frac{|A|}{|\Omega|}$$

for all events $A \subseteq \Omega$. Given a row index $i \in \{1, \ldots, n\}$, we use the notation

$$X_{i_1}, \ldots, X_{i_m} \tag{2}$$

as well as

$$X_{1_i}, \ldots, X_{m_i} \tag{3}$$

to denote the *data point*

$$\Omega(i),$$

i.e.,

$$\Big\langle \pi_1\big(\Omega(i)\big), \ldots \pi_m\big(\Omega(i)\big) \Big\rangle.$$

Please note, that according to Def. 2:

- We use *factor*, *variable* and *column* as synonyms throughout the thesis.
- We use *data point* and *row* as synonyms throughout the thesis.

Let us demonstrate, how Def. 2 works in practice by considering the example in Table 1. Table 1 shows a dataset consisting of citizen records [62], for each citizen recording their salary, the city in which they live, their profession, their educational level, their age group, and whether they are a freelancer or not. The data set is heterogeneous in terms of the statistical scale of the columns, i.e., the factor *Salary* is a numerical factor, whereas the factors *Profession*, *Education*, *AgeGroup* are categorical variables, each taking values from a finite set of distinct categories, and *Freelance* is a binary variable, as follows:

$$\begin{aligned}
\mathit{City} &: \Omega \longrightarrow \{\mathit{Boston}, \mathit{LA}, \mathit{NY}, \mathit{Seattle}, \ldots\} && (4)\\
\mathit{Profession} &: \Omega \longrightarrow \{\mathit{Chef}, \mathit{Builder}, \mathit{IT}, \mathit{Lawyer}, \ldots\} && (5)\\
\mathit{Education} &: \Omega \longrightarrow \{\mathit{High\ School}, \mathit{Bachelor}, \mathit{Master}, \mathit{PhD}\} && (6)\\
\mathit{AgeGroup} &: \Omega \longrightarrow \{18-25, 25-30, 30-40, 40-50, 50-58, 59-65, >65\} && (7)\\
\mathit{Freelancer} &: \Omega \longrightarrow \mathbb{B} && (8)\\
\mathit{Salary} &: \Omega \longrightarrow \mathbb{R}_0^+ && (9)
\end{aligned}$$

Next, let us have a look, how some of the OLAP and ARM terminology would show in our notation. As an example from OLAP terminology, a *drill-down* of the salary into a particular value of the *dimension* city, e.g., *Boston*, would show as stepping from the expectation of the salary

$$\mathrm{E}(\mathit{Salary}) \tag{10}$$

to the conditional expectation

$$\mathrm{E}(\mathit{Salary} \mid \mathit{City}{=}\mathit{Boston}). \tag{11}$$

Further *drill-down* into the dimension *Profession* with value *Lawyer* would show as stepping further to the conditional expectation

$$\mathrm{E}(\mathit{Salary} \mid \mathit{City}{=}\mathit{Boston}, \mathit{Profession}{=}\mathit{Lawyer}). \tag{12}$$

Next, in terminology of ARM, an example of an association rule in terms of Table 1 could be

$$\{\mathit{Profession}{=}\mathit{IT}, \mathit{AgeGroup}{=}\mathit{40\text{-}50}\} \Longrightarrow \{\mathit{City}{=}\mathit{Seattle}\} \tag{13}$$

Table 1: Example table in regards of Def. 2.

City	Profession	Education	Age Group	Freelancer	Salary
NY	Lawyer	Master	25–30	0	3.800
Seattle	IT	Bachelor	18–25	1	4.200
Boston	Lawyer	PhD	40–50	1	12.700
LA	Chef	High School	30–40	0	3.700
...	...	...	...	...	...

Now, according to ARM terminology, the *confidence* of the *association rule* (13) corresponds to the conditional probability

$$\mathrm{P}(\mathit{City{=}Seattle} \,|\, \mathit{Profession{=}IT}, \mathit{AgeGroup{=}40\text{-}50}), \tag{14}$$

whereas the *lift* of the *association rule* (13) would show as the following change factor:

$$\frac{\mathrm{P}(\mathit{City{=}Seattle} \,|\, \mathit{Profession{=}IT}, \mathit{AgeGroup{=}40\text{-}50})}{\mathrm{P}(\mathit{City{=}Seattle})} \tag{15}$$

Having random variables for representing factors instead of choosing a more specific notation such as relational algebra or the itemset apparatus is a good choice, as it gives us the most flexibility in argumentation, independent from any more specific terminology, and, even more important, directly unlocks all the whole apparatus of statistics and probability theory for our argumentation. For example, the concept of lift presented in terms of conditional probabilities in (15), can be easily adopted to numerical target values by stepping from probabilities to expectations, e.g., for salaries as follows:

$$\frac{\mathrm{E}(\mathit{Salary} \,|\, \mathit{Profession{=}IT}, \mathit{AgeGroup{=}40\text{-}50})}{\mathrm{E}(\mathit{Salary})}, \tag{16}$$

whereas, there exist no "lift" as presented by (16) in standard ARM. This is so, because standard ARM is merely about bit tables and an association rule such as

$$\{\mathit{Profession{=}IT}, \mathit{AgeGroup{=}40\text{-}50}\} \Longrightarrow \{\mathit{Salary}\} \tag{17}$$

is simply not available in standard ARM. Many generalization of ARM exist, among those a variety of so-called *numerical association rule mining* (NARM) approaches[13], however, in this thesis, we usually prefer to treat concepts directly in the terminology of probability theory as outlined above.

2.2 Bilateral Impact Measure Graphs

On various occasions, we make use of so-called *bilateral impact measure graphs*, or just *impact graphs* for short. The nodes of an impact graph are the factors of

[13] For an exhaustive discussion of numerical generalizations of association rule mining, see [118, 117, 116, 119].

a given data set. An edge between nodes A and B of the impact graph visualizes some measured impact of A onto B in terms of some directed impact measure, see Def. 3. Impact graphs can be used for different impact measures; further details of what information is shown by an impact graph (such as used thresholds or additional labels) is always explained in the context where they appear. Indeed, we will use impact graphs for ad-hoc measure from Section 3.2, linear regression slopes, see Chapter 4, Oaxaca-Blinder decomposition, see Section 3.3, and our novel C-IA measure from Chapter 5.

In Section 4.2, we use a form of impact graph called *multiplicative edges diagram* to systematically elaborate an interpretation of the liner regression model.

Throughout the presentation of the data case studies in Sections 8.3, 8.4, 8.5, and 8.6, the purpose of an impact graph is to give a first impression and overview of the dataset and the existing associations between factors in it. It is not meant as a analytical result (although it can be used, to a certain extent, as an analytical tool), rather, it is meant to accompany the presentation of our combinatorial experiments as described in Sect. 7.2 to ease its understanding. For the same reason, the information of an impact graph is not necessarily complete, i.e., it might be used to highlight some dependencies in a data set and omit some others.

Definition 3 (Bilateral Impact Measure Graph)**.** The *bilateral impact measure graph*, or just *impact graph* for short, of a data set D and an impact measure M is a directed graph as follows. The nodes of the graph are the columns of the data set D. An arrow between two nodes c_1 and c_2 represents some measured impact of c_1 onto c_2 in terms of M.

It is important to grasp that impacts graph are *bilateral*, i.e., they visualize impact measure only at the outermost level of analysis. For example given an impact graph of three nodes A, B, and C, together with edges $A \longrightarrow C$ and $B \longrightarrow C$, it visualizes only some measured impact of A onto C as well as another measured impact of B onto C, but never some joint impact A and B onto C.

Furthermore, it is important to grasp that the edges in an impact graph do not necessarily visualize *causal impact*, where the long name of impact graph, i.e., *bilateral impact measure graph* is actually less misleading. In the same vain, our impacts graphs must not be confused with Bayesian Networks [159, 160, 161, 165]. A Bayesian Network is a DAG (directed acyclic graph), where an arrow from node n_1 to n_2 indicates that n_1 belongs to the Markovian parents of n_2 in the context of a fixed sequence of the nodes of the Bayesian network. In contrast to Bayesian network, the following holds for impact graphs:

- We use impact graphs for visualizing and analyzing any kind of impact measure, i.e., not only for indicating Markovian parents as in Bayesian Networks. Actually, we use impact graphs for such impact measures that are about the impact of a primary influencing factor in the context of potential confounders.
- An impact graph is not necessarily a DAG, i.e., it can have cycles. Impact is understood as potential impact and is not drawn from domain knowledge. Both directions of potential impact are analyzed between any two nodes, which might lead to cycles. Actually, the all instances of impacts graphs that we will present in the sequel are indeed DAGs, but this is only due to their more specific definitions, not to due the general definition of impact graph in Def. 3.

- For an impact graph, the set of nodes is not ordered, i.e., there is no sequence of nodes that would matter. Again, this is do to the fact that we do not pre-assume causalities from domain knowledge during our analyses.

2.3 Basic Measures

In the course of the thesis we need to compare values in regard of notions of degree of change. And these degree of changes can become again subject of comparison. For example, the impact of an event onto X onto a target random variable Y can be made formal by comparing the conditional probability $\mathsf{E}(Y|X)$ with either the marginal value $\mathsf{E}(Y)$ or with the "dual" conditional probability $\mathsf{E}(Y|\neg X)$. For the comparison itself, various basic measures could be used, see Table 2. Probably, the most common one from everyday world is the *relative difference*, which is about adding or subtracting a percentage of the base value to resp. from the base value, e.g., 3% inflation or 20% discount. However, in defining impact measures, studies usually use other basic measures. For an extensive account of various basic measures see, e.g., [148]. Unfortunately, terminology of basic measures is not standardized, see again Table 2, and studies might use different names for the same concept or even the same name for different concepts.

In this thesis, we use the *change factor* as defined in Def. 4 and the *percentage difference* as defined in Def. 5. The percentage difference is similar to the relative difference, just, it uses the mean of the two compared values as a base instead of one of the two compared values. The percentage difference is a basic measure that is particularly often used scientific studies, in particular, when it comes to comparing confounding effects.

Definition 4 (Change Factor)**.** Given real numbers me $a \in \mathbb{R}$ and $b \in \mathbb{R}$ such that $b \leq a$, we define their *change factor* as follows:

$$\frac{b}{a} \tag{18}$$

Definition 5 (Percentage Difference)**.** Given real numbers $a \in \mathbb{R}$ and $b \in \mathbb{R}$ such that $b \leq a$, we define their *percentage difference* as follows:

$$\frac{b-a}{\frac{a+b}{2}} \tag{19}$$

Table 2: Concepts of degree of changes.

absolute difference	$\lvert b-a\rvert$
relative difference, relative change, percentage change, increase ratio, *percentage difference*	$\frac{b-a}{a}$
change factor, change multiplier, change coefficient, lift (in ARM [4])	$\frac{b}{a}$
percentage difference	$\frac{b-a}{\frac{a+b}{2}} = 2\frac{b-a}{b+a}$
relative change (w.r.t. the arithmetic mean)	$\frac{b-\frac{a+b}{2}}{\frac{a+b}{2}} = \frac{b-a}{b+a}$

2.4 Inter-Rater Reliability, Cohen's κ and Yule's ϕ

2.4.1 Proportionate Agreement and Cohen's κ Defined

In brief, a measure of inter-rater reliability is a measure of the degree of agreement between instances of two categorical factors. Here, *rater* is another name for *factor*, as are *variable* and *column*. We need measures of inter-rater reliability for our comparison of confounding measures in Section 8.1.

The most basic and obvious measure of inter-rater reliability is *proportionate agreement*. In our study, we will use Cohen's κ coefficient [48, 142], because Cohen's κ is widely used[14] and has been widely discussed [97] throughout the literature. We define both proportionate agreement (Def. 6) and Cohen's κ (Def. 7) for the special case of 2×2 contingency tables only, i.e., two binary categorical variables, as this is for what we use the coefficient in our study. We further discuss Cohen's κ in Sections 2.4.2 and 2.4.3. We discuss significance tests for Cohen's κ in Sect. 2.4.4.

Definition 6 (Proportionate Agreement (Binary Case))**.** Given two events $A \subseteq \Omega$ and $B \subseteq \Omega$, called *rater*, we define their *proportionate agreement* p_{AB} as follows:

$$p_{AB} = \mathrm{P}(AB) + \mathrm{P}(\neg A \neg B) \tag{20}$$

Definition 7 (Cohen's κ (Binary Case))**.** Given two events $A \subseteq \Omega$ and $B \subseteq \Omega$, called *rater*, we define their *observed proportionate agreement* p_{AB}, *hypothetical random agreement* (*agreement by chance*) $\widetilde{p}_{AB}$ and *Cohen's* κ_{AB} as follows:

$$\kappa_{AB} = \frac{\overbrace{\Big(\mathrm{P}(AB) + \mathrm{P}(\neg A \neg B)\Big)}^{p_{AB}} - \overbrace{\Big(\mathrm{P}(A)\mathrm{P}(B) + \mathrm{P}(\neg A)\mathrm{P}(\neg B)\Big)}^{\widetilde{p}_{AB}}}{1 - \underbrace{\Big(\mathrm{P}(A)\mathrm{P}(B) + \mathrm{P}(\neg A)\mathrm{P}(\neg B)\Big)}_{\widetilde{p}_{AB}}} \tag{21}$$

[14] The Scopus query TITLE-ABS-KEY("Cohen's kappa") on www.scopus.com yields 13,092 documents as of 22 January 2005.

For the sake of completeness and comparison, we also provide the definition of the measure of *accuracy* as widely used in epidemiology and machine learning in Def. 8.

Definition 8 (Accuracy)**.** Given two events $A \subseteq \Omega$ called actual condition (testee, disease etc.) and $B \subseteq \Omega$, called prediction (test, diagnosis etc.), furthermore, a with n individuals, *true positives* $\mathrm{TP} = n\,\mathrm{P}(AB)$, and *true negatives* $\mathrm{TN} = n\,\mathrm{P}(\neg A \neg B)$ as usual, the measure of *accuracy* a_{AB} is defined as follows:

$$a_{AB} = \frac{\mathrm{TP}+\mathrm{TN}}{n} \tag{22}$$

Let us compare proportional agreement in Def. 6 and accuracy in Def. 8. Mathematically, i.e., merely in terms of data, both measures are simply the same. The difference is only in their application. When A is considered as a ground truth and B is a prediction, the measure is usually called *accuracy*. When both A and B are predictions of a third ground truth, the measure it usually called *proportional agreement*.

2.4.2 Intuition Behind Cohen's κ

The idea of Cohen's κ is as follows. Given any two independent events $A \subseteq \Omega$ and $B \subseteq \Omega$, i.e., such that $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$, they would show some proportionate agreement $p_{AB} = \mathrm{P}(A)\mathrm{P}(B) + \mathrm{P}(\neg A)\mathrm{P}(\neg B)$. However, intuitively, a measure of agreement should assess two independent factors as having neutral agreement, i.e., no agreement. Otherwise, for example, if we have two independent factors A and B with $\mathrm{P}(A) = \mathrm{P}(B) = 0.9$ and, furthermore, two independent factors A' and B' with $\mathrm{P}(A) = \mathrm{P}(B) = 0.5$, we have that their proportionate agreements significantly differ with $p_{AB} = 0.82$ and $p_{A'B'} = 0.5$, although they are both just independent variables and the proportionate agreements result just from chance in this case. What we would like to have is a measure of inter-rater reliability that amounts for the agreement *beyond* the agreement that is given by chance anyhow, i.e., we would like have a measure that filters out the agreement that is given by chance to achieve robust comparability of dependent variables. This is exactly what Cohen's κ aims at as follows.

Intuitively, the hypothetical random agreement $\widetilde{p}_{AB}$ can be interpreted as the agreement that is achieved by chance anyhow. Now, $1 - \widetilde{p}_{AB}$ is the room (or playground) for any *potential agreement beyond chance* and $p_{AB} - \widetilde{p}_{AB}$ is the *actual (or observed) agreement beyond chance*. Now, Cohen's κ is the *relative change* of *actual* and *potential* agreement beyond chance, see (21), yielding a robust measure of factor agreement.

2.4.3 Interpretation of Cohen's κ

In this study, we deal only with κ-values that are close to zero. We interpret κ-values that are zero as *no agreement*, and κ-values that are close to zero as *almost no agreement*, see Def. 9.

Definition 9 (Study's Interpretation of Cohen's κ)**.** We interpret values of $\kappa = 0$ as *no agreement* (also: *no agreement/disagreement*) and values of $\kappa \approx 0$ close to zero, i.e., $-0.1 \leq \kappa \leq 0.1$, as *almost no agreement* (also: *any relevant agreement/disagreement beyond chance*), where the thresholds of -0.1 and 0.1 are our arbitrary choice.

Several classification schemes for the interpretation of κ have been suggested. In Table 3, we have summarized the classification of Fleiss et al. [77]. Fleiss' classification should be taken as authoritative source, as Fleiss co-authored a seminal paper on the standard error of Cohen's κ [76]. See Table 4, for two other widely used classification schemes for κ , i.e., those of Landis and Koch [128] (the paper [128] has 60,129 citations on Scopus[15]) as well as McHugh [142] (the paper [142] has 12,796 citations on Scopus[15]). Note that such classifications of κ are mere heuristics. For example, Landis and Koch [128] state: "Although these divisions are clearly arbitrary, they do provide useful »benchmarks «[...]" ([128], p. 165). In the same vein, Fleiss et al. [76] state: "For most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance." ([77], p. 604).

Table 3: Interpretation of Cohen's κ (rounded) according to Fleiss et al. ([77], p. 604).

Range	Interpretation
0.00–0.40	Poor
0.40-0.75	Fair/Good
0.75-1.00	Excellent

Table 4: Interpretation of κ according to Landis and Koch, and Mary McHugh.

Range	Landis and Koch [128]	McHugh [142]
< 0.00	Poor	–
0.00–0.20	Slight	None
0.21–0.40	Fair	Minimal
0.41–0.60	Moderate	Weak
0.61–0.80	Substantial	Strong
0.81–1.00	Almost perfect	Almost Perfect

Both Fleiss et al. [76] (Table 3) and Landis and Koch [128] (Table 4) classify $\kappa = 0$ as *poor agreement*, whereas McHuge classifies $\kappa = 0$ as *none*, i.e., *no agreement* (actually, McHuge classifies $\kappa \leq 0.2$ as *none*). We follow McHuge in this regard. We argue that understanding $\kappa = 0$ as *no agreement* is the key assumption of the intuition behind the κ statistics, see the discussion in Section 2.4.3, therefore, the interpretation of $\kappa = 0$ should be *no agreement*. And indeed, in [76], p. 325, Fleiss, Cohen and Everitt talk about $\kappa = 0$ as "the case of no association", which confirms our classification of $\kappa = 0$ as *no agreement*.

In the same vein, we argue that κ-values that are close to zero can be interpreted as *almost no agreement*, and that this is a valid interpretation for both

[15] as of 25 Feb 2025

positive and negative values of κ. We have set arbitrary thresholds of -0.1 and 0.1 for κ-values to be considered as *almost zero* and, therefore, as showing *no agreement* (note, that McHuge has choosen a weaker threshold of 0.2, see Table 4), , see Def. 9. In light of the interpretation of κ, see Section 2.4.3, we argue that negative κ-values should be interpreted as *disagreement* rather than *poor agreement* as suggested by Landis and Koch, see Table 3. Therefore, in Def. 9, we say that the case $\kappa = 0$ shows *no agreement* and equally can be said to show *no agreement/disagreement* (and for $-0.1 \leq \kappa \leq 0.1$ alike).

2.4.4 Significance Tests for Cohen's κ

In Def. 10, we define the standard error of Cohen's κ under the null hypothesis that $\kappa = 0$ ($\mathrm{H}_0 : \kappa = 0$), for 2×2 contingency tables as a special case of the definition provided by Fleiss et al. [76, 77], in the version of [77], page 605, Equation (18.13).

Definition 10 (Standard Error of Cohen's κ for $\mathrm{H}_0 : \kappa = 0$ (Binary Case) ([76], Eqn. (18.13))**.** Given N observations with contingency table of two events $A \in \Omega$ and $B \in \Omega$, the standard error of κ_{AB} under the assumption of no association $\mathrm{H}_0 : \kappa = 0$, denoted as SE_0, is defined as follows ([76], Eqn. [14], p. 325):

$$SE_0 = \frac{1}{(1-\widetilde{p}_{AB})\sqrt{N}} \sqrt{\widetilde{p}_{AB} + \widetilde{p}_{AB}^2 + \mathrm{P}(A)\mathrm{P}(B)\Big(\mathrm{P}(A)+\mathrm{P}(B)\Big) + \mathrm{P}(\neg A)\mathrm{P}(\neg B)\Big(\mathrm{P}(\neg A)+\mathrm{P}(\neg B)\Big)} \tag{23}$$

Now, the z-statistic for κ under the null hypothesis that $\kappa = 0$ ($\mathrm{H}_0 : \kappa = 0$) is defined as follows, see [77], page 605, Equation (18.14):

$$z = \frac{\kappa}{SE_0} \tag{24}$$

In our study we will use the z-statitic in (24) to obtain the p-value for κ under the null hypothesis that $\kappa = 0$ ($\mathrm{H}_0 : \kappa = 0$).

2.4.5 The ϕ Coefficient

The ϕ-coefficient, introduced by Udny Yule in 1912 [234], and commonly known as *Matthews correlation coefficient* (MCC) [140], is a widely used correlation coefficient. The ϕ-coefficient is a correlation measure for 2×2 contingency tables that ranges between -1 and $+1$, see Def. 11. We will use the ϕ-coefficient in our study in addition to Cohen's κ, to reconfirm our observations through the lens of an additional measure.

Definition 11 (ϕ-Coefficient)**.** Given two events $A \subseteq \Omega$ and $B \subseteq \Omega$, we define their *ϕ-coefficient* ϕ_{AB} as follows:

$$\phi_{AB} = \frac{\mathrm{P}(AB)\mathrm{P}(\neg A\,\neg B) - \mathrm{P}(A\neg B)\mathrm{P}(\neg AB)}{\sqrt{\mathrm{P}(A)\mathrm{P}(\neg A)\mathrm{P}(B)\mathrm{P}(\neg B)}} \tag{25}$$

The interpretation of the ϕ-coefficient is similar to the interpretation of the Pearson coefficient ρ [166], with $+1$ and -1 indicating *perfect correlation*, and 0 indicating *no correlation*. Indeed, in case of independence of events A and B, we have that their ϕ-coefficient equals 0, as, in this case:

$$\phi_{AB} = \frac{\mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(\neg A)\mathrm{P}(\neg B) - \mathrm{P}(A)\mathrm{P}(\neg B)\mathrm{P}(\neg A)\mathrm{P}(B)}{\sqrt{\mathrm{P}(A)\mathrm{P}(\neg A)\mathrm{P}(B)\mathrm{P}(\neg B)}} = 0 \tag{26}$$

3 Background

3.1 Confounding Effects

Science is the inexhaustible process of mankind aiming at systematically gaining knowledge. Scientific activities have been unfolding as they reveal facts about phenomena to humans over the course of their existence. The discussion that occurs among scientists when learning about each others' results and assess the precision and interpretation of those results is, to my mind, a fascinating process. Data integrity and accuracy of data analytics are still among the most critical issues among the scientific research community.

As far as data analysis is concerned, the issues of data accuracy and adequacy of used tools determine the quality of the investigation's outcome. At the same time, in the field of data analysis, utilizing experimental methods is promising and might become relevant, however, in order to implement them using observational data, some existing methods need to be modified or entirely new methods might have to be introduced. In any case, genuine empirical information must be assigned with specific properties so that the appropriate research tools can be used to discover the essence of the data and enable their understanding.

Let us consider the distinction between experimental and observational data. Experimental data collection involves carefully arranging and maintaining the conditions under which the experiment is conducted. so that the experimental results are, ideally, uninfluenced by external factors. The ambition is to allow researchers to control the variables and trace the relationships with highest possible precision. With observational data, information is generated from pure observations only, which are rather not deliberately created and controlled. Likewise, unobserved relevant factors that naturally exist might not be observed, i.e., overlooked. Over the time, statistical methods have been advanced and improved, in regard of allowing for the adjustment for confounding effects occurring in the data.

3.1.1 Today's Data Landscape

Statistical hypothesis testing and regression analysis have been indispensable in the context of experimental data analysis in scientific studies. These tools allow you to present not only the result of the experiment, but also the relations between variables. However, the limits of the existing methods are often hindered by the problem of non-randomized sampling, measurement errors, and uncontrollable variables among other factors.

Alongside this, we have the pathway of observational data analysis, which nowadays utilizes big data analysis [43, 188] to handle and process large amounts of data that could be handled by human hands. A current trend is the application of modern distributed computation tools [189, 190, 191, 27] such as Apache Hadoop[16] and Apache Spark[17] to enable parallel execution of jobs on clusters of machines or nodes located across a distributed system that would be impractical with traditional resources to solve analytical challenges.

In the landscape of modern data analysis, there is an immense volume of data that researchers want to investigate, comprehend, and that can occur in both structured and unstructured forms. Regardless of whether these data are well

[16] https://hadoop.apache.org/
[17] https://spark.apache.org/

ordered in a database or are unstructured and distribute over different sources, they still lead to paramount discoveries and insights. However, to fulfill this potential and make these data calculable and prepared for subsequent analysis, they should go through extensive data pre-processing [98, 223]. Moreover, sparse data form an important element of many of today's research studies; and the inquiry into sparse data creates additional challenges.

Structured data follow predefined schemas and is organized in strict data structures, e.g., in tables with rows and columns [46]. Therefore, structured data are easy to store and retrieve with the help of database management systems. Typical structured data are master data (staff data, customer data, financial data etc.) and transaction data (orders, sales etc.) of today's enterprise resource planning (ERP) systems, furthermore, sensor data in today's Internet of Things (IoT) [187]. On the other hand, unstructured data do not come with an entirely fixed format. Examples of unstructured data are text documents, images, audio recordings, social media posts, and Web pages. Both structured and unstructured data require pre-processing and transformations, in order to make data manageable and to utilize its information.

Text processing is a sequence of operations aimed at standardization, purification and structuring text data so as to enable more advanced analysis. Text processing is an important facilitator that allows for computing both structured and unstructured data for subsequent analysis. By systematic pre-processing, transformation, and structuring of text data, researchers can uncover the latent value that these data carry, from which they derive useful knowledge to drive scientific discovery, support decision-making, and foster innovation. As the volume and diversity of textual information steadily grows, the development of text processing techniques becomes a must.

3.1.2 Relevance of Confounding

Biased data and confounding variables pose a great challenge determining the most appropriate data analyses procedures and the sensitivity analyses since the quality of the outcome of a study will be primarily determined by the quality of the study design. Sensitivity analysis [181] is concerned with the relationship between uncertainty in assumptions and parameters and uncertainty of outcomes. Consequently, sensitivity analysis represents the entirety of judgments of the accuracy of results drawn by utilizing observational data as foundation of analysis.

In today's studies, we see increasingly more joint efforts of statisticians, data scientists, domain specialists, and subject matter experts illustrating a multidisciplinary mode of reasoning which passes beyond disciplinary boundaries. The multiplicity of perspectives and interdisciplinary sensibilities can lead to the creation of composite solutions to problems, and they can utilize the richness of observational data as a powerful source underlying scientific knowledge. Flexing creativity, joining multiple disciplines, and incorporating scientific data meticulously will surely help researchers overcome the challenge of observational data and find solutions to previously unresolved scientific issues.

A guiding principle of this thesis is to understand that confounding variables act as a main barrier for validity and reliability of multidisciplinary research. Confounding factors play a disturbing role in various disciplines such as epidemiology, psychology, sociology, and economics. Therefore, causal inferences in research [162, 164] should be made methodologically strict [177]. Ethical and

practical problems of confounding in research practice are an important concern, because the implications of uncontrolled confounding for public policies, clinic practices, and decision-making are often severe. Therefore, this thesis aims at helping researchers to deal with their ethical responsibility by providing tools to identify potential sources of bias and control confounding.

Confounding is the main theme of this thesis, in particular, it focuses on the importance of variables that can produce bias if not properly controlled [92, 36]. Confounding variables are extraneous factors that can obfuscate the relationship between independent and dependent variables of a research study and can lead to wrong conclusions and interpretations of the results. Through a careful analysis of existing measures of confounding, the thesis aims to increase the awareness of confounding problems and to show ways on how to deal with them.

Confounding effects and ways on how to take them into account during statistical analyses have always been crucial topics in research studies and statistical research, and have evolved continuously over time. With time, confounding effects have been recognized as a source of confusion of research results and it became clear that more precise methods of data analysis are needed to identify them. For example, this has been accentuated by Greenland and Morgenstern [92] who have called for the control of confounding variables to ensure the validity and reliability of research.

Despite the numerous complicated issues related to the misleading factors, many can be done to cope with them. For example, the identification and inclusion of covariates in the analyses should be done with care as to aim at getting realistic relations that are more than mere statistical artefacts.

Fundamentally, confounding factors, which are not known, are among the most challenging problems that affect the validity of study results. However, careful scrutiny, critical thinking, and contemplation of any possible confounders are essential to combat analytical complexity and to extract truthful associations. Incorporating control variables into the analytical model, improves the strength of research results and, in addition, also facilitates the development of a comprehensive presentation of existing problems. The systematic incorporation of control variables is one to way to characterize our subject of investigation.

3.1.3 Examples of Confounding

As a first example, take the case of the drug rosiglitazone [111], which is used more frequently for the treatment of type 2 diabetes, but surprisingly shows a significant connection with acute heart attack [18] [55]. Although, on the first sight, it looks as if the correlation between rosiglitazone and acute heart attache is dominating, when scrutinising closer, we would find that in addition to this correlation, there exist confounding factors. Type 2 diabetes itself is one such confounder [100]. The inclusion of type 2 diabetes as a covariate in the analysis will weaken the strength of the association between rosiglitazone and acute heart attache stressing the fact that there is a confounding effect due to this morbidity. This phenomenon shows how important it is to include the outside context for understanding relationships between factors and for studying the interconnectedness of variables which, in turn, can mask and obscure the research outcomes.

Another well-known example of the confounding effect encountered in epidemiological studies is in studies that focus on understanding the link between

[18] myocardial infarction

coffee consumption and health. As a case in point, researchers should note the positive relationship between coffee intake and the possibility of heart disease [94, 144]. In spite of it, when a more detailed analysis is carried out, the relationship between heavy coffee drinkers and non-coffee drinkers reveal that the heavy coffee drinkers always tend to smoke cigarettes at higher frequencies [96, 28, 224]. In this case, the smoking status is a confounding variable [85] because it is connected both with coffee consumption and the risk of cardiovascular disease. Not considering the smoking status can cause bias, leading to the perception that coffee consumption increases the risk of heart disease, whereas smoking is actually the more severe risk factor.

Another classic example of confounding stems from educational research, in studies that examine the impact of small class sizes on academic performance of students [167, 199]. Suppose a research which shows that the students who attend a class with 20 or fewer students achieve better grades than students who attend a class with more than 20 students. However, when this situation is examined in more detail, this fact is more closely connected to the small classes present in more affluent schools that have well-equipped laboratories, libraries, and experienced teachers [167, 219, 103]. Confounding variables reflecting the socioeconomic status of students affect both class sizes and student achievements. If socioeconomic status is not taken into account, people might assume that the size of the class itself is the reason for the positive correlation between class size and academic achievement, whereas, in fact, socioeconomic inequalities are the most relevant factor.

3.1.4 Definition of Confounding

Confounding occurs when the true effect of an exposure on an outcome is obscured or distorted by one or more extraneous factors, known as confounders. Traditionally, a confounder is treated as a proxy for confounding. To address this issue, researchers seek to identify and adjust for confounders.

A confounder must satisfy three criteria[19]:

1. It must be a risk factor for the outcome.
2. It must be associated with the exposure within the study population,i.e., it must effect the exposure (directly or indirectly via further variables).
3. It must not be effected by the exposure (directly or indirectly via further variables), i.e., it must not be an intermediary step in the causal pathway from the exposure to the outcome.

Confounders can play two roles in scientific studies: they can be seen as troublemakers as well as troubleshooters. As troublemakers, confounders are seen as the underlying causes of confounding. This perspective stems from the belief that identifying confounders pinpoints the causes of confounding. As troubleshooters, they are viewed as variables that can help mitigate confounding.

[19] For example, Jager et al. [109] have stated it as follows: "[...] to be a potential confounder, a variable needs to satisfy all three of the following criteria: (1) it must have an association with the disease, that is, it should be a risk factor for the disease; (2) it must be associated with the exposure, that is, it must be unequally distributed between exposure groups; and (3) it must not be an effect of the exposure; this also means that it may not be part of the causal pathway." [109]

This dual role is based on the idea that by controlling confounders, which are thought to induce confounding, one can effectively manage confounding itself. Unfortunately, research has indicated that merely meeting the above three criteria does not necessarily qualify a variable as a confounder, nor does controlling such variables guarantee the control of confounding. This suggests that these criteria should be considered characteristic properties rather than a definitive identification of confounders.

Confounding significantly increases the difficulty of causal inference [159, 162, 163, 165, 164]. The name confounding comes from the Latin verb 'confoundere,' which means mixing [150], and it is indeed often described as mixing of effects [93, 226]. According to [231], p. 110, the "earliest published account of an experiment in which certain interactions were confounded" [231] appears in a paper on an experiment with confounded interactions by Eden & Fisher [68] in 1929.

Confounding is one of the main concepts used in statistics to represent the association between multiple variables. The association between two variables are confused or corrupted by the presence of a third variable, which is then called confounding. In other words, a confounder (the third variable) influences both dependent (target) and independent (influencing) variables to generate misleading associations or correlations between them. The presence of a confounder can make it seem as if there is a correlation between the influencing and target variables, or it may hide or obscure the true correlation between them.

3.1.5 Causality vs. Correlation

Causality and correlation are fundamental concepts in research, but they represent distinct phenomena with crucial implications for understanding relationships between variables. While correlation describes the statistical association [102] between two variables, causality implies a direct cause-and-effect relationship, where changes in one variable directly influence changes in another. However, distinguishing between causality and correlation can be challenging, especially in observational studies where confounding factors may obscure the true nature of the relationship. Sections 3.1.5.1 and 3.1.5.2 aim to explore the differences between causality and correlation, whereas Section 3.1.5.3 aims at highlighting how confounding can blur these lines and offering insights into methods for disentangling causal relationships from mere associations.

3.1.5.1 Correlation

Correlation refers to the degree to which two variables are related or move together in a systematic manner. It quantifies the strength and direction of the relationship between variables but does not imply causality. Concrete correlation coefficients, such as the Pearson correlation coefficent ρ [166] or Spearman's rank correlation coefficient [203], measure the extent to which changes in one variable correspond to changes in another. A correlation coefficient close to +1 indicates a strong positive correlation, while a coefficient close to -1 indicates a strong negative correlation. A coefficient near zero suggests little to no correlation between the variables.

For example, consider a study examining the relationship between hours of study and exam scores among students. A positive correlation between these variables would indicate that students who study more hours tend to achieve

higher exam scores, while a negative correlation would suggest the opposite. However, correlation alone cannot determine whether studying more hours causes higher exam scores or whether other factors, such as intelligence or motivation, influence both variables.

3.1.5.2 Causality

Causality goes beyond mere association by asserting that changes in one variable directly give rise to changes in another. As such, pure causality can only be observed in experiments with highly accurately designed experiments conditions, i.e., under laboratory conditions. A concept that is usually associated with causality is temporal precedence. Temporal precedence refers to the notion that the cause must precede the effect in time [174].

In experimental research, e.g., in the field of epidemiology, a standard way to control causality is through randomized controlled trials (RCTs), where participants are randomly assigned to different treatment groups. Randomization helps minimizing confounding variables and allows researchers to draw causal inferences about the effect of the treatment on the outcome. However, in observational studies, establishing causality is more challenging due to the presence of confounding variables that may distort the observed associations between variables. Lastly, ruling out confounding variables involves controlling for alternative explanations and ensuring that the observed relationship is not due to extraneous factors.

3.1.5.3 How Confounding Distorts Causal Inference

Confounding can distort causal inference in several ways. First, confounding can create the illusion of a causal relationship where none exists. This occurs when the observed association between the exposure and the outcome is actually due to the confounding variable. For example, suppose a study that finds a positive correlation between the consumption of sugary beverages and the risk of obesity. Without controlling for factors such as overall diet quality or physical activity levels, the study may erroneously conclude that sugary beverage consumption directly causes obesity whereas, in reality, dietary habits and lifestyle factors are confounding the association.

Second, confounding can mask a true causal relationship by obscuring the observed association between the exposure and the outcome. This occurs when the confounding variable is associated with both the exposure and the outcome but is not accounted for in the analysis. For instance, suppose a study investigates the relationship between socioeconomic status and cardiovascular disease risk. If the study fails to control for lifestyle factors such as smoking, diet, and physical activity, which are more prevalent among individuals of lower socioeconomic status and also independently contribute to cardiovascular disease risk, it may underestimate the true impact of socioeconomic status on cardiovascular health.

Third, confounding can lead to biased estimates of the strength or direction of the observed association between the exposure and the outcome. This occurs when the confounding variable distorts the relationship between the exposure and the outcome, making it appear stronger or weaker than it actually is. For example, suppose a study examines the association between alcohol consumption and liver disease risk. If the study includes individuals with pre-existing liver conditions or who are heavy drinkers, it may overestimate the true impact of alcohol

consumption on liver disease risk, as these individuals are more likely to develop liver disease regardless of their alcohol intake.

Confounding presents a significant challenge in research, as it can distort the observed relationship between an exposure and an outcome, leading to inaccurate conclusions about causality. By understanding how confounding operates and employing appropriate methods for controlling it, researchers can improve the validity and reliability of their study findings.

3.1.6 A Historical Example

Smoking and lung cancer are among the most famous historical examples of the confounding effect. At the middle of the 20th century, a pronounced relationship between cigarette smoking and lung cancer prevalence started being noted by the researchers. Most of the early studies on this association were complicated by a number of issues, which include the tar content in cigarettes and the prevalence of smoking in some demographic groups.

During the 50s, a series of epidemiological studies conducted by Richard Doll and Bradford Hill [56] in the UK, as well as Ernest Wynder and Evarts Graham [228] in the US, amongst many others, proved beyond doubt how deeply cigarette smoking is linked to lung cancer. These studies discovered that smokers were at much higher risk of getting lung cancer when compared to non-smokers and, therefore, triggered widespread public health worries and campaigns against smoking.

Yet, early results were obscured by the chemical composition of cigarettes and the complexity of smoking habits. A single cigarette consists of thousands of chemical compounds, including tar, nicotine, and various carcinogens, that can be present in different amounts depending on the brand and manufacturing process. The tar content was the initial confounding factor that researchers focused on, suggesting that it might be the most significant carcinogenic element of cigarettes.

Furthermore, the occurrence of smoking differed between various demographic groups, where men, people with lower socioeconomic status, and certain occupational groups had more tobacco users. The demographic diversity of smokers is as an external factor that complicates the interpretation of study results as it brings in various confounding variables such as occupation and lifestyle factors.

Therefore, in early studies, epidemiological evidence was not yet settled due to these confounding factors, and further research was needed to provide more evidence for the hypotheses of a cause-and-effect relationship. Exactly in that vain, Richard Doll et al. [57, 58, 59, 61, 60] initiated a longitudinal cohort study, which provided strong evidence of the dose-response relationship between smoking intensity and lung cancer risk. These studies were conducted on large cohorts of smokers who were observed over time for their smoking behaviour and health outcomes; and all of the studies concluded that heavy smokers had a significantly higher risk of developing lung cancer than non-smokers and lighter smokers. The later studies in the longitudinal research of Richard Doll et al. were focused on separating the particular ways through which smoking contributes to lung cancer development. Experimental studies with laboratory animals and cell cultures have proved that certain tobacco smoke ingredients such as *polycyclic aromatic hydrocarbons* (PAHs) [237] and nitrosamines [105] are carcinogenic.

As the empirical evidence for the relation between smoking and lung cancer became stronger and more widespread, public health interventions and tobacco

control policies were introduced to minimize the prevalence of smoking and alleviate the burden of smoking-related diseases. Such initiatives included warning signs on the packets of cigarettes, smoking bans, taxation of tobacco and programs for smokers who were willing to quit.

Smoking and lung cancer are a cornerstone example of how confounding variables can mess up the interpretation of research results, which is especially true for observational studies. Although some early confounders included tar content and demographic distribution of smoking behaviour, epidemiological research, along with mechanistic studies [6, 20], resolved the actual causes in the relationship between smoking and lung cancer.

3.1.7 Confounding Across Disciplines

The identification and mitigation of confounding effects are crucial in various disciplines, each offering unique insights and methods to address these challenges. Confounding can significantly skew results and mislead researchers if not properly identified and managed. To fully appreciate the lessons learned across different fields, it is essential to explore how various disciplines approach the issue of confounding, what methodologies they employ, and how these can be synthesized to enhance overall research quality.

3.1.7.1 Epidemiology

In epidemiology, confounding is a well-recognized problem [129, 220, 73, 210, 202] because of the complexity of human health and disease processes. Researchers in this field are highly aware of the need to control for confounders to ascertain true causal relationships between exposures and outcomes. One common method is stratification, where data is divided into subgroups that are homogeneous concerning the confounder. By analyzing these subgroups separately, epidemiologists can more accurately assess the exposure-outcome relationship. Another technique is multivariable regression analysis, which adjusts for multiple confounding variables simultaneously, see Section 4. Advanced methods, such as propensity score matching [178], also play a significant role. This technique involves creating a *statistical twin* for each subject based on their probability of receiving a particular treatment or exposure, thereby isolating the effect of the treatment from confounding variables. The rigorous use of these techniques has demonstrated the importance of meticulous planning and the adoption of robust statistical methods to mitigate confounding.

3.1.7.2 Economics

In economics, confounding is often encountered in observational studies where controlled experiments are impractical. Economists frequently use the instrumental variables (IV) approach [208, 141] to address this issue. An instrumental variable is a variable that is correlated with the potentially confounding independent variable but is not correlated with the errors in the dependent variable equation [141]. This approach helps to isolate the causal impact of the independent variable on the dependent variable. Another method employed is the difference-in-differences (DiD) approach [1, 133], which compares the changes in outcomes over time between a treatment group and a control group. This method can help account for confounding factors that vary over time but are constant between groups [213]. The lessons from economics highlight the necessity of innovative

statistical tools and the creative use of naturally occurring experiments to discern causal relationships amid potential confounders.

3.1.7.3 Psychology

In psychology, the issue of confounding is addressed through both experimental design and statistical control. Randomized controlled trials (RCTs) are considered the gold standard for eliminating confounding by randomly assigning participants to different conditions, thus ensuring that confounders are equally distributed across groups. However, when RCTs are not feasible, psychologists often rely on techniques like *analysis of covariance* (ANCOVA) [147], which adjusts the dependent variable for the effects of confounders. Structural equation modeling (SEM) [207] is a sophisticated approach that allows for the examination of complex relationships among variables, including the control of multiple confounders simultaneously. Psychology's emphasis on rigorous experimental control and advanced modeling techniques provides valuable insights into maintaining internal validity in the face of potential confounders.

3.1.7.4 Social Sciences

In social sciences, particularly sociology and political science, the identification and mitigation of confounding are addressed through careful study design and sophisticated analytical techniques. Social scientists often use panel data [14, 32], which follows the same subjects over multiple time points, to control for unobserved confounders that are constant over time. They also employ fixed-effects models [168], which control for all time-invariant characteristics of the individuals, thus reducing the risk of confounding. Additionally, natural experiments [67], where external events or policies create conditions similar to randomized experiments, are utilized to infer causality [52]. The social sciences underscore the value of leveraging natural variations and longitudinal data to control for confounding factors.

3.1.7.5 Environmental Science

In environmental science, researchers frequently encounter confounding factors when studying the impacts of pollutants or climate change on ecosystems. One common approach to mitigate confounding in this field is the use of longitudinal studies [42], which track changes over time and can help differentiate between correlation and causality [121]. Additionally, geographic information systems (GIS) [232] are often employed to control for spatial confounding by allowing researchers to analyze data in relation to specific locations and account for geographic variability. The use of ecological models that incorporate multiple variables and simulate different scenarios also aids in understanding the complex interplay of factors influencing environmental outcomes. Environmental science thus teaches the importance of long-term data collection and spatial analysis in managing confounders.

3.1.7.6 Biostatistics

Biostatistics offers a wealth of methods for dealing with confounding in medical research and beyond. Techniques such as logistic regression, Cox proportional hazards models [114], and generalized estimating equations (GEEs) [101] are commonly used to adjust for confounders in complex datasets. Biostatisticians

also advocate for the use of sensitivity analysis to assess the robustness of results to potential unmeasured confounding. The field of biostatistics highlights the importance of advanced statistical methods and rigorous sensitivity checks to ensure the validity of research findings.

3.1.7.7 Education

In education, confounding can arise from a myriad of student, teacher, and school-related factors. To address this, researchers often employ hierarchical linear modeling (HLM) [171], which accounts for the nested structure of educational data (students within classes within schools). Randomized field trials are also used to evaluate educational interventions while controlling for confounding variables. Moreover, the use of propensity score analysis [178] helps to create comparable groups in observational studies. Education research illustrates the significance of considering the hierarchical nature of data and employing robust statistical techniques to mitigate confounding.

3.1.7.8 Public Health

From the domain of public health, the concept of the social determinants of health emphasizes the need to account for a wide range of confounding factors, such as socioeconomic status, education, and access to healthcare. Public health researchers often use community-based participatory research (CBPR) [107, 108, 205] methods, which involve the community in the research process to better understand and control for confounding factors relevant to the population being studied [184]. Additionally, the use of big data analytics in public health [122] enables the integration and control of vast amounts of potential confounders. Public health teaches the importance of comprehensive data collection and community involvement in identifying and mitigating confounders.

3.1.7.9 Marketing

In marketing, researchers frequently deal with confounding variables related to consumer behaviour and market trends. Techniques such as experimental designs, including A/B testing [125, 172] and randomized controlled trials [206, 173], are used to control for confounders in marketing campaigns. Econometric models, such as multivariate regression [79] and structural models, help to isolate the effects of marketing variables on consumer outcomes. A method in widespread use in marketing research is conjoint analysis [90]. Again, conjoint analysis studies can be severely affected by confounding effects [215, 214]. Marketing research demonstrates the value of controlled experimentation and sophisticated econometric techniques in addressing confounding.

3.1.7.10 Computer Science

In the field of computer science, especially in artificial intelligence (AI) [123] and machine learning , confounding can lead to biased models and inaccurate predictions [236]. Techniques such as cross-validation [209], regularization, and the use of balanced datasets are employed to mitigate confounding effects. Feature selection and engineering also play crucial roles in ensuring that models are not unduly influenced by confounders. Moreover, causal inference methods, such as causal trees and causal forests [211], are increasingly used to better understand

and control for confounding variables. Computer science highlights the importance of methodological rigor and the integration of causal inference techniques in dealing with confounding.

3.1.7.11 Conclusion on Confounders Across Disciplines

Synthesizing lessons from these diverse fields, several overarching themes emerge regarding the identification and mitigation of confounding effects.

First, the importance of study design cannot be overstated. Randomization, when feasible, remains one of the most effective ways to control for confounders. In the absence of randomization, techniques such as stratification, matching, and the use of control groups are essential. These methods ensure that confounders are evenly distributed across comparison groups, allowing for more accurate estimates of causal relationships.

Second, the use of advanced statistical techniques is crucial in adjusting for confounders. Methods such as multivariable regression, instrumental variables, fixed-effects models, and propensity score matching provide powerful tools for researchers to account for multiple confounding variables simultaneously. These techniques, when properly applied, can significantly enhance the validity of research findings.

Third, longitudinal data and repeated measures offer valuable opportunities to control for confounders that vary over time. By tracking the same subjects over multiple time points, researchers can differentiate between correlation and causality more effectively. This approach is particularly useful in fields like environmental science, public health, and social sciences, where confounding factors may change over time.

Fourth, the incorporation of spatial analysis and hierarchical models helps to account for confounding related to geographic and nested data structures. Geographic information systems and hierarchical linear modeling allow researchers to control for spatial and hierarchical confounders, respectively, providing a more nuanced understanding of the relationships under study.

Fifth, the involvement of the community and stakeholders in the research process, as seen in public health and education research, can enhance the identification and control of confounders. Community-based participatory research ensures that the perspectives and knowledge of those affected by the research are incorporated, leading to a more comprehensive understanding of potential confounding factors.

Finally, the integration of causal inference methods into traditional statistical and machine learning techniques offers a promising avenue for addressing confounding.

3.2 The Familiar Ad-Hoc Method for Confounding Adjustment

This section is about a well-known technique for adjusting confounding effects, which is in widespread in the statistics community and has been used in a plethora of studies of in various domains. According to Judea Pearl: *"If we do have measurements of the third variable, then it is very easy to deconfound the true and spurious effects. For instance, if the confounding variable Z is age, we compare the treatment and control groups in every age group separately. We can then take an average of the effects, weighting each age group according to its percentage in the target population. This method of compensation is familiar to all*

statisticians; it is called »adjusting for Z« or »controlling for Z.«" [165].

We call this technique *familiar ad-hoc method for confounding adjustment*, or just *ad-hoc method* for short. We also call it *standard categorical adjustment* or *categorical adjustment* for short, as the influencing factors are always categorical.

Usually, this adjustment technique is presented for categorical target factors, actually for the impact on a single instance of a categorical factor, i.e., for the impact on an event, in terms of conditional probabilities. However, we introduce the technique in a generalized form, i.e., for the impact on numerical target factors utilizing conditional expectations.

We proceed as follows. First, we explain categorical adjustment in Section 3.2.1. Then, we briefly discuss Judea Pearl's do-calculus as important background literature in Section 3.2.2.

3.2.1 Standard Categorical Adjustment

Given a random variable $z : \Omega \longrightarrow \mathbb{R}_0^+$ and a partition[20] $p_1, \ldots, p_n$ of Ω, we have the following:

$$\mathsf{E}(z) = \sum_i \mathsf{P}(p_i)\mathsf{E}(z|p_i) \tag{27}$$

(27) is called *law of total expectations*. Given a random categorical variable $y : \Omega \longrightarrow \{v_1, \ldots, v_n\}$, we have that $(y = v_1), \ldots, (y = v_n)$ forms a partition of Ω. In terms of y, the law of total expectation therefore shows as:

$$\mathsf{E}(z) = \sum_i \mathsf{P}(y = v_i)\mathsf{E}(z|y = v_i) \tag{28}$$

Together with a further event x, we have that

$$\mathsf{E}_x(z) = \sum_i \mathsf{P}_x(y = v_i)\mathsf{E}_x(z|y = v_i) \tag{29}$$

Due to the fact that $\mathsf{P}_x(a|b) = \mathsf{P}(a|x, b)$ for any events a and b, we can rewrite (29) as

$$\mathsf{E}(z|x) = \sum_i \mathsf{P}(y = v_i|x)\mathsf{E}(z|x, y = v_i) \tag{30}$$

Given a numerical random variable $z : \Omega \longrightarrow \mathbb{R}$, called *target variable (or dependent variable)*, an event x, called *impacting variable (or impact factor)*, and a categorical random variable $y : \Omega \longrightarrow \{v_1, \ldots, v_n\}$, called the *confounder (or confounding variable, or confounding factor)*, we define the *adjustment of the conditional expectation* $\mathsf{E}(z|x)$ *to the (impact of the) confounder* y, denoted by $\widehat{\mathsf{E}^y}(z|x)$, as follows:

$$\widehat{\mathsf{E}^y}(z|x) = \sum_{i \leq n} \mathsf{P}(y = v_i)\mathsf{E}(z|x, y = v_i) \tag{31}$$

Note that the adjustment of the expectation $\mathsf{E}(z|x)$ is achieved by the *coercion* of $\mathsf{P}(y = v_i|x)$ in (30) to the value $\mathsf{P}(y = v_i)$ in (31). This coercion is the only (but

[20] As usual a *partition* of a set A is a collection of subsets $A_1 \subseteq A, \ldots, A_n \subseteq A$ of A such that $A = A_1 \cup \cdots \cup A_n$ and $A_i \cap A_j = \emptyset$ for any two distinct $A_i \neq A_j$.

crucial) difference between the expectation $\mathsf{E}(z|x)$ and the adjusted expectation $\widehat{\mathsf{E}^{y}}(z|x)$.

Accordingly, given a numerical random variable $z : \Omega \longrightarrow \mathbb{R}$, an event x, and a series $\vec{y}$ of confounding categorical random variables $y_1 : \Omega \longrightarrow I_1$ to $y_m : \Omega \longrightarrow I_m$, we first define the multivariate random variable $y : \Omega \longrightarrow I_1 \times \ldots \times I_m$ as usual, i.e., $\mathsf{P}(y = \langle v_1, \ldots, v_m \rangle) = \mathsf{P}(y_1 = v_1, \ldots, y_m = v_m)$; then, we define the adjustment $\widehat{\mathsf{E}^{\vec{y}}}(z|x)$ as follows:

$$\widehat{\mathsf{E}^{\vec{y}}}(z|x) = \widehat{\mathsf{E}^{y}}(z|x) \tag{32}$$

In [65, 62], partial conditionalization [63] has been generalized from partial conditional probabilities [63] to partial conditional expectations. Then, the adjusted expectation has been explained as a partial conditional expectation. In [65, 62], the quotient of the adjusted expectation $\widehat{\mathsf{E}^{y}}(z|x)$ and the marginal expectation $\mathsf{E}(z)$ has been called the *genuine impact of x (onto z)*. In terms of association rule mining (ARM) [4], the genuine impact could be also called *adjusted lift*, as the quotient $\mathsf{P}(z|x)/\mathsf{P}(x)$ is known as *lift* in ARM (for the specialized cases that z and x are events).

3.2.2 Judea Pearl's Do-Calculus

Judea Pearl, a pioneer in the field of artificial intelligence and statistics, has significantly advanced our understanding of causal inference, in particular, also in regards to confounding and confounding adjustment. Pearl's methods for addressing confounding have contributed to how researchers approach causal questions, offering a robust framework to disentangle complex relationships in data.

A key concept in Pearl's framework is the do-operator, denoted as $do(X = x)$. This operator allows to simulate interventions by setting a variable X to a specific value x, effectively breaking the usual causal pathways and isolating the direct effect of X on other variables. The do-operator distinguishes between observational and interventional distributions, which is crucial for causal inference.

For instance, in an observational study, we observe the distribution $\mathsf{P}(Y|X)$, which might be confounded by other variables. By applying the do-operator, we aim to estimate $\mathsf{P}(Y|do(X = x))$, the distribution of Y when X is set to x through an intervention, thus eliminating the influence of confounders.

The primary goal of causal inference is to identify causal effects from observational data. Using do-calculus, Pearl developed algorithms that determine whether a causal effect is identifiable and, if so, provide a method to compute it. One of the key algorithms is the ID algorithm, which systematically applies the rules of do-calculus to derive expressions for causal effects in terms of observational quantities.

Eventually, to estimate the causal effect using observational data, we can use the Judea Pearl's adjustment formula[21]:

$$\mathsf{P}(z|\,\mathrm{do}(x)) = \sum_i \mathsf{P}(y_i)\mathsf{P}(z|x, y_i) \tag{33}$$

See, how the r.h.s. of (33) meets the r.h.s. of (31) by generalizing the probabilities $\mathsf{P}(z|x, y_i)$ to expectations $\mathsf{E}(z|x, y_i)$ (with the only (notational) difference, that

[21] In [160], Eqn. (14), Judea Pearl uses $\mathsf{P}_x(z)$ to denote $\mathsf{P}(z|\,\mathrm{do}(x))$. In this study, we use, as usual, both $\mathsf{P}_x(z)$ as an alternative notation for the conditional probability $\mathsf{P}(z|x)$.

we use an instance of a random variable $y = v_i$ in (31) to denote an instance of the partition y_i in (33)). In that sense, (31) is a generalization of (33) from a target event z to a numerical target random variable.

3.3 Oaxaca-Blinder Decomposition for Confounding Adjustment

This section provides an overview of several papers available within the last 40 years beginning with Oaxaca (1973) [156] and Blinder (1973) [29][22]. Since these first efforts, the technique has been increasingly used as decomposition approach in service of finding impacts in studies, in particular in the field of labour economics [78, 72].

For example, health disparities refer to differences in the quality of health care provided to individuals of different status in society. Such disparities should not exist and are unfair [35]. Unfortunately, despite the number of federal and state-directed campaigns that have been implemented in recent decades to eliminate the disparity, these disparities have been known to persist [22, 201].The Oaxaca-Blinder decomposition approach is used to identify such disparities. It helps researchers disaggregate these differences with respect to various variables.

In confounding analysis, sources of inference moderate target variables and adjust for confounding. The analysis aims to examine the impact of an exposure on an outcome by teasing out direct and indirect effects of the influencing variables on the exposure.

The described disparities issues, therefore, emphasize a severe complexity that continues to hinder improvements in the society. Finally, it would be pertinent to note that broad and structured approaches are needed to analyse the fundamental reasons for inequalities.

3.3.1 Theoretical Foundations

The Oaxaca-Blinder decomposition which is frequently applied in labor economics enables to explain how various characteristics of groups influence the differences in results, including wages, between various demographic groups. Wage discrimination is understood as the act of paying employees with similar skills and experiences different wages depending on factors such as gender and race, instead of their performance. This is important because it defines the need to know and bring change in the wage inequalities.

Oaxaca–Blinder decomposition approaches the problem by decomposing the existing average wage differentials between groups into two components. The first component (the so-called 'explained part', also called 'endowment effect') deals with the differences in qualifications, which the model explains [157]. These include personal characteristics such as education, experience, and skills that can explain the difference in wages. The second component (the so-called 'unexplained part', also called the 'coefficient effect') shows variation in the model structure that is not attributed to the qualifications [157]. Consequentially, the second component is commonly blamed on discrimination in the labour market, as it captures inequalities in earnings that cannot be explained by observable characteristics.

[22] Oaxaca-Blinder decomposition and Blinder-Oaxaca decomposition are used synonymously throughout the literature, and seemingly equally often – e.g., with 785 hits (Oaxaca-Blinder) and 729 hits (Blinder-Oaxaca) respectively on Scopus, as of 18th Feb. 2024.

In general, it is universally accepted that wage discrimination exists when men receive higher wages than women, regardless of their skills and performance. Oaxaca (1973) [156] developed the notion of wage discrimination formally and described the *discrimination coefficient* (D) as a measure of this wage discrimination as follows:

$$D = \frac{\frac{X_h}{X_l} - \left(\frac{X_h}{X_l}\right)^*}{\left(\frac{X_h}{X_l}\right)^*} \tag{34}$$

where:

- D is the discrimination coefficient.
- X_h and X_l are the average characteristics of the higher and lower categories, respectively.
- $\frac{X_h}{X_l}$ is the observed relationship between the wages of the higher and lower categories.
- $\left(\frac{X_h}{X_l}\right)^*$ it is the wage ratio between the higher and lower categories when there is no discrimination.

Oaxaca explains that in the absence of discrimination, the structure of wage factors would impact womens' wages in the same way as mens' wages [156]. However, Blinder also discussed that it is common knowledge that Whites were paid more than Blacks as men are paid better than women. So, the presumption that wages reflected marginal productivity did not apply well.

Therefore, Blinder suggests that for calculating the decomposition, it makes sense to conduct certain estimations as follows. In the first step, a regression analysis is conducted to estimate the earnings equations separately for the higher (h) and lower (l) categories:

$$Y_i^h = \beta_0^h + \sum_{1 \le j \le n} X_{ji}^h \beta_j^h + \epsilon_i^h \tag{35}$$

$$Y_i^l = \beta_0^l + \sum_{1 \le j \le n} X_{ji}^l \beta_j^l + \epsilon_i^l \tag{36}$$

where:

- Y_i is the earned salary for each observations (Target variable).
- $X_{ji}^h and X_{ji}^l$ are the explanatory variables of the higher and lower categories, respectively (influencing variable).
- β_j is the slope of the regression line, β_0 is a constant.
- ϵ is the error term.

One of the most straightforward ways to determine discrimination is by calculating the difference between the equation of the low-wage group and the high-wage group. However, Blinder (1973) noted that the unexplained portion of the

income difference arises not only from the coefficient difference but also from the variations in the average characteristics of the minority group. Blinder (1973) revised his equation to reflect this consideration as follows:

$$\Delta = \overline{Y}^{h} - \overline{Y}^{l} \tag{37}$$

$$\overline{Y}^{h} = \widehat{\beta}_0^{h} + \sum_{1 \le j \le n} \overline{X}_j^{h} \widehat{\beta}_j^{h} \tag{38}$$

$$\overline{Y}^{l} = \widehat{\beta}_0^{l} + \sum_{1 \le j \le n} \overline{X}_j^{l} \widehat{\beta}_j^{l} \tag{39}$$

Equations (37), (38) and (39) provide information about the profiles of higher and lower categories by explaining the distribution in between mean wage and numerous observed average traits between the two categories.

To work with the decomposition method, we need to identify a non-discriminatory benchmark [26]. Here, we consider the higher category (h) as the benchmark, and the β values of the equation (39) of the lower category are replaced with the higher category as follows:

$$\overline{Y}^{l'} = \widehat{\beta}_0^{h} + \sum_{1 \le j \le n} \overline{X}_j^{l} \widehat{\beta}_j^{h} \tag{40}$$

Now, the decomposition of the difference between the higher and the lower category is as follows:

$$\Delta = (\overline{Y}_i^{h} - \overline{Y}^{l'}) + (\overline{Y}^{l'} - \overline{Y}^{l}) \tag{41}$$

$$\Delta = \underbrace{\left(\sum_{1 \le j \le n} \widehat{\beta}_j^{h} (\overline{X}_j^{h} - \overline{X}_j^{l}) \right)}_{\textit{explained part}} + \underbrace{\left(\widehat{\beta}_0^{h} - \widehat{\beta}_0^{l} + \sum_{1 \le j \le n} \overline{X}_j^{l} (\widehat{\beta}_j^{h} - \widehat{\beta}_j^{l}) \right)}_{\textit{unexplained part}} \tag{42}$$

In (42), the first part represents the explained part (endowment effect), and the second part represents the unexplained part (coefficient effect).

3.3.2 Limitations and Criticisms

The econometric regression analysis technique proposed by Blinder and Oaxaca to deduce the causes of the gender wage gap has been subject to considerable criticism that revolves around the model specification and the choice of the independent variables [176].

According to Rosenzweig and Morgan [179], the use of age and age squared instead of work experience and squared experience in the structural equation developed by Blinder creates a differential bias in estimated returns to education for men; and, therefore, it is likely that the Blinder results reflect an exaggeration of the size of the component of education that explains the difference in income between men and women attributable to discrimination.

In [157], Oaxaca and Ransom report, that Blinder's decomposition method [29] to separate the contribution of the discrimination was defective in the presence of a set of fictional variables, since the magnitude of the estimated constant term depends on the reference out-of-sample group.

In [113], Jones brings examples, in which Oaxaca-Blinder decomposition shows highly arbitrary results for the explained and unexplained part. He identifies "arbitrary decisions about how to impose a metric on the variables" [113] as the reason for this arbitrariness.

According to David Madden [137], the Oaxaca-Blinder standard approach tends to underestimate the discrimination degree if there are differences in access to endowments that are rewarded in the labor market, e.g., if men have better access to higher education than women, or even if, ceteris paribus, men are more likely to work than women.

3.3.3 Conclusion on Oaxaca-Blinder Decomposition

The Oaxaca-Blinder decomposition is a powerful tool for understanding wage differentials, but it has several limitations that need to be considered. These include linear assumptions, endogeneity, omitted variable bias, the choice of non-discriminatory wage structure, interpretation of the unexplained component, aggregation issues, lack of dynamic analysis, and context-specific factors. Researchers must be cautious in interpreting the results and acknowledge these limitations to avoid misleading conclusions.

Despite these criticisms, the Oaxaca-Blinder decomposition remains a valuable method for labor economists, providing a structured approach to dissect wage differentials and identify potential sources of inequality. Future research should focus on addressing these limitations, possibly through the development of more sophisticated models that incorporate nonlinearities, endogenous variables, and dynamic elements, and by ensuring a comprehensive set of variables that reflect the complex nature of labor markets.

4 Utilization of Linear Regression for Confounding Adjustment

Linear regression is a standard, widespread statistical model [87, 112, 149]. When investigating the impact of an influencing factor (exposure, independent factor) onto a target factor[23] (outcome, dependent factor), the experienced analyst is aware of potential confounding [182, 183, 165] and aims at controlling confounding effects as effectively as possible.

In the context of linear regression, controlling confounding effects means to add further influencing, potentially *confounding factors* to the analysis [143, 170], see [235, 153, 158, 135, 132, 136] for some recent concrete example studies. If the added factors actually have an impact on the target value, the coefficient estimator of the *primary influencing factors* (the factors that we are actually interested in) change [124, 37, 130, 185]. The coefficient estimator of a primary influencing factor analysed in the context of additional confounding factors is then usually called *adjusted coefficient*, whereas the the coefficient estimator of a primary influencing factor analysed without confounders added is usually called *crude coefficient*. Furthermore, we call the coefficient estimator of a confounding factor a *confounder coefficient*. [170] give the following example: *"For example, in a research seeking for relationship between LDL cholesterol level and age, the multiple linear regression lets you answer the question: How does LDL level vary with age, after accounting for blood sugar and lipid (as the confounding factors)?* [..] *The process of accounting for covariates is also called adjustment* [..] *and comparing the results of simple and multiple linear regressions can clarify that how much the confounders in the model distort the relationship between exposure and outcome."*

The difference between a crude coefficient and its corresponding adjusted coefficient is usually called *confounding effect*, no matters how this difference is actually measured, e.g., as *increase factor*, *percentage difference*, or, simply as the (arithmetic) difference of the two estimators.

The aim of this chapter is to exactly explain confounding effects in terms of the involved coefficients. To do so, we need the estimations of further relationships, i.e., between the primary influencing factors and each of the additionally added confounding factors. We call the coefficient estimators stemming from such analyses *latent coefficients*[24] – see Table 5 for an overview of all coefficient estimators involved in our analysis.

[23] We prefer to talk about influencing factors and target factors over talking about independent factors and dependent factors, as usual. In case of several influencing factors, these are, in general, not independent among each other, neither any of them is independent from the so-called dependent factor, i.e., it makes actually little sense to distinguish between them as independent and dependent factors. When distinguishing between influencing factors and target factors, we do not want to express or impose any notion of causality. We distinguish between influencing factors and targets factor only to express a direction of analysis between factors. The relationship between the factors is to be understood purely stochastically, as data. Only later, whenever it comes to the interpretation of a statistical model or machine learning model, concepts of causality become relevant in the argumentation, still remaining outer-mathematical concepts, either appealing to intuition or drawn from a specific expert domain or experiment. Still, however, from a strict viewpoint the analysis remains purely stochastically.

[24] The choice of *latent coefficient* might not be the best, as neither the primary influenc-

Table 5: Coefficient estimators involved in our analysis of confounding adjustment in linear regression.

crude coefficient	Coefficient estimator of a primary influencing factor before adjustment, i.e., before adding confounding factors to the analysis. A primary influencing factor is subject to the original analysis problem, whereas a confounding factor is is added to the analysis for the purpose of controlling confounding.
adjusted coefficient	Coefficient estimator of a primary influencing factor after adjustment, i.e., after adding confounding factors to the analysis.
confounder coefficient	Coefficient estimator of a confounding factor.
latent coefficient	Coefficient estimator of a primary influencing factor in a different role, i.e., as impacting a confounding factor as target factor instead of the original target factor.

Our analysis and findings are based on a conjecture that is well-known from the literature [88, 83]:

- A crude coefficient equals its adjusted coefficient plus the weighted sum of all confounding coefficients, each weighted by its latent coefficient in regard of the crude coefficient's factor.
- The conjecture is noise-independent, i.e., holds independent of any concrete noise ϵ in the linear regression problem, i.e., it holds for the estimators found after least-squares optimization and is the same for each noise.

Our discussion aims at improving the explainability of linear regression models, in general, and in the contexts of the domains where they are utilized. Confounding effects are typically dealt with by rule of thumbs such as a 10%-rule. In practice, little effort is usually spent to investigate the significance of confounding effects. Here, our findings can help to improve the theoretical basis and to improve the maturity of discussions. For example, it follows immediately from our conjecture, that the confounding effect onto a crude coefficient (here: measured simply as arithmetic difference) is a latent-coefficient-weighted sum of all confounding coefficients, i.e., we can explain what the confounding effect is as opposed to merely measuring it. Such explanations can then help to avoid misinterpretation of experimental setups, diagnostic tests, clinical studies*clinical study*, and observational studies in general (interpretations of observations on natural phenomenae, social phenomenae etc.).

We proceed as follows. In Sect. 4.1, we recap, from the literature, a conjecture on the exact value of the omitted variable bias in linear regression. We state the conjecture in the form of two Lemmas, i.e., Lemma 1 for the case of single variable

ing factors for which the latent coefficients are estimated are hidden nor their impact on the confounders is latent in the moment of analysis. Still, we think that the terminology is intuitive as it is in accordance with usual terminology – the impact they represent remains the same and gets latent as soon as the confounders are screened off.

regression and Lemma 2 for the general case of multiple linear regression. In Section 4.2, we discuss an interpretation of the linear regression model by introducing so-called multiplicative edges diagrams and elaborating well-defined linear-regressive scenarios, i.e., error-free/error-prone mediator/confounder scenarios, heavily relying on Lemma 1 and Lemma 2 from Sect. 4.1. In Section 4.3, we discuss two kinds of utilization of linaer regression, i.e., predictions and assessments of interventions. Based on our argumentation in Section 4.2 and Section 4.3 and again the conjecture from Section 4.1, we discuss in how far our findings could contribute to the explainability of linear regression models and their utilization in various contexts in Section 4.4.

4.1 Equations for Screening Off Dimensions in Linear Regressions

Each multiple linear regression problem is based on n given observations

$$(y_i, x_{i0}, x_{i1}, \ldots, x_{im}),$$

where $i \in \{1, \ldots, n\}$ and $y_i, x_{i0}, \ldots, x_{im} \in \mathbb{R}$ are real numbers. The goal is to find regression coeficcients $\beta_0, \ldots, \beta_m \in \mathbb{R}$ that minimize the error function

$$\epsilon(\beta_0, \ldots, \beta_m) = \sum_{i=1}^{n} (\underbrace{\beta_0 x_{i0} + \ldots + \beta_m x_{im}}_{\hat{y}_i} - y_i)^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \ . \tag{43}$$

In vector notation, having vectors $Y = (y_1, \ldots, y_n)^T, X_0 = (x_{10}, \ldots, x_{n0})^T, \ldots, X_m = (x_{1m}, \ldots, x_{nm})^T$ in the n-dimensional Euclidean space $\mathbb{R}^n$, find a linear combination $\hat{Y} = \beta_0 X_0 + \ldots + \beta_m X_m$ that minimizes the Euclidean distance

$$d(\hat{Y}, Y) = \|\hat{Y} - Y\| = \sqrt{\langle \hat{Y} - Y, \hat{Y} - Y \rangle} \ ,$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product of $\mathbb{R}^n$, i.e. $\langle U, V \rangle = u_1 v_1 + \ldots + u_n v_n$ for every two vectors $U = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$ and $V = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$.

In econometric applications of linear regression, it is usually assumed that the vectors $X_0, \ldots, X_m$ are linearly independent, which guarantees the uniqueness of the solution $\beta_0, \ldots, \beta_m$.

Notational Conventions and Remarks:

- As it is common in the linear regression literature, we have chosen to start the index of the column vectors by zero, i.e., we express the linear regression problem in terms of vectors $X_0, \ldots, X_m$ and not in terms of vectors $X_1, \ldots, X_m$.

- In the linear regression literature, it is assumed that the vector X_0 is always set to the all-ones-vector $(1, \ldots, 1)^T \in \mathbb{R}^n$. Then, the regression coefficient β_0 is called *intercept* and all other regression coefficients $\beta_1, \ldots, \beta_m$ are called *slopes*. We call this the *standard linear regression model*. We do not make any such assumption on X_0 in our linear regression model. It would only complicate the discussed Lemmas technically, without benefit. All of our discussion also applies immediately to the standard linear regression model.

- In texts on linear regression, the results β_i of a linear regression optimization are often denoted as $\hat{\beta}_i$, in particular, when discussed as *estimator*. In our lemmas, there is no need for such extra notation, as it is always clear from the lemmas' conditions that all β_i are part of a minimal solution to a respective equation.

The following lemmas are well-known from the literature, see [83], equation (11) on page 521, and [88], equation (17.4) on page 184:

Lemma 1 (Single Variable Case)**.** *Given $Y, X_0, \ldots, X_m \in \mathbb{R}^n$ such that the vectors $X_0, \ldots, X_m$ are linearly independent and*

$$\begin{aligned} \hat{Y} &= \beta_0 X_0 + \ldots + \beta_{m-1} X_{m-1} \text{ minimizes } \|\hat{Y} - Y\| && (44)\\ \tilde{Y} &= \beta'_0 X_0 + \ldots + \beta'_{m-1} X_{m-1} + \beta'_m X_m \text{ minimizes } \|\tilde{Y} - Y\| && (45)\\ \hat{X}_m &= \beta''_0 X_0 + \ldots + \beta''_{m-1} X_{m-1} \text{ minimizes } \|\hat{X}_m - X_m\| && (46) \end{aligned}$$

then for every $i \in \{0, \ldots, m-1\}$ the regression coefficients satisfy the relations:

$$\beta_i = \beta'_i + \beta'_m \beta''_i \tag{47}$$

Lemma 2 (General Case)**.** *Given $Y, X_0, \ldots, X_m \in \mathbb{R}^n$ such that the vectors $X_0, \ldots, X_m$ are linearly independent and*

$$\begin{aligned} \hat{Y} &= \beta_0 X_0 + \ldots + \beta_k X_k \text{ minimizes } \|\hat{Y} - Y\| && (48)\\ \tilde{Y} &= \beta'_0 X_0 + \ldots + \beta'_k X_k + \beta'_{k+1} X_{k+1} \ldots + \beta'_m X_m \text{ minimizes } \|\tilde{Y} - Y\| && (49)\\ \hat{X}_j &= \beta^j_0 X_0 + \ldots + \beta^j_k X_k \text{ minimizes } \|\hat{X}_j - X_j\| \text{ for each } j \in \{k+1, \ldots, m\} && (50) \end{aligned}$$

then for every $i \in \{0, \ldots, k\}$ the regression coefficients satisfy the relations:

$$\beta_i = \beta'_i + \beta'_{k+1} \beta_i^{k+1} + \ldots + \beta'_m \beta_i^m \ . \tag{51}$$

4.2 An Interpretation of the Linear Regression Model

With the discussion in this section, we want to come closer to the *semantics* (or to what [40, 41] has called *explication* of a theory) of the linear regression model, i.e., what could be *meant* by the equations of a linear regression model, and, specifically, what could be meant by *confounding* and *confounders* in the context of linear regression. Albeit there is an immediate intuition about these concepts, it turns out to be a difficult endeavour to grasp a more precise understanding of their intended meanings.

Figure 2 illustrates what we call an *error-free linear-regressive mediator scenario*. The diagram in Fig. 2 has to be understood as what we call a *multiplicative edges diagram*, i.e., values are multiplied by labels of edges as navigating via edges. Multiplicate edges diagrams are instances of bilateral impact measure diagrams, see Def. 3, where we want to have the scenario in Fig. 2 to be interpreted as described in Table 6. Considering Table 6 and Fig. 2, we can describe the scenario consistently by the following equations:

$$\begin{aligned} X' &= cX && (52)\\ Y &= aX + bX' && (53)\\ Y &= (a + bc)X && (54) \end{aligned}$$

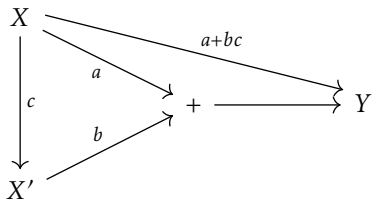


Figure 2: A two-step error-free linear-regressive mediator scenario: multiplicative edges diagram (compare to Table 6).

Note, that (54) can be gained by substituting the RHS of (52) for X' in (53):

$$Y \overset{(53)}{=} aX + bX' \overset{(52)}{=} aX + b(cX) = (a + bc)X \tag{55}$$

Let us assume that we screen off X' from the scenario in Fig. 2, i.e., that we turn X' into a *latent variable*. The concepts of *screening off* and *latent variable* only make sense relatively to a specific external observer. In particular, for such an observer, the factors a, b, and c would be unknown. Still, (54) would remain valid. For the external reviewer, Y still would be determined perfectly, i.e., error-free, by X via some factor d, i.e., for the external observer:

$$Y = dX \tag{56}$$

The external observer, not being aware of the latent variable X', would be able to determine this factor d from his observations. But he cannot determine the values of a, b and c. Once the observer learns about X' together with the values a, b and c, he is able to refine his understanding of the impact of X to Y in terms of (54), i.e., henceforth, he can use that $d = a + bc$. It can be said, that the root cause of impact in Fig. 2 all stems from X, first, directly via $X \xrightarrow{a} +$, then, second, indirectly via $X \xrightarrow{c} X' \xrightarrow{b} +$

All of this becomes different, when there would be no causal impact of X and X' in the scenario of Fig. 2 (and also not vice versa, i.e., no causal impact of X' onto X). In empirical data, X and X' would than show as (nearly) stochastically independent (assuming that empirical data set is large enough, and that there exist no further latent variables). This scenario is illustrated in Fig. 3 by dropping the edge $X \xrightarrow{c} X'$ from Fig. 2. Here, X' is not a mediator anymore. Instead, X and X' immediately, i.e., in one step together impact Y. Now, Y cannot be anymore perfectly determined solely in terms of X as in Fig. 2 and (54). Now, when screening off X', the stochastic distribution of X', is perceived as varying (stochastically distributed) error ε, which disturbs the impact of X on Y as illustrated *informally* in Fig. 3 (i). Henceforth, we will use a *dotted arrow* for the purpose of illustrating such an error-prone impact, see Fig. 3 (ii).

Next, we switch our attention from the mediator in Fig. 2 the effect of *confounding*, see Fig. 4. Actually, Fig. 4 depicts the same scenario as Fig. 2, in the sense that the process described in Table 6 an (52), (53), and (54) apply to it. The difference is that we are now also interested in the role of X as being a *confounder* to X' in addition to understanding X' as being a mediator to X. The edge $X \xrightarrow{c} X'$ still represents a *causal impact*, however, when we want to describe the

Table 6: A two-step error-free linear-regressive mediator scenario: assumptions and process of impact (compare to Fig. 2).

- X and X' together *causally impact* Y.
- We can think of the impact of X and X' on Y as a two-step process.
  - Before the impact, X has the value v_1. X' and Y have values v_2 resp. y, which do not influence the impact process.
  - First, X impacts X' so that its value changes to $v_2' = cv_1$, illustrated by the edge $X \xrightarrow{c} X'$ in Fig. 2.
  - Second, Y is impacted so that its value changes to $av_1 + bv_2'$ (illustrated in Fig. 2 by the of edges $X \xrightarrow{a} + \xleftarrow{b} X'$ and $+ \longrightarrow Y$).
- Note that during the whole process, the value v_1 of X remains unchanged. Therefore, after the second step, we have that the value of Y equals $(a + bc)v_1$.
- Consistently, the edge $X \xrightarrow{a+bc} Y$ in Fig. 2 represents the impact of X onto Y when screening off X'. X' is called *mediator* in the given scenario.
- The impact process is *perfectly* error-free in all of its parts. We assume that it is neither disturbed by the environment nor suffers measurement mistakes.

impact on Y merely in terms of X', the reciprocal of c matters as depicted by the edge $X' \dashrightarrow X$ in Fig. 3. We have used a *dashed* arrow for $X' \dashrightarrow X$ to indicate that it does not a represent a causal impact but merely an observable behaviour. First, we can rewrite (52) as follows:

$$X = \frac{1}{c}X' \tag{57}$$

Next, we can describe the impact on Y merely in terms of X' as follows:

$$Y \overset{(53)}{=} aX + bX' \overset{(57)}{=} a\frac{1}{c}X' + bX' = \left(a\frac{1}{c} + b\right)X' \tag{58}$$

To represent (58) accordingly, we have added a respective edge $X' \longrightarrow Y$ to Fig. 3. There is an important difference in *mediator effects* and *confounding effects*. After screening of the mediator X', we keep still control over the impact process, i.e., we can trigger the impact process with own values, as we have still control over the variable X. When we screen off the *confounder* X', we are forced into the role of a pure observer. The whole process is triggered by the now latent variable X and we can only record changes of Y as they come by.

Such considerations show, that our informal explanations are limited. To be more precise, we needed to explain more about the experimental setup, its conditions and ways to manipulate it. Still, we think that our discussion can help to understand better how a linear regression model could be interpreted, and also, to show the limitations of its interpretation.

Now, we turn to the usual *linear regression model*, in which all effects are assumed to be error-prone as depicted in Fig. 5. Here, all factors $\beta_1, \beta_1', \beta_1''$ and

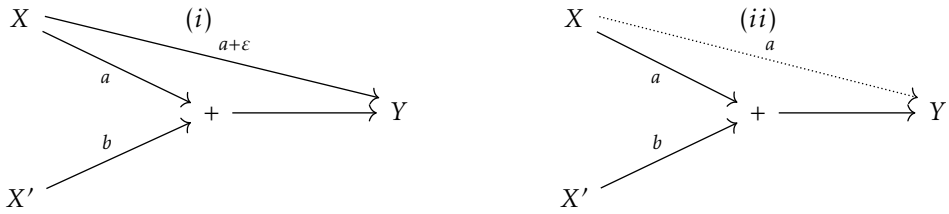


Figure 3: A one-step error-free linear-regressive mediator scenario leading to an error-prone impact after screening off the mediator.

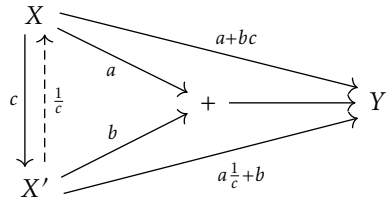


Figure 4: A two-step error-free linear-regressive mediator and confounder scenario.

$\beta_2, \beta_2', \beta_2^{**}$ are respective estimators. Let us compare Fig. 5 with Fig. 4, which show important correspondences under some obvious *renamings*. When we consider the renaming of a as β_1', b as β_2', and c as β_1'', we have an exact correspondence between the following edges[25]:

$$X \xrightarrow{a+bc} Y \tag{59}$$

$$X \xrightarrow{\beta_1'+\beta_2'\beta_1''} Y \tag{60}$$

See, how (60) is an instance of (47). Similarly, when we again match a with β_1', b with β_2', but, now, $1/c$ with β_2^{**}, we again have an exact correspondence between the following edges:

$$X' \xrightarrow{a\frac{1}{c}+b} Y \tag{61}$$

$$X' \xrightarrow{\beta_2'+\beta_1'\beta_2^{**}} Y \tag{62}$$

See, how (62) is again an instance of (47).

Despite the similarities, there are also important differences between the scenarios in Fig. 5 with Fig. 4. We want to highlight the following. With (59) and (60),

[25]Furthermore, we always assume X is renamed to X_1 and X' is renamed to X_2. Before, we have used X, X' instead of X_1, X_2 only to enhance readability and, furthermore, to highlight that, with X, X' we are not working with a usual regression model but specific error-free equations.

we have mapped c to $\beta_1^{''}$, and with (61) and (62), we have mapped $1/c$ to β_2^{**}, however, a simultaneous mapping of c to $\beta_1^{''}$ and $1/c$ to β_2^{**} is, in general not possible; as in the regression model, we have, in general:

$$\beta_2^{**} \neq \frac{1}{\beta_1^{''}} \tag{63}$$

Given a linear regression problem in matrix notation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, we have its sum-of-least-squares optimization has the solution $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, see, e.g., [87, 149]. Therefore, assuming that the factors in Fig. 5 correspond to observation vectors $X = (x_1, \ldots, x_n)^T$ and $X' = (x'_1, \ldots, x'_n)^T$, we actually have (with $\mathbf{Y} = X'$ and $\mathbf{X} = X$ for $\beta_1^{''}$, as well as $\mathbf{Y} = X$ and $\mathbf{X} = X'$ for β_2^{**}):

$$\beta_1^{''} = \frac{\sum_{i=1}^{n} x_i x'_i}{\sum_{i=1}^{n} x_i x_i} \qquad \beta_2^{**} = \frac{\sum_{i=1}^{n} x'_i x_i}{\sum_{i=1}^{n} x'_i x'_i} \tag{64}$$

Clearly, according to (64) $\beta_1^{''}$ and β_2^{**} are, in general, no reciprocals, rather, we have that:

$$\beta_2^{**} = \frac{\sum_{i=1}^{n} x_i x_i}{\sum_{i=1}^{n} x'_i x'_i} \beta_1^{''} \tag{65}$$

Equation (64) means that linear regression optimization is asymmetric. It matters, in which *direction* we analyse a relationship between two factors X_1 and X_2. Similarly, let us turn our scenario in Fig. 5 into a regression problem according to the *standard regression model* by adding an observation vector $X_0 = (1_1, \ldots, 1_n)^n$ to account for the estimation of an *intercept*. Now, both $X \xrightarrow{\beta_1^{''}} X'$ and $X' \xrightarrow{\beta_2^{**}} X$ represent two *simple regression problems* in opposite directions. Again, the scenario is asymmetric. With ρ_{XY} denoting the Pearson correlation coefficient between X and Y and σ_X denoting the standard deviation of X as usual, we have, in general, the following (where the first equation (i) is due to Lemma B in Appendix 3):

$$\beta_2^{**} \overset{(i)}{=} \rho_{X'X} \frac{\sigma_X}{\sigma_{X'}} = \rho_{X'X} \frac{\sigma_X}{\sigma_{X'}} \cdot \frac{\rho_{XX'} \frac{\sigma_{X'}}{\sigma_X}}{\rho_{XX'} \frac{\sigma_{X'}}{\sigma_X}} = \rho_{XX'}^2 \frac{1}{\beta_1^{''}} \tag{66}$$

Again, in (66), the estimators stemming from two opposite directions are not simply reciprocal. Only in case of perfect correlation, i.e., whenever $\rho_{XX'}$ equals one, the estimators are reciprocals.

4.3 On the Utilization of Linear Regression Models

Adjusting for confounding by adding additional factors is in widespread use, it is then called, e.g., *"Multiple linear regression with adjustment for confounding factors"* [135], *"linear regression analyses adjusting for confounding factors"* [235],

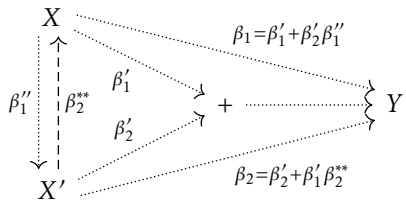


Figure 5: An error-prone linear-regressive undetermined scenario (multiplicative edges diagram). Note that $\beta_1'' = \rho_{X_1X_2}/\beta_2^{**}$ *and* $\beta_2^{**} = \rho_{X_1X_2}/\beta_1''$.

"regression analysis [..] *after adjusting for confounders"* [153], *"Multiple linear regression analyses* [..] *after adjustments for confounding factors"* [132], or *"multiple linear regression analysis* [..] *after adjusting for potential confounders"* [158].

Adjusting for confounding is considered a way of *controlling confounding*, i.e., the adjusted coefficient is considered a *better* value than the crude coefficient in the sense that the adjusted coefficient would better reflect the *genuine* [65, 62] or *isolated* impact of the factor.

There are two ways to utilize the findings of a regression model:

- *Utilization in Predictions* The found model is used in outcome prediction in future cases. This seems to be the utilization of the regression model *per se* – understanding "regression" as a synonym for *prediction*.
- *Utilization for Assessments of Interventions* The overall aim to intervene, i.e., to affect the influencing factor, in order to have a desired impact on the outcome. For example, a doctor would recommend to decrease body weight in order to increase life expectancy. Now, the linear regression model would tell, in how much a recommended treatment is relevant.

In any case, it is usually assumed that the data analysts have an understanding of the causalities in the study. When it comes to confounder adjustments, it is usually assumed that the factors, with which the original model is adjusted, are actually confounders and not mediators. For example [130] states: *"Specifically, confounders are variables that are associated with both exposure and outcome but not affected by either the exposure or outcome* [180]."

4.3.1 Utilization in Predictions

In case of predictions, adjustment is about finding a better model. In future predictions, the adjusted coefficients alone would not be useful, even harmful, i.e., value is added only together with the additionally added confounding coefficient. When you use the only the adjusted values of the extended model for predictions, this would mean that you predict each future case as it would be at level zero for all confounder, which renders such predictions obviously as inadequate. If you have no access to confounder data in future cases, you are urgently advised to stay with the crude coefficients. If you have access to confounder data in future cases, you are advised to use the full extended model, i.e., adjusted coefficient plus confounder coefficients together, to achieve more accurate predictions.

Therefore, if a study aims at predictions, a statement such as *"We found that, after adjustment for confondunding factors* $X, Y, Z, \ldots$*, the exposures* $A, B, \ldots$*, im-*

pact the outcome C with $\beta_A = v_A, \beta_B = v_B, \ldots$" rather seems to make little sense, as the interesting information is about all coefficients, adjusted coefficients plus confounding coefficients.

Similarly, it could be argued that is makes little sense to restrict adjustment to confounders in this case. In a multiple regression model with a known causal direction between any two of the influencing factors, we have that one them is a confounder from the perspective of the other, and the other is a mediator from the perspective of the previous, see Fig. 4. So, a multiple regression model is made of confounders and mediators anyhow, and the model is utilized regardless of whether the one or the other is a confounder or a mediator. So, why restricting the refinement of a given model to the addition of confounders only?

4.3.2 Utilization for Assessment of Interventions

In scenarios that aim at intervention, i.e., at strengthening or weakening a primary influencing factor in service of a wanted effect onto the outcome, a report such as *"We found that, after adjustment for confounding factors $X, Y, Z, \ldots$, the exposures $A, B, \ldots$, impact the outcome C with $\beta_A = v_A, \beta_B = v_B, \ldots$"* (compare to Section 4.3.1) makes sense. Enforcing a change of a primary influencing factor would not change the level of the confounder. The idea would be that the impact increases according to the adjusted coefficient for each fixed level. Therefore, it makes sense to consider the adjusted coefficient at the more realistic impact than the crude impact. Whether the system actually behaves as such after an intervention from the outside is merely an assumption, more sophisticated observations would be needed to provide evidence for such than provided by the plain linear regression data. This means, as the data, on which such assessment is based, stems from a study in which the particular intervention has not been enacted, such utilization is speculative. Ideally, the effects needed to be evaluated on the basis of new, typically longitudinal, data from after invention.

In this case, it also makes sense to exclude mediators from the adjustment, as mediators impact the exposure on behalf of the primary influencing factor, i.e., the primary influencing factor remains the original cause of the impact and, therefore, adjusting for a mediator would not be appropriate.

4.4 On Cutoff Rules in Confounder Adjustment and Improvements

Many studies use linear regression with adjusting for confounders. However, often, the resulting confounding effects themselves are not discussed further or only little, in particular, there might be no discussion of the significance of the confounding effects [235, 153, 158, 135, 132, 136]. In such studies, it might simply be taken for granted that adding further factors to the ones which the study is primarily interested in, would increase the quality of the findings by controlling the confounding effects. The additional factors might be selected in regards of common sense or domain knowledge. When the study is not conducted with a specifically collected primary data, but with an already existing data set such as NHANES[26] (National Health and Nutrion Examination Survey) [235], it would be naturally to extend the model with additional factors that are available in the data anyhow.

In the literature, it is stated that studies often use a *change-in-estimate* cutoff

[26] https://www.cdc.gov/nchs/nhanes/

rule to distinguish significant from non-significant confounders, see, e.g., [130, 180, 37, 138]. For example, [130] states: *"A cutoff of 10% is commonly cited in the literature* [37]." Various studies [37, 143, 138, 145, 91, 146] have investigated various ways to optimally select confounders to be included into the model. In regard of the cutoff rule, [130] furthermore states: *"However, the appropriateness of this cutoff has never been evaluated."*

Our approach is not to rely on rule of thumbs such as the cutoff rule at all. Instead of merely measuring a confounding effect and base decisions on it, we suggest to investigate and explain the confounding effect. Let us assume that a confounding effect (change-in-estimate) is measured as the (absolute) percentage change $|(\beta - \beta')/\beta|$ of a crude coefficient β and its adjusted coefficient β'. Now, please consider the example in Fig. 6. Figure 6 illustrates three different data scenarios (i–iii) with a primary influencing factor A and a confounder B as multiplicative edges diagrams, compare to Figs. 2 through 5. Here, we assume that the regression problem is given as a standard regression problem including the estimation of an intercept, however, we have omitted the intercept as a factor from the diagram to keep the illustration simple. This explains the occurrence of the Pearson correlation coefficient ρ_{AB} in Fig. 6, compare to (66). All of the three scenarios show exactly the same confounding effect, as the crude coefficients as well as the adjusted coefficients are always the same, i.e., $A \xrightarrow{1.1} C$ and $A \xrightarrow{1} +$ respectively. Despite the same confounding effect, the scenarios are very different. In scenario (i), the individual impact of A and B in the full (adjusted) regression model are perfectly balanced, i.e, their coefficients are the same, they are both *one*; whereas, the causal impact of B onto A is relatively high ($\rho^2_{AB} \times 10$) – resulting into a relatively small latent coefficient of 0.1. In contrast, in scenario (ii), both the causal impact (ρ^2_{AB}) and the latent coefficient (1.0) are relatively moderate; whereas the adjusted coefficient (1.0) and differs a lot from the confounder coefficient (0.1). So, it can be said that, in scenario (i), the confounding effect can be explained rather by the large causal impact of the confounder, whereas, in scenario (ii), it can be rather explained by the smallness of the confounder coefficient. From the perspective of (ii), scenario (iii) is a merely more extreme form of scenario (i). The latent coefficient is further decreased, whereas, this time, the confounder coefficient is relatively large, i.e., twice as much as the adjusted coefficient.

Now, given a 10% cutoff rule, the confounder B would be rejected for all three scenarios in Fig. 6. However, given the explained differences between the scenarios, an analyst might want to decide differently in each scenario. A discussion of how is beyond the scope of this study, as concrete decisions needed to be based on the domain knowledge surrounding the respective study, the concrete study design (experimental setup, clinical study*clinical study* setup etc.) and, last but not least, the intended purpose of the analysis, i.e., utilization in predictions vs. utilization in assessments – compare to Sects 4.3.1 and 4.3.2.

Next, let us assume that the confounding effect is measured particularly simple, as (arithmetic) difference $\Delta(\beta) = \beta - \beta'$ of a crude coeffienct β and its adjusted coefficient β'. Now, our findings (47) and (51) show as

$$\Delta(\beta_i) = \beta'_m \beta''_i \tag{67}$$

$$\Delta(\beta_i) = \beta'_{k+1} \beta_i^{k+1} + \ldots + \beta'_m \beta_i^m \tag{68}$$

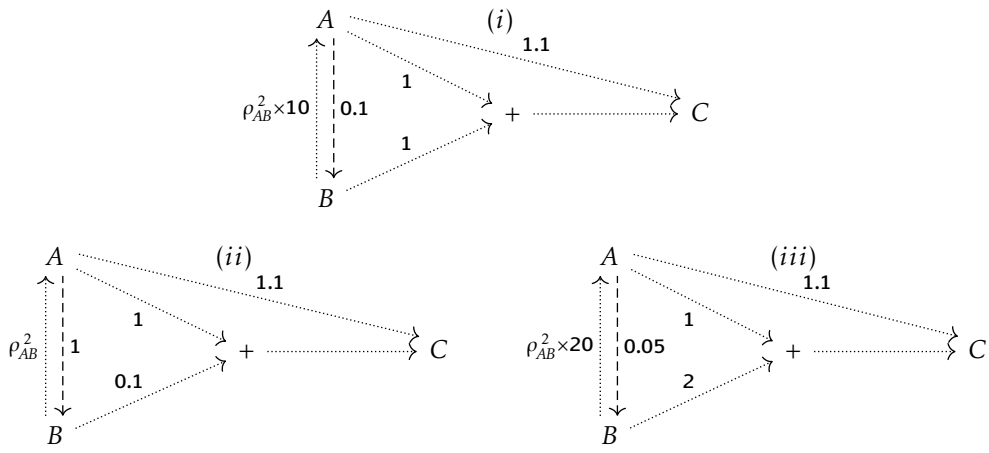


Figure 6: Three different data scenarios, each showing the same confounding effect (change-in-estimate) of less than 10% (≈9%, absolutely) for the primary influencing factor A after adjusting for the confounder B (in each of the data scenarios (i–iii), the crude coefficient is 1.1., whereas the adjusted coefficient is 1).

Given, (67) and (68), it could be said that a confounding effect $\Delta(\beta_X)$ equals the *direct* impact of the confounders *speed up* by their *indirect* latent influences (where a latent influence of a confounder stems from reversal of its impact onto the primary influencing factor X). Now, let us assume that we invert the direction of causality for (67) and (68). This is simply a question of perception and terminology, i.e, we consider the factors that are added to the regression model to form an extended model as *mediators* instead of as *confounders*. Now, it is fair to talk about $\Delta(\beta)$ as a *mediator effect*. Now, it could be said that a mediator effect $\Delta(\beta_X)$ equals the *direct* impact of the mediators *speed up* by their *indirect* latent impacts.

5 Measuring the Impact of a Categorical Factor in its Entirety

5.1 On the Impact of a Categorical Factor in its Entirety

Given a categorical factor, i.e., a random variable $X : \Omega \longrightarrow \{v_1, ..., v_n\}$, it is not obvious, how to assess the impact of X as a whole, i.e., *in its entirety*. To be used in data science methods such as linear regression, logistic regression, or association rule mining, categorical factors have to be prepared or transformed first. One option is to turn each category, i.e., each instance of a categorical random variable, into a single binary factor, before processing the whole group of resulting variables jointly within some data science method. This splitting of categorical factors is exactly what is done by leading ARM tools such as RapidMiner[27]. Similarly, as just one important example, this technique was also used by the gender pay gap analysis of the EU commission [72], as part of utilizing Oaxaca-Blinder decomposition – see Section 3.3. Another option is to impose an artificial order on the categories and assign numbers to them, so that the categorical factor is coerced into a numerical factor, and this is exactly what we do later in Chapters 7 and 8, when we deal with mixed-scaled data.

To explain the issue in more depth, let us a have a more concrete example, i.e., our introductory example from Table 1 on page 19. In the terminology of association rule mining the impact of each single profession onto the salary could be assessed as a lift, i.e.,

$$\begin{aligned}
&\mathrm{E}(\mathit{Salary} \mid \mathit{Profession{=}Chef})/\mathrm{E}(\mathit{Salary}),\\
&\mathrm{E}(\mathit{Salary} \mid \mathit{Profession{=}Builder})/\mathrm{E}(\mathit{Salary}),\\
&\mathrm{E}(\mathit{Salary} \mid \mathit{Profession{=}IT})/\mathrm{E}(\mathit{Salary}), \mathit{and}\\
&\mathrm{E}(\mathit{Salary} \mid \mathit{Profession{=}Lawyer})/\mathrm{E}(\mathit{Salary}).
\end{aligned}$$

And similarly, the impact of each single city onto the salary could be assessed as:

$$\begin{aligned}
&\mathrm{E}(\mathit{Salary} \mid \mathit{City{=}\ Boston})/\mathrm{E}(\mathit{Salary}),\\
&\mathrm{E}(\mathit{Salary} \mid \mathit{City{=}NY})/\mathrm{E}(\mathit{Salary}),\\
&\mathrm{E}(\mathit{Salary} \mid \mathit{City{=}LA})/\mathrm{E}(\mathit{Salary}), \mathit{and}\\
&\mathrm{E}(\mathit{Salary} \mid \mathit{City{=}Seattle})/\mathrm{E}(\mathit{Salary}).
\end{aligned}$$

Now we can compare the impact of each two instances of the factors among each other, both, homogeneously, i.e., between instances of the same factor such as

$$\begin{aligned}
\mathit{Profession{=}Chef} \quad &\mathit{vs.} \quad \mathit{Profession{=}Builder},\\
\mathit{Profession{=}Lawyer} \quad &\mathit{vs.} \quad \mathit{Profession{=}IT},\\
\mathit{City{=}NY} \quad &\mathit{vs.} \quad \mathit{City{=}\ Boston}, \mathit{and}\\
\mathit{City{=}Seattle} \quad &\mathit{vs.} \quad \mathit{City{=}LA},
\end{aligned}$$

or, heterogeneously, i.e., between instances of the various factors such as

[27] https://altair.com/altair-rapidminer

$$\textit{Profession=Chef} \quad \textit{vs.} \quad \textit{City= Boston,} \tag{69}$$
$$\textit{Profession=Builder} \quad \textit{vs.} \quad \textit{City=LA,} \tag{70}$$
$$\textit{Profession=IT} \quad \textit{vs.} \quad \textit{City=NY, and} \tag{71}$$
$$\textit{Profession=Lawyer}) \quad \textit{vs.} \quad \textit{City=Seattle.} \tag{72}$$

Whether comparisons such as (69) through (72) might be useful resp. meaningful in some context is not important here, however, the example shows what cannot be done without further endeavors. It is not clear how to compare the impact of the factor *Profession* onto *Salary as a whole* with the impact of the factor *City* onto *Salary as a whole*, i.e., *in their entireties*. If the factors *Profession* and *City* were binary variables in our model (5) through (9) on page 18, we could compare their impacts on the basis of the conditional expectations

$$\mathrm{E}(\textit{Salary} \mid \textit{Profession})/\mathrm{E}(\textit{Salary}), \textit{and} \tag{73}$$
$$\mathrm{E}(\textit{Salary} \mid \textit{City})/\mathrm{E}(\textit{Salary}), \tag{74}$$

however, in our model, the factors *Profession* and *City* are non-binary, categorical variables, and the expectations (73) and (74) simply do not exist in our model.

In linear regression, see Chapter 4, such questions do not appear, because, here, the influencing factors are numerical factors. In linear regression, the effect of an influencing measure is indeed always assessed as a whole, i.e., in its entirety. However, whenever the influencing factors are categorical variables, the question on how to deal with them as a whole appears. One approach has been described above, i.e., transforming them into numerical variables. In Section 5.2, we develop a novel measure to assess the impact of a categorical variable in its entirety.

5.2 A Novel Measure: Coupled Impact Assessment (C-IA)

We introduce a novel measure that quantifies the "impact" of a categorical influencing factor X on a real-valued target variable y by aggregating the relative differences between the conditional expectations of y given $X = v_i$ and y given $X = v_j$ across all possible pairs values v_i and v_j of X, each weighted with the probabilities of X being v_i and X being v_j.

First, we provide the formal definition of the measure in Def. 15, then, we provide an explanation on how we came up with the measure in Section 5.3.

For our impact definition in Def. 15, we first need to define the notions of *sheer value* and *sheer lift*. The sheer value of a real number turns any fraction, that has an absolute value smaller one, into its reciprocal, and behaves like the identity on all other real numbers in $\mathbb{R}\backslash]-1, 1[$, see Def. 13. Sheer values are to quotients, what *absolute values* are to (subtractive) differences. For the sake of convenience, we also include the definition of absolute value, see Def. 12.

Definition 12 (Absolute Value)**.** Given a real number $x \in \mathbb{R}$, we define the *absolute value* x, denoted as $|x|$, as usual, as follows:

$$|x| \quad = \quad \begin{cases} x & , x \geqslant 0 \\ -x & , \textit{else} \end{cases} \tag{75}$$

Definition 13 (Sheer Value)**.** Given a real number $x \in \mathbb{R}$, we define the *sheer value of* x, denoted as $\|x\|$, as follows:

$$\|x\| \quad = \quad \begin{cases} x & , |x| \geqslant 1 \\ \dfrac{1}{x} & , \textit{else} \end{cases} \tag{76}$$

The idea of sheer values is to utilize them in the definition of what we call *sheer lifts*, i.e., yielding quotient-based measures for the distinctness of two effects $m_1 \in \mathbb{R}_0^+$ and $m_2 \in \mathbb{R}_0^+$ that are independent of the order of m_1 and m_2, i.e.:

Definition 14 (Sheer Lift)**.** Given real numbers $m_1 \in \mathbb{R}_0^+$ and $m_2 \in \mathbb{R}_0^+$, we define their *sheer lift* as the sheer value (Def. 13) of their quotient, i.e.:

$$\left\| \frac{m_1}{m_2} \right\| = \left\| \frac{m_2}{m_1} \right\| \geq 1 \tag{77}$$

Definition 15 (Coupled Impact Assessment (C-IA))**.** Given a target variable $y : \Omega \longrightarrow \mathbb{R}_0^+$ and an influencing factor $X : \Omega \longrightarrow \{v_1, \ldots, v_n\}$, we define the *impact of* X *on* y, denoted as $\iota(X \Rightarrow y)$ as follows:

$$\iota(X \Rightarrow y) \quad = \quad \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n} \mathsf{P}(X = v_i)\mathsf{P}(X = v_j) \left\| \frac{\mathsf{E}(y \mid X = v_j)}{\mathsf{E}(y \mid X = v_i)} \right\| \tag{78}$$

The sheer lift and coupled impact assessment are defined only for positive real numbers, as it is not obvious, how to compare and aggregate sheer values of different sign. For the course of this thesis, this poses no practical problem, as all of our approx. 700 datasets have only positive numerical target values. However, in future work, it might be interesting to develop a measure that works for real numbers in general.

5.3 Stepwise Development of the C-IA Measure

A first attempt to provide a measure for the impact of a categorical factor X onto a numerical factor y, would be to simply take the average of the impacts of the individual instances of X onto y. Here, the individual impact of a single instance of X onto y can be, for example, oriented towards the *lift* in association rule mining, just in its generalized form adopted to conditional expectations[28], as follows:

$$\frac{1}{n} \sum_{1 \leq i \leq n} \frac{\mathsf{E}(y \mid X = v_i)}{\mathsf{E}(y)} \tag{79}$$

The problem with the attempt in (79) is that it entirely neglects the probabilities $\mathsf{P}(X = v_1), \ldots, \mathsf{P}(X = v_n)$ of the various instances of X. This leads to counter-intuitive effects, as can be constructed by some example as follows. Table 7 shows two factors of a dataset X and X', each having the same number of three instances. Now, X and X' have the same conditional expectations for their first, second and third instance, but significantly different values for the probabilities of their corresponding instances[29]. Now, in therms of the measure in (79), X

[28] see the discussion on (16) on page 19.

[29] Note, that the categories of X are, in general, different from those of X'. The number of categories of X and X' is the same just for the sake of this example. The correspondence between the variables is according to their indices.

Table 7: Scenario with two factors X and X' showing the same conditional expectations but significantly different values for the probabilities of their instances, each with its impact value (rounded) according to (79), its impact value (rounded) according to (80), its impact value (rounded) according to (81) and its C-IA impact $\imath(X \Rightarrow y)$ (rounded) according to Def. 15, (78).

X	$\mathsf{P}(X{=}v_1)$	$\mathsf{P}(X{=}v_2)$	$\mathsf{P}(X{=}v_3)$	$\mathsf{E}_{X=v_1}(y)$	$\mathsf{E}_{X=v_2}(y)$	$\mathsf{E}_{X=v_3}(y)$	(79)	(80)	(81)	$\imath(X \Rightarrow y)$
	1/3	1/3	1/3	1.000	2.000	3.000	1.00	1.5	1.5	1.78

X′	$\mathsf{P}(X'{=}v'_1)$	$\mathsf{P}(X'{=}v'_2)$	$\mathsf{P}(X'{=}v'_3)$	$\mathsf{E}_{X'=v'_1}(y)$	$\mathsf{E}_{X'=v'_2}(y)$	$\mathsf{E}_{X'=v'_3}(y)$	(79)	(80)	(81)	$\imath(X' \Rightarrow y)$
	0.99	0.005	0.005	1.000	2.000	3.000	1.97	1.98	1.03	1.03

has an impact value of 1 and X' has an impact value of ≈ 1.97. The impact of X' is, according to (79), higher than the impact of X. This can be considered as counter-intuitive. We argue that the factor X, considered in its entirety, shows significantly more *activity* than factor X'. Actually, the factor X' is almost *inactive*, as most of its individuals are concentrated in just one instance of its categories, i.e., instance $X' = v_1$. The concepts of *active* versus *inactive* heavily rely on intuition, and we make no attempt to formalize it here. We try to provide some more intuition about it later, when we introduce a 2-step game on the dataset. First, we want to analyse (79) further.

We identify two potential reasons for the fact that the impact of X' is higher than the impact of X according to (79), i.e., (i) the fact that the several corresponding instances of X and X' have different probabilities and (ii) the fact that (79) is not defined in terms of sheer lift, but straightforward as generalization of lifts from ARM. Considering the factor X, the lift of its first instance category v_1 is smaller than one, therefore, outbalancing the effect of the lift of the third instance category. We can try to fix this by adopting the impact definition of (79) to sheer lifts (Def. 14), resulting into:

$$\frac{1}{n} \sum_{1 \leq i \leq n} \left\| \frac{\mathsf{E}(y \mid X = v_i)}{\mathsf{E}(y)} \right\| \tag{80}$$

Applying the new measure (80) to the data in Table 7 yields an impact value of 1.5 for X, and an impact value of ≈ 1.98 for the factor X', see Table 7. This means that the impact of X' is still significantly higher than that of X, i.e., the situation is basically the same as with the measure from (80). The situation only changes when we incorporate the probabilities of the category instances in the measure as with our measure C-IA, see Table 7. Our measure yields 1.78 for X and 1.03 for X'. The value of X' is very close to 1, which also fits our intuition that the factor X' is almost *inactive*.

A next straightforward try would be to incorporate probabilities of category instances as weights into (80) yielding weighted averages instead of direct averages, resulting into:

$$\sum_{1 \leq i \leq n} \mathsf{P}(X = v_i) \left\| \frac{\mathsf{E}(y \mid X = v_i)}{\mathsf{E}(y)} \right\| \tag{81}$$

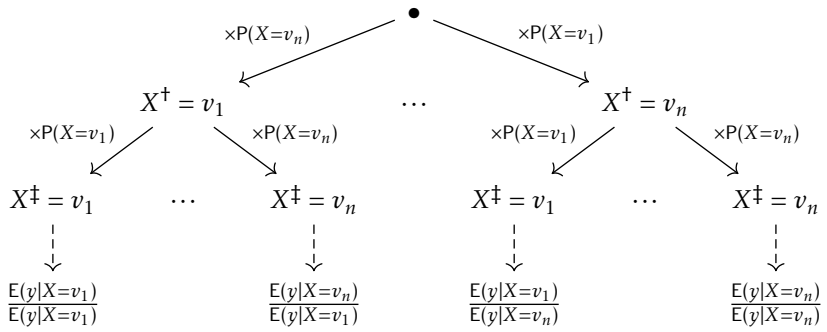


Figure 7: Impact of a categorical factor according to Def. 15: A two-step game is played on a dataset D for its factor X via two random variables $X^\dagger$ and $X^\ddagger$ (of an appropriately constructed data set $D^\star$) that have both the same distribution as X w.r.t D.

It turns out that (81) resolves the counter-intuition of our example in Table 7, i.e., the factor X receives an impact value of 1.5, whereas X' receives a smaller impact value of ≈ 1.03. The measure in (81) is worth to be investigated in further work, however, for the purpose of this thesis, we step further and provide the more fine-grained measure C-IA. Considering our example of Table 1 on page 19, we could choose that the target factor y stands for salary, X stands for professions and X' stands for cities. Now, the following could be said:

- The measure (81) stands for the impact of selecting a profession or a city for living, whereas
- the measure C-IA stands for the impact of changing the profession or moving from one city to another.

Disclaimer: As a disclaimer we need to say that such explanation is only for illustrative purposes, to give an appeal to intuition. We cannot tell from the data at all, whether the causalities in the real world behave as such. Our explanation is only for the purpose of creating intuition about the measure.

We can think of our measure C-IA as a two-step game as illustrated in Fig. 7. The dataset is taken as basis for conducting two experiments as a thought experiment: each experiment tossing from the individuals represented by the dataset, i.e., according to the probabilities of dataset. Technically, you can think of this as constructing a larger data set by duplicating rows appropriately, i.e., preserving probabilities, and having two i.i.d. random variables $X^\dagger$ and $X^\ddagger$ that have both the same distribution as X w.r.t. the original data set.

6 Utilizing Coupled Impact Assessment for Confounding Adjustment

In order to utilize the C-IA measure from Def. 15 for confounding adjustment, we need to clarify how to apply it to more than one influencing factor simultaneously.

In Def. 15, we have defined the measure C-IA for measuring the impact of a categorical factor X, in its entirety, onto a numerical factor y. Next, we would like to have a similar measure for the simultaneous impact of n categorical factors $X_1, \ldots X_n$ onto a numerical factor. It turns out that Def. 15 is completely sufficient for this purpose as we can apply it to an appropriate concept of tuple factors. Each set of factors can be treated in a natural way as a single tuple factor as follows.

Definition 16 (Tuple Factors)**.** *Given factors $X_1 : \Omega \longrightarrow V_1, \ldots X_1 : \Omega \longrightarrow V_n$, we define their tuple factor $\langle X_1, \ldots, X_n \rangle : \Omega \longrightarrow V_1 \times \cdots \times V_n$ for all instances $\langle v_1, \ldots, v_2 \rangle \in V_1 \times \cdots \times V_n$ as follows:*

$$\mathsf{P}(\langle X_1, \ldots, X_n \rangle = \langle v_1, \ldots, v_2 \rangle) = \mathsf{P}(X_1 = v_1, \ldots, X_n = v_n) \tag{82}$$

Now given a target factor y, a primary impacting categorical factor X and a series of further, potentially confounding, categorical factors, $C_1, \ldots, C_2$, we can use the C-IA measure to assess the impact X onto y alone, and, also to assess the simultaneous impact of $X, C_1, \ldots C_n$ onto y, i.e.:

$$\iota(X \Rightarrow y) \tag{83}$$

$$\iota(\langle X, C_1, \ldots C_n \rangle \Rightarrow y) \tag{84}$$

Now, we say that (84) adjusts the confounding effect of $C_1, \ldots, C_n$, and we call (84) the adjusted value of (83). By comparing (83) and (84), we can assess the existence of potential confounding effects, and utilize this in our experiments, see Chapters 7 and 8.

7 Experimental Setup

The experimental setup of this study consists of the usual well-defined stages of data analytics [99] that help to achieve credible outcomes. After an intensive search for suitable data by screening existing repositories, we have selected a total of 694 dataset as follows:

- 657 datasets from the Harvard Dataverse[30] repository, plus,
- 37 datasets from the New Zealand government repository[31].

To eliminate possible problems with missing values, outliers, and other inconsistencies, we meticulously cleaned these data.

See Figure 8 for an overview of the experimental setup that will be described in detail in Sections 7.1 through 7.4. We proceed as follows. In Section 7.1, we provide explanations about the datasets used in our experiments. In Section 7.2, we describe the experiments that we have conducted. In Section 7.3, we describe the technical setup with Google Colab. In Section 7.4, we discuss the key features of the Harvard Dataverse as important background information.

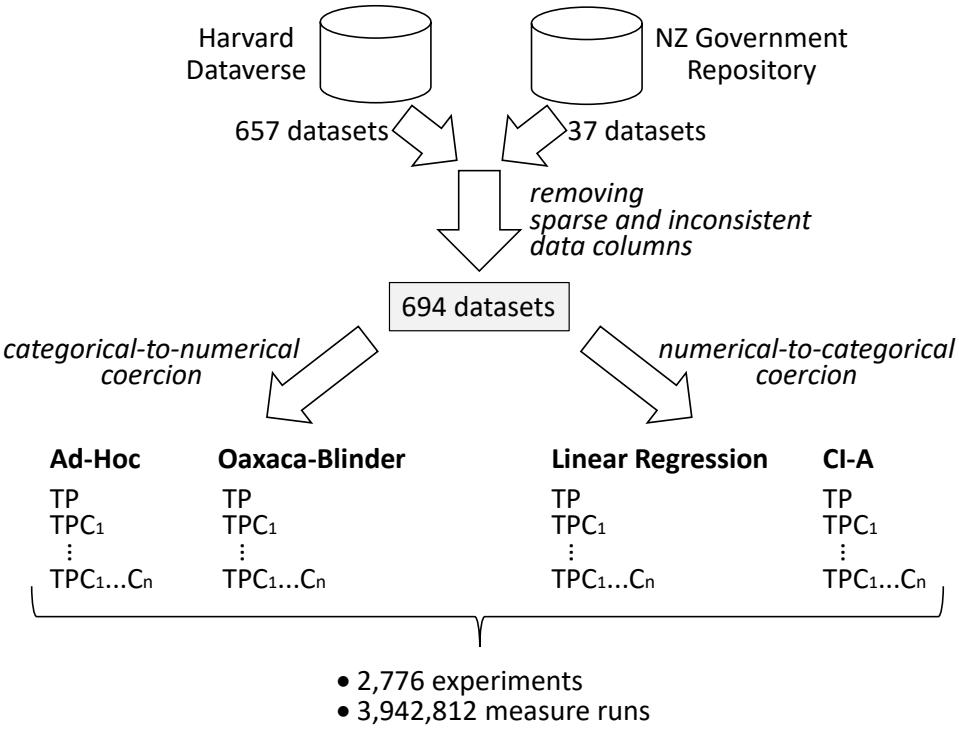


Figure 8: Experimental setup.

[30] https://dataverse.harvard.edu/
[31] https://data.govt.nz/

7.1 The Used Datasets

7.1.1 Heterogeneity

A heterogeneous dataset refers to a collection of data that combines information from diverse sources, formats, or domains [2, 8, 115, 25]. This diversity allows for a more comprehensive analysis, capturing complex relationships and interactions across different dimensions. The datasets come from various fields, such as health, economics, social sciences, environmental studies, and technology. The integration of such diverse data provides a broader perspective, enabling researchers to study multifaceted problems that span multiple domains [21].

The main advantage of heterogeneous datasets lies in their ability to reflect real-world complexity more accurately [218, 69, 86]. They allow researchers to explore interactions between variables that might otherwise be overlooked in isolated datasets. As a result, studies based on heterogeneous data are often more robust, generalizable, and capable of addressing intricate questions involving multiple factors [230].

To achieve more credible and precise results, we integrated 694 datasets spanning diverse categories. This extensive dataset selection enhances our understanding of interactions among multiple variables, enabling stronger and more general conclusions. As illustrated in Figure 9, the datasets encompass a wide range of domains, including health data, environmental factors, social and economic status, income levels, and other demographic characteristics. Such comprehensive coverage ensures that our data capture the full spectrum of variable values, contributing to the depth and specificity of our analysis.

By incorporating data from multiple fields, we address complex relationships that may not be evident in smaller or more narrowly focused datasets. This approach not only increases the robustness of our findings but also minimizes biases arising from limited data sources. The inclusion of diverse datasets supports broader generalizations, making the results applicable across different domains.

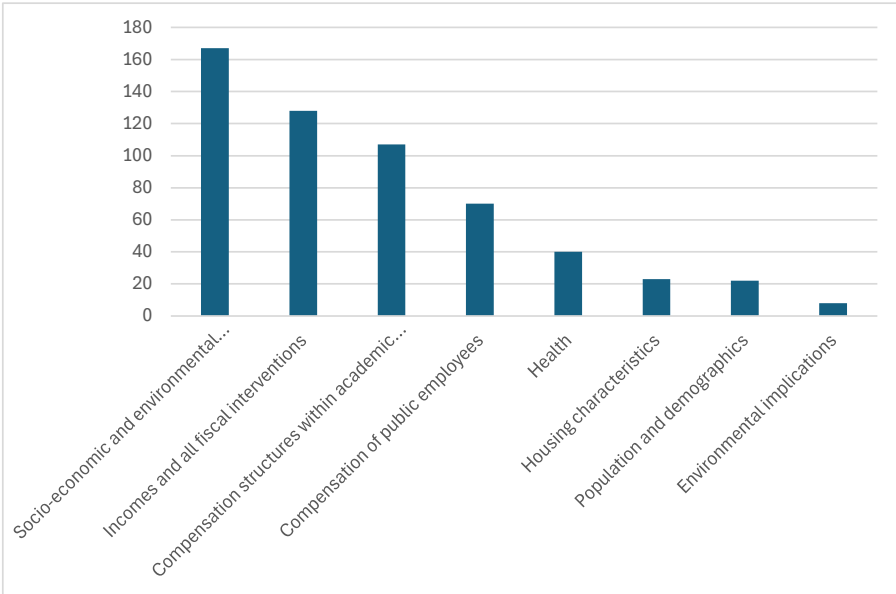


Figure 9: Heterogeneity of the 694 used datasets: Number of datasets per domain.

Table 8: Meta data about sizes of the used 694 datasets.

Min. number of rows	41
Max. number of rows	1,048,576
Avg. number of rows	76,160
Min. number of columns	3
Max. number of columns	16
Avg. number of columns	9
Number of measure runs	3,942,812

7.1.2 Meta Data

In Table 8 we have collected meta data about the sizes of the 694 used data sets. The number of rows is relevant for the statistical significance, in particular, when as we heavily drill-down as part of our experiments. The number of columns is relevant as it reflects the high number of combinations needed by the combinatorial design of our experiments, showing in the number of 3,942,812 measure runs as part of our experiments, as will be explained in Section 7.2.2.

7.2 Conducted Experiments

7.2.1 Data Cleansing

We checked each data set for sparse data and inconsistencies, see Figure 9, as follows:

- Whenever the values in a column have been too sparse, the column has been removed.
- In regard of inconsistencies, we checked previous studies of other researchers who have used a given data set. We also checked the manual given along with the dataset, whether it reports inconsistencies. Whenever previous studies or the manuals mentioned inconsistencies for some columns, we have removed these columns from the dataset.

7.2.2 Combinatorial Design of the Conducted Experiments

All datasets that we have selected for this study have at least one numerical column. For each dataset, we have chosen one numerical column as the *target factor*, and the remaining columns as the *influencing factors*. From those influencing factors, we have selected a column as the *primary influencing factor* and the remaining as the *potential confounders*. We have selected the target factor and the primary influencing factor on the basis of investigating previous studies of other researchers who used this dataset. Whenever such studies were not available, we tried to make a rational choice on the basis of the structure of the dataset, e.g., selecting naturally appearing columns such as salary.

Then, we conducted experiments on the target and the primary factor with the four measures from Section 3.2, 3.3, 4 and 5, i.e., the Ad-Hoc measure, linear regression, OB-decomposition, and our novel Coupled Impact Assessment measure

(C-IA). Each of these experiments yielded a marginal impact value (henceforth, also just marginal value for short) to be used as baseline for further experiments with potential confounders. We consider a confounding effect as detected, whenever there is at least a 10% change (percentage difference), during drill-down as compared to the marginal value, see Def. 17. These 10% of required change are also minimum threshold in our study.

Definition 17 (Study's Cutoff Rule)**.** In all our experiments (and for all used measures), a confounding effect is detected, whenever there exists at least 10% percentage difference between the marginal value (excluding the potential confounders) and the drill-down value (including the potential confounders).

A 10% difference is a commonly used minimum threshold to identify confounding effects, see Publication II and [130, 104] as well as Section 4.4.

For drill-down, we combine the potential confounders (remaining columns) with the primary influencing factor. We create all possible combinations of potential confounding variables with the primary influencing factor, e.g., let us consider a dataset consisting of 5 columns P, A, B, C, T as an example. Here, P is the primary influencing factor, T is the target factor and A, B and C are the potential confounding factors. The combination of the potential confounder with the primary influencing factor and target factor for such table will be:

$$
\begin{aligned}
&T, P, A\\
&T, P, B\\
&T, P, C\\
&T, P, A, B\\
&T, P, A, C\\
&T, P, B, C\\
&T, P, A, B, C
\end{aligned}
$$

Then, we use these combinations for further analyses (drill-down), and to compare their results with the marginal value to detect confounding effects.

In our study, we call the execution of a measure towards a single of the above drill-down combinations a *measure run*, see Def. 18. Furthermore, in our study, we say that a *single experiment* consists of running a measure towards a dataset, where running a measure towards a dataset means to execute it combinatorially often as explained above, i.e., to execute all of its measure runs, see Def. 19.

As we conducted experiments with four measures against 694 datasets with various numbers of columns, see Figure 8 and Table 8, our experimental setup resulted into:

- 2,776 experiments
- 3,942,812 measure runs

Definition 18 (Measure Run)**.** Given any dataset together with a selection of the target variable T, the primary influencing factor P and potential confounders $C_1, \ldots C_n$, *measure run* is the name for the execution of a measure towards P, T and a single combination $C_{i_1}, \ldots, C_{i_m}$ such that $m \leq n$.

Definition 19 (Study's Experiment)**.** In the context of this study, we say that an *experiment* consists of running a measure towards all possible measure runs, see Def. 18, against a dataset and specific selection of a target value, primary influencing factor and collection of potential confounders.

7.2.3 Transformation of Columns

The used datasets have both numerical and non-numerical, i.e., categorical columns. However, the measures that we tested in our experiments do not work for mixed numerical/categorical influencing factors, i.e., Oaxaca-Blinder decomposition and linear regression work only with numerical influencing factors, whereas the ad-hoc method and C-IA only work with categorical influencing factors. Therefore, conflicting columns need to be transformed into the needed scale before processing them in an experiment. We call these transformations *categorical-to-numerical coercion* and *numerical-to-categorical coercion* and they are needed as follows:

- *Oaxaca-Blinder decomposition, linear regression*: Categorical-to-numerical coercion.
- *Ad-Hoc, C-IA*: Numerical-to-categorical coercion.

We explain the details of how we have conducted the categorical-to-numerical coercion and numerical-to-categorical coercion in Sections 7.2.3.1 and 7.2.3.2.

7.2.3.1 Categorical-to-Numerical Coercion

The first step of transforming categorical values into numerical values involved examining whether the categorical variables exhibited a natural order such as levels (e.g., low, medium, high) or rankings. If such a natural order was present, we directly mapped these categories to corresponding numerical values.

However, in cases where no inherent order existed, we imposed an order based on the relationship between the categorical variable and the target variable, see Table 9 for an example with artificial data. To achieve this, we calculated the average of the target variable for each distinct category in the categorical column. This distinctive average served as the basis for ordering the categories. Once the averages were computed, we sorted the categories in an order according to their average target values. We then assigned numerical values to the categories in alignment with this sorted order, ensuring a meaningful representation of their influence on the target variable. Finally, these numerical values replaced the original categorical labels and were used in subsequent analyses.

Table 9: Example of categorical-to-numerical coercion (artificial data).

Profession	Avg. Salary	Numerical Label
Lawyer	7,000	4
Software Developer	5,500	3
Teacher	4,500	2
Chef	3,500	1

7.2.3.2 Numerical-to-Categorical Coercion

The Ad-Hoc measure and the C-IA measure are specifically designed to function only with categorical influencing factors. However, numerical variables often play a significant role in data analysis. To incorporate these variables, we convert them into categorical data by stratifying them into smaller groups, see Table 10 for an example with artificial data. This approach ensures that numerical data can be effectively analysed within the framework of categorical measures.

The stratification process begins by identifying the minimum and maximum values within each numerical influencing column. The range between these values is then calculated to determine the overall distance. To simplify the data, this distance is divided into five groups of equal length, creating distinct intervals that categorize the numerical data into discrete classes. These newly defined groups serve as categorical representations of the original numerical values, allowing us to integrate them seamlessly into our analytical model. We have chosen a number of five to create groups, as this is a common number of categories used in the widely used Likert scale [134, 39].

Table 10: Example of a numerical-to-categorical coercion (artificial data) showing the result of stratifying original data in the form of individuals with an age (in the range of 15 to 65 years) together with their salary. The stratified data shows age groups (instead of age) and avg. salaries (instead of salaries).

Age Group	Avg. Salary
15-25	2,500
25-35	3,000
35-45	3,500
45-55	4,200
55-65	4,100

By adopting this method, we retain the variability and distribution patterns present in numerical data while aligning it with the categorical requirements of the combinatorial measure. This transformation not only preserves meaningful distinctions between data points but also enables the exploration of relationships between numerical and categorical variables in a unified manner.

7.3 Google Colab

We have conducted all of our experiments in Google Colab. Google Colab[32] is a cloud-based tool that allows users to write and execute Python code along with sharing Jupyter notebooks[33]. Created by Google Research, Colab has become popular among data scientists, machine learning engineers, teachers, and researchers. Some of the advantages include a user-friendly interface, and that it is easily accessible and has powerful computational capability.

The platform creates a space where it is easy to write and run Python code without having to download any program. Colab's integration with cloud services

[32] https://colab.research.google.com/

[33] https://jupyter.org/

also enables it to provide the users with powerful computational resources, and also gives out free access to *graphical processing units* (GPUs) and *tensor processing units* (TPUs)[34], which are crucial in the training and deployment of machine learning models.

We utilized the free version of Google Colab for our analysis. The CPU configuration available to free users includes:

- a dual-core Intel Xeon processor,
- 16 GB of memory, and
- 100 GB of storage.

These resources, while modest, provided sufficient computational power for our needs.

The total runtime needed to complete the experiments for all the four measures amounted to 125 hour. Despite the limitations of the free-tier environment, such as session timeouts and restricted resource availability, we effectively managed the workflow to optimize performance.

The entire analysis of these results spanned approximately 100 days. This timeline accounts for various stages of the process, including calculating the percentage difference, identifying the confounding effects (by eliminating combinations of potential confounders whose percentage differences fall below our cutoff rule (Def. 17)), identifying various patterns of confounding, and making comparisons between the measures. By leveraging Google Colab's accessible platform, we were able to conduct the analysis without incurring additional infrastructure costs. While the free-tier resources required careful management and planning, they ultimately proved to be a practical and effective solution for our computational needs.

7.4 Harvard Dataverse

Harvard Dataverse is an advanced open source data repository solution for storing and sharing research data at its best. Developed by the Institute for Quantitative Social Science at Harvard University, it is designed to provide a secure web-based environment where researchers can identify suitable data for preservation and sharing. They can do so easily and quickly irrespective of the discipline they belong to. Originally conceived as a Harvard initiative, Harvard Dataverse is now one of the world's largest and most impactful repositories.

The platform greatly assists the paradigm of open science, providing an open data deposit which increases the amount of data openly available for analysis. For each dataset that is stored on the Harvard Dataverse, the data receive a Digital Object Identifier(DOI). Versioning and editing facilities of a dataset enable a researcher to always have the most updated and accurate dataset.

Key features of the Harvard Dataverse are:

- *Foundational Support and Features* The principal objective of using the Harvard Dataverse is to provide researchers, regardless of their fields, efficient tools and services for managing their data. It is easy for researchers to upload their datasets and the datasets can be made public accessible for other scholars to use.

[34] `https://cloud.google.com/tpu`

- *Comprehensive Data Management* An important advantage of Harvard Dataverse is its data management system. The data management system has a flexible structure that allows the user to define complex data fields, versions, and metadata. Scholars can easily edit and modify the embodied data as well as metadata; this way, the data collected remains up-to-date and credible.

- *Security and Access Control* Harvard Dataverse pays a lot of attention to authorised user access and data security. Researchers have flexibility in managing datasets, including controlling access to protect sensitive or proprietary information. This capability allows researchers to agree with set ethical standards and legal provisions during data dissemination.

- *Integration and Interoperability:* The platform is easily integrated with many different formats of research utilities and services, such as data analysis software and institutional repositories. This integration also benefits the data stored in Harvard Dataverse by expanding its accessibility and making it easily integrated into researchers' current processes.

- *Community and Collaboration:* Harvard Dataverse fosters scientific collaboration by enabling scholars to create libraries and curate datasets by theme, project, or team. For example, the leading data catalogue service *data.world*[35] ,[36] supports integration with Harvard Dataverse. Such libraries help identify related datasets and connect diverse research fields, promoting interdisciplinary cooperation and advancing the common good.

Thus, Harvard Dataverse is a key structure in the academic and research society. Thematically, it is rich in terms of covered functions, is completely committed to the principles of open science, and has a rather large impact on the practices of researchers. Thus Harvard Dataverse also substantially helps the multiple-researcher cooperation as well as the identification of the higher objectives of science and development.

[35] https://data.world/

[36] https://www.gartner.com/reviews/market/active-metadata-management/vendor/data-world/product/data-world

8 Experimental Results

8.1 Inter-Rater Reliabilities of the Investigated Methods

Table 11: Proportional agreement p, Cohen's κ, ϕ-coefficient and p-value of the χ^2 test for independence $p(\chi^2_{A \perp\!\!\!\perp B})$ for all combinations of the investigated methods, i.e., the Ad-Hoc method, Oaxaca-Blinder Decomposition (OB-Dec.), linear-regression-based (Regr.), and C-IA-based.

A	B	p_{AB}	κ_{AB}	ϕ_{AB}	$p(\chi^2_{A \perp\!\!\!\perp B})$
Ad-Hoc	OB-Dec.	0.5765	0.0027	0.0027	0.0064
Ad-Hoc	Regr.	0.6526	-0.0001	-0.0001	0.9270
Ad-Hoc	C-IA	0.6165	0.0491	0.0491	0.0000
OB-Dec.	Regr.	0.6094	0.0001	0.0001	0.9018
OB-Dec.	C-IA	0.5875	0.0424	0.0426	0.0000
Regr.	C-IA	0.6636	0.0648	0.0689	0.0000

Table 11 presents the main discovery of this study[37]. When mutually compared, the four investigated methods for detecting confounders do not show any relevant agreement or disagreement in the sense of Cohen's κ (Def. 6), i.e., they do not show any relevant agreement or disagreement beyond chance, see the discussion in Section 2.4.1 and Def. 9.

Actually, for all pairs of used methods, Cohen's κ is almost zero ($\kappa \approx 0$), see Table 11. As the quality of Cohen's κ has been discussed in the literature [53, 44], we also provide the ϕ coefficients (Def. 11) for each pair of methods in Table 11. The ϕ coefficients re-confirm the result, i.e., their are almost no correlations ($\phi \approx 0$) between any pair of methods.

The fact that the investigated methods do not show any relevant agreement or disagreement is a surprising and relevant discovery at the same time:

- *Surprisingness:* The fact is surprising, as all of the investigated methods have been specifically designed for *exactly* the same target: to detect confounders, where, actually, three of the methods (Ad-Hoc, OB-decomposition, linear-regression-based) have been in widespread use over the decades in a plethora of scientific studies for detecting confounders.

- *Relevance:* The fact is highly relevant for the working data scientist in choosing methods for the detection and adjustment of confounders. A dominating decision criterion remains good fit to the data science technique that is used in a specific study. Decision criteria for data science techniques, in general, are in the composition of heterogeneous data, i.e., the scale of the most important factors in a study, and, on the other hand, best practices of the respective study's research field.

[37] the precise calculations (30 decimal points) are available as Excel-file for download: `https://github.com/istaltech/Measures-of-Impact-and-Confounding`

Given that both the κ-values and ϕ-coefficients are all very close to zero for all combinations, this immediately leads to the assumption that the factors are all pairwise independent, or at least close to independency. A look into the corresponding contingency tables and confusion matrices, see in Table 12 and 13, immediately reconfirms us with this assumption (with values 48.50 vs. 48.45, 60.73 vs. 60.74, 52.83 vs. 51.84 etc.), at least from a practical, data-analyst viewpoint.

Given that the observed values are seemingly close to independence, we have also conducted a X^2 test for independence for each combination, see Table 11. From a statistical viewpoint, only the combinations of Ad-Hoc/Regression and OB-Decomposition/Regression would usually be argued to be close to independence, with very high p-values of 0.9270 and 0.9018 respectively[38]. All other combinations show very small p-values[39], and, therefore, must not be considered as independent from the viewpoint of statistical significance. The results of the X^2 test are due to the very large sample sizes, with close to 1 Million experiment runs for each measure, see Fig. 8.

The calculated p-values of the X^2 tests (Table 11) are a good example for ASA's warning [225], that p-values should not be interpreted barely without context[40]. Given the large sample sizes, the X^2 test for independence is simply a too fine-grained analytical device ("too sharp knife"). Instead the κ-statistic and, similarly, the ϕ-coefficient, provide appropriate (pragmatic/heuristic) measures from the viewpoint of the working data analyst. In case of our data, the κ-values and ϕ-coefficients are even almost equal for all combinations, see Table 11. It is known from the literature, that Cohen's κ and the ϕ coefficient behave similarly under certain conditions [44, 53, 89, 7], e.g., they are exactly the same in case of perfectly balanced contingency tables[41] [7]. In our case, it can be assumed that the similarities of the κ-values and ϕ-coefficients are rather due to the fact, that they are all already very close to zero.

For the sake of completeness, we have also conducted the z-test for Cohen's κ under the null hypothesis that $\kappa = 0$ ($H_0 : \kappa = 0$) for all combinations of measures, see Table 14 for the calculated p-values (calculated according to [76, 77], see Sect. 2.4.4, Def. 10 and (24)). Given the large sample sizes, the p-values of the κ z-tests and the χ^2 tests are almost perfectly the same for all combinations, see their differences in terms of 20 decimal points in Table 14. This means, in our case, that the κ z-tests are just another device for testing independence, reconfirming the analysis of the χ^2 tests through the lens of inter-rater reliability.

[38] Formally, the p-values merely mean, that at least 92.70% resp 90.18% of data sets would show observed values as least as extreme (as least as different from independent values (H_0 hypothesis)) as the current data set. Note, the American Statistical Association (ASA) has clarified that "a relatively large p-value does not imply evidence in favour of the null hypothesis" [225].

[39] Actually, the p-values of all three combinations of the C-IA measure with other measures are zero up to 30 decimal points.

[40] "Researchers should recognize that a p-value without context or other evidence provides limited information." [225]

[41] In case of 2×2-contingency tables with factors A and B, the contingency table is said to be perfectly balanced in case that $P(A) = P(B)$ (and, consequentially, $P(\neg A) = P(\neg B)$).

Table 12: Contingency tables for all combinations of the investigated methods, i.e., Ad-Hoc, Oaxaca-Blinder decomposition (OB-Dec.), regression (Regr.), Coupled Impact Assessment (C-IA), each contingency table showing also the product (hypothetical independent probability) of the marginals.

A	B	P(AB)	P(A$\overline{\text{B}}$)	P($\overline{\text{A}}$B)	P($\overline{\text{A}}$ $\overline{\text{B}}$)	P(A)	P(B)	P(A)P(B)
Ad-Hoc	OB-Dec.	48.50	24.43	17.92	9.15	72.93	66.43	*48.45*
Ad-Hoc	Regr.	60.73	12.20	22.54	4.53	72.93	83.28	*60.74*
Ad-Hoc	C-IA	52.83	20.10	18.25	8.82	72.93	71.08	*51.84*
OB-Dec.	Regr.	55.32	11.11	27.96	5.62	66.43	83.28	*55.32*
OB-Dec.	C-IA	48.13	18.29	22.95	10.62	66.43	71.08	*47.22*
Regr.	C-IA	60.36	22.91	10.72	6.00	83.28	71.08	*59.20*

Table 13: Confusion matrices for all combinations of the investigated methods,i.e., Ad-Hoc, Oaxaca-Blinder decomposition (OB-Dec.), regression (Regr.), Coupled Impact Assessment (C-IA), with N*=985,703 overall existing combinations of potential confounders.*

A	B	AB	A$\overline{\text{B}}$	$\overline{\text{A}}$B	$\overline{\text{A}}$ $\overline{\text{B}}$	A	B	A $\perp\!\!\!\perp$ B
Ad-Hoc	OB-Dec.	478,103	240,793	176,662	90,145	718,896	654,765	477,535
Ad-Hoc	Regr.	598,652	120,244	222,201	44,606	718,896	820,853	598,667
Ad-Hoc	C-IA	520,774	198,122	179,907	86,900	718,896	700,681	511,023
OB-Dec.	Regr.	545,283	109,482	275,570	55,368	654,765	820,853	545,261
OB-Dec.	C-IA	474,432	180,333	226,249	104,689	654,765	700,681	465,436
Regr.	C-IA	594,987	225,866	105,694	59,156	820,853	700,681	583,498

Table 14: p-value $\text{p}(z_{\kappa=0})$ *of the z-test for Cohen's* κ *under the null hypothesis that* $\kappa = 0$ *(*$H_0 : \kappa = 0$*), p-value of the* χ^2 *test for independence* $\text{p}(\chi^2_{A \perp\!\!\!\perp B})$*, and their difference up to 20 decimal points for all combinations of the investigated methods, i.e., the Ad-Hoc method, Oaxaca-Blinder Decomposition (OB-Dec.), linear-regression-based (Regr.), and C-IA-based.*

A	B	$\text{p}(z_{\kappa=0})$	$\text{p}(\chi^2_{A \perp\!\!\!\perp B})$	$\text{p}(z_{\kappa=0}) - \text{p}(\chi^2_{A \perp\!\!\!\perp B})$
Ad-Hoc	OB-Dec.	0.0064	0.0064	0.00000000000000122125
Ad-Hoc	Regr.	0.9270	0.9270	-0.00000000000030819791
Ad-Hoc	C-IA	0.0000	0.0000	0.00000000000000000000
OB-Dec.	Regr.	0.9018	0.9018	-0.00000000000024724667
OB-Dec.	C-IA	0.0000	0.0000	0.00000000000000000000
Regr.	C-IA	0.0000	0.0000	0.00000000000000000000

Table 15: Investigated drill-down patterns of confounding behaviour: pattern name; pattern description; method used in data case study.

Pattern	Pattern Description	Measure
Increased Confounding	A confounding effect is substantially increased as the result of adding further potential confounders to the analysis, see Section 8.3.	Ad-Hoc
Negative Explained Effect (specific to OB-decomposition)	The unexplained effect increases as the result of adding a further potential confounder, i.e, a negative explained effect occurs, see Section 8.4.	Oaxaca-Blinder Decomposition
Cancelling Out	A confounding effect is substantially reduced or even vanishes as the result of adding further potential confounders to the analysis, see Section 8.5.	Linear Regression
Time-Wise Change	A confounding effect changes over time, see Section 8.6.	C-IA

8.2 Identified Drill-Down Patterns of Confounding Behaviour

Overall, we have conducted 2,776 experiments, see Def. 19 and Figure 8. Beyond the *quantitative* investigations presented in Sect. 8.1, we started to carefully study the experiment outcomes *qualitatively*. In the beginning, we aimed at discovering interesting phenomena such as statical paradoxes [196] and any kind of data patterns [50]. After a while, our focus was narrowed on identifying patterns of confounding during drill-down, i.e., to be more precise, during drill-down into multiple potential confounders.

In a first round, we have briefly studied the outcomes of each experiment individually in order to gain an overview. In a second round, we delved deeper into investigating the experiment outcomes, individually and comparatively against each other. In our investigation, we used our Python code base, see Appendix A (see Publication V, IV, and II), enriched with data visualizations[42].

After these studies, we found four patterns of confounding worth to be communicated that we have summarized in Table 15.

The first pattern, called *Increased Confounding*, is about accumulative confounding effects. Here, some confounders have been found, and then, adding further potential confounders to the analysis increases the confounding effect. This pattern might seem rather intuitive and straightforward, but is important, in particular when compared to its dual pattern *Cancelling Out* to be discussed in due course. In Section 8.3, we present two data case studies that detect this pattern with the help of the Ad-Hoc method of confounder adjustment.

The second pattern, called *Negative Explained Effect*, is specific to Oaxaca-Blinder decomposition. Here, the unexplained effect found in an Oaxaca-Blinder decomposition increases as the result of adding a further potential confounder, i.e, a negative explained effect occurs. We investigate this pattern in two data

[42] `https://matplotlib.org/`

case studies in Section 8.4.

The third pattern, called *Cancelling Out*, is dual to the pattern *Increased Confounding*. Here, some confounders have been found, and then, adding further potential confounders to the analysis *decreases* the confounding effect. This pattern shows, how important it is, to conduct fine-grained, combinatorial analyses of confounding effects. In Chapter 4, we have discussed that it is common practice in many studies to utilize multiple linear regression to detect confounders. Here, usually, multiple potential confounders are added simultaneously to adjust for confounding. However, as the *Cancelling Out* pattern shows, important confounding effects might be overlooked, if individual potential confounders are not also analysed isolated or in smaller groups. Therefore, in Section 8.5, we present two data case studies that detect this pattern with the help of linear-regression-based confounder adjustment.

The fourth pattern is called *Time-Wise Change*. Here, the confounding effect of a group of confounders changes over time. In Section 8.6, we present two data case studies that detect this pattern with the help of linear-regression-based confounder adjustment.

Each of the confounding patterns shows in one of the data case studies in Sections 8.4 through 8.6. Table 15 includes information, in which of the Sections 8.4 through 8.6 each pattern is shown, and also, which measure is used to detect the pattern in the respective data case study.

8.3 Case-Studies on the Ad-Hoc Method

In this analysis, the Ad-Hoc method (32) is used to assess causal effects within 694 datasets. The analysis focuses on evaluating the causal impact of the primary influencing variable by systematically controlling for confounding variables. Incorporating potential confounding variables into the analysis ensures that the estimated causal effect represents the direct influence of the primary influencing variable more accurately, thereby minimizing distortions caused by external influences.

A key observation from the analysis is the role played by confounding variables in refining causal estimates. The confounding effect is indicated by the increase in percentage difference of potential confounders when adding to the analysis as a combination. Some variables show negligible or non-confounding effects after adding them individually to the analysis. To gain a deeper understanding of this effect, we will explore specific examples from our dataset.

8.3.1 Child Weight Dataset

The dataset captures details about children's age, weight z-scores, expenditures on food and health, and gender. The analysis focus on identifying the causal effect of age (in months) on weight z-scores while accounting for potential confounders.

The dataset consists of 9,168 observations with the following variables:

- Gender (X): Gender, serving as the primary influencing factor, measured as a categorical variable encoded as integers *(Primary Influencing Factor)*.
- weight_z (Y): Weight z-score, serving as the target variable, measured as continuous variable *(Target Variable)*.

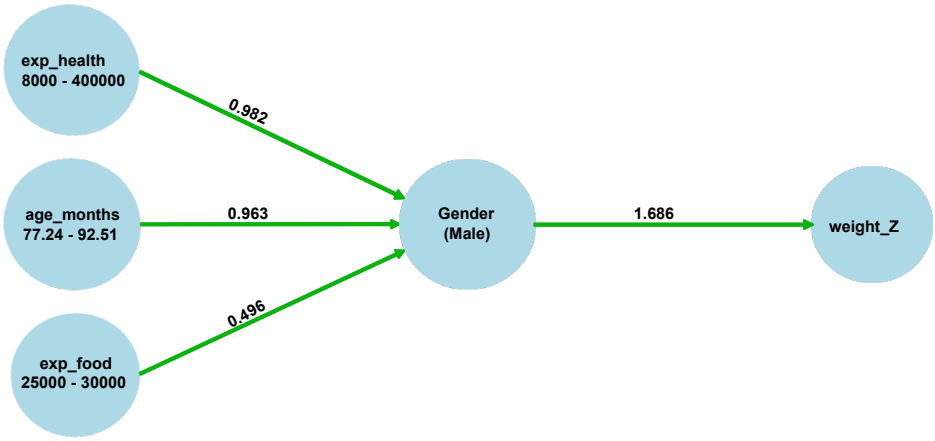


Figure 10: Impact graph of the Child Weight Dataset: Edges are annotated with lifts as labels. The information in the impact graph is incomplete. It highlights only the ARM lifts [4] between selected instances of potential confounders and a selected instance of the primary influencing factor Gender=Male*, as well as the generalized ARM lift (see (16) and [62]) between a selected instance of the primary influencing factor* Gender=Male *and the target value* weight_z*.*

This analysis aims to estimate the causal effect of *Gender* on *weight_z* while accounting for potential confounders using the Ad-Hoc formula (32). Several scenarios were considered, where potential confounders were introduced incrementally to adjust for their influence on the causal relationship between *Gender* and *weight_z*.

The initial analysis examined the relationship between *Gender* and *weight_z* without adjusting for any potential confounders. The observed value of *weight_z* for $Gender = 0$ was -1.686, while for $Gender = 1$, it was -1.677. The difference between these two values was -0.009. These results represent the marginal values, as no adjustments for confounding variables were made.

When the variable *age_months* was introduced as a confounder, the adjusted values of *weight_z* changed slightly. For $Gender = 0$, the adjusted value was -1.687, while for $Gender = 1$, it was -1.676. The difference between these adjusted values was -0.0114. This adjustment indicated a percentage difference of 23.56% compared to the marginal values. This notable percentage difference suggested that *age_months* had an influence on the causal relationship between

Gender and *weight_z*. Including this confounder brought to light the importance of age in months as a factor affecting weight outcomes.

The next confounder examined was *exp_food*, representing expenditure on food. After adjusting for this variable, the adjusted value of *weight_z* for $Gender = 0$ was -1.684, and for $Gender = 1$, it was -1.677. The difference between these two values was -0.007, corresponding to a percentage difference of -24.76% compared to the marginal values. The results indicated that expenditure on food had an impact on the causal effect of *Gender* on *weight_z*.

When *exp_health*, representing expenditure on health services, was introduced as an individual confounder, the results were more or less the same as the marginal values. The percentage difference was below of our cutoff rule (Def. 17). This result suggested that *exp_health* did not have a substantial confounding effect when considered individually.

The analysis was extended to examine combinations of potential confounders. When *age_months* was combined with *exp_health*, the percentage difference increased from 23.56% (the individual effect of *age_months*) to 28.57%, indicating a 5% increase. This highlighted the synergistic effect of combining these two variables in the adjustment process. Similarly, when *exp_food* was combined with *exp_health*, the percentage difference shifted significantly, reaching -38.84%. This substantial change demonstrated that *exp_food* and *exp_health* together influenced the causal effect of *Gender* on *weight_z* stronger than any of the potential confounders individually.

Finally, when adjustments were made for all potential confounders simultaneously, the pattern of increasing percentage differences persisted. The combined adjustment resulted in a percentage difference of 28.83%. This result underscored the importance of considering multiple confounders together, as their combined effects might reveal stronger confounding influences even when individual effects appeared negligible. The findings demonstrated that a drill-down approach, where adjustments are made incrementally and systematically, is essential for accurately identifying and accounting for confounding effects.

8.3.2 CEQ Assessment Guatemala

The analysis aimed to estimate the causal effect of *urban* on *tot_exp* while adjusting for potential confounders using the Ad-Hoc formula (32). To ensure comprehensiveness, the analysis included various combinations of potential confounders to assess their individual and combined influence on the relationship between *urban* living and total expenditure.

The dataset consists of 11,530 observations with the following variables:

- urban (X): Indicator of whether the household is located in an urban area, serving as the primary influencing factor, measured as a categorical variable encoded as integers *(Primary Influencing Factor)*.
- tot_exp (Y): Total expenditure of the household, serving as the target variable, measured as continuous variable *(Target Variable)*.
- final_Income (Z1): Total income, measured as a continuous variable *(Potential Confounder)*.
- no_persons (Z2): Persons in the household, measured as a continuous variable *(Potential Confounder)*.

- elec_pc (Z3): Electricity expenditure per capita, measured as a continuous variable *(Potential Confounder)*.
- fee_hlth_pc (Z4): Health-related expenditure per capita, measured as a continuous variable *(Potential Confounder)*.
- fee_educ_pc pc (Z5): Education-related expenditure per capita, measured as a continuous variable *(Potential Confounder)*.
- alc_pc pc (Z6): Alcohol expenditure per capita, measured as a continuous variable *(Potential Confounder)*.
- nal_pc (Z7): non-alcoholic beverages expenditure per capita, measured as a continuous variable *(Potential Confounder)*.
- cigr_pc (Z8): Tobacco expenditure per capita, measured as a continuous variable *(Potential Confounder)*.

For an overview over the CEQ Assessment Guatemala dataset and relationships among its factors, see Figure 11.

The initial phase of the analysis focused on the unadjusted relationship between *urban* and *tot_exp*. The observed value of *tot_exp* for $urban = 0$ was 212.92 , while for $urban = 1$, it was 152.72. The difference between these two values was 60.2. Since no adjustments were made for potential confounders at this stage, these values represented marginal effects. These results served as a baseline for subsequent comparisons after adjusting for confounders.

In the next stage, the analysis incorporated individual confounders to adjust the observed relationship. When *final_Income* was introduced as a confounder, the adjusted *tot_exp* for $urban = 0$ decreased to 167.78, and for $urban = 1$, it increased to 194.86. The difference between these adjusted values narrowed to 27.08, representing a substantial percentage change of 75.89% compared to the marginal values. Similarly, when *elec_pc* (electricity expenditure per capita) was included as a confounder, the adjusted *tot_exp* values were 165.4 for $urban = 0$ and 197.72 for $urban = 1$, resulting in a difference of 32.32. This represented a percentage change of 60.27% relative to the marginal values. These findings suggested that both *final_Income* and *elec_pc* had a significant confounding effect on the relationship between *urban* living and total expenditure, indicating their strong influence on the observed causal relationship.

In contrast, the introduction of *fee_educ_pc* (expenditure on education) as an individual confounder yielded results closely resembling the marginal values. The adjusted percentage difference was merely 1.69%, falling below our cutoff rule ("Def. 17"). This negligible change suggested that *fee_educ_pc* did not exert a substantial confounding effect when considered in isolation.

Further analyses examined the impact of additional potential confounders, including *no_persons*, *fee_hlth_pc*, *alc_pc*, *nal_pc*, and *cigr_pc*. When these variables were individually introduced as confounders, the adjusted values of *tot_exp* shifted notably. The percentage differences relative to the marginal values were 34.33% for *no_persons*, 40.82% for *fee_hlth_pc*, 27.88% for *alc_pc*, 25.79% for *nal_pc*, and 15.27% for *cigr_pc*. These results demonstrated that, except for *fee_educ_pc*, all other potential confounders had a measurable confounding effect on the relationship between *urban* and *tot_exp* when considered individually.

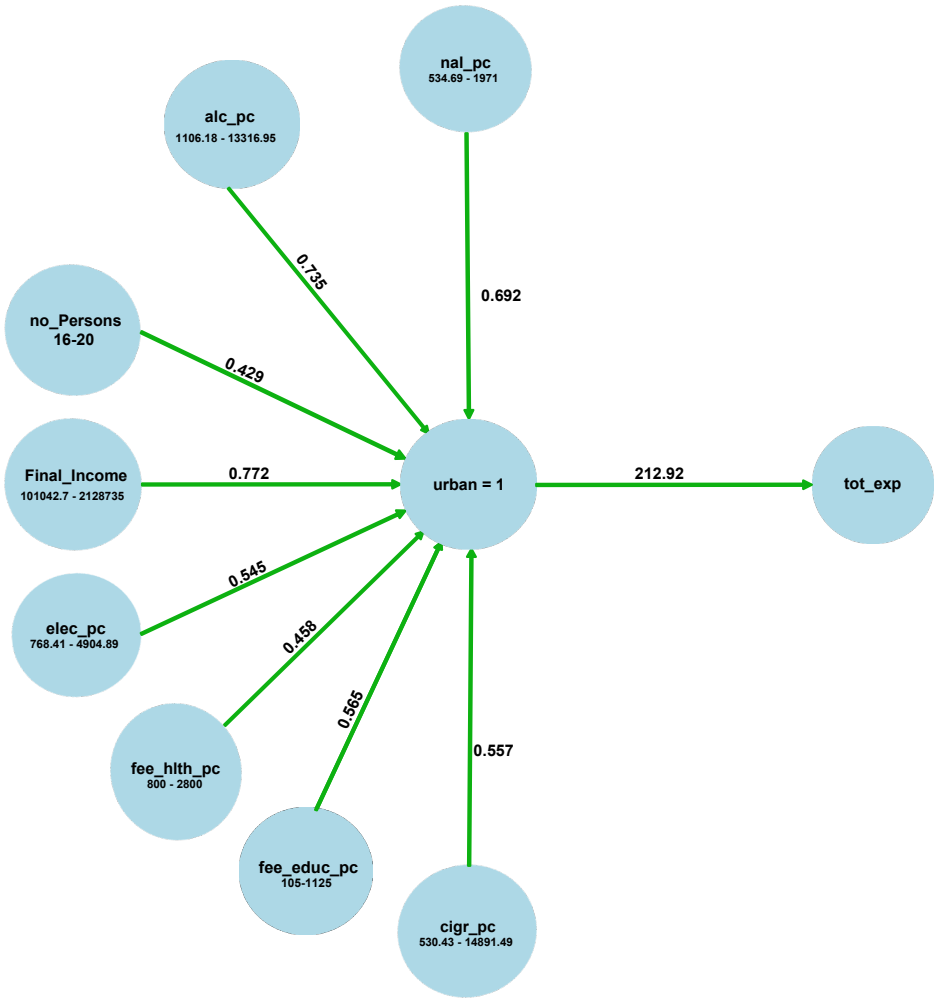


Figure 11: Impact graph of the CEQ Assessment Guatemala Dataset: Edges are annotated with the lift as labels. It highlights only the ARM lifts [4] between selected instances of potential confounders and a selected instances of the primary influencing factor urban = 1*, as well as the generalized ARM lift (see (16) and [62]) between the primary influencing factor* urban *and the target value* tot_exp*.*

The analysis then progressed to examine combinations of potential confounders to explore their combined effects on the relationship between *urban* living and *tot_exp*. Interestingly, when *fee_educ_pc*, which showed a negligible individual confounding effect, was combined with other variables, its impact became more pronounced. For instance, the combination of *fee_educ_pc* and *fee_hlth_pc* resulted in a percentage difference of 43.01% compared to the marginal values. This was a significant increase from the individual effects of *fee_educ_pc* (1.69%) and *fee_hlth_pc* (40.82%), suggesting a synergistic interaction between these two confounders. Similar patterns were observed when *fee_educ_pc* was combined with other variables. These findings highlighted the importance of considering combinations of confounders, as their joint effects might reveal underlying

interactions that are not apparent when variables are analysed individually.

To deepen the understanding of these interactions, the analysis considered all potential confounders simultaneously. The combined adjustment resulted in a dramatic increase in the percentage difference, reaching 152.46% relative to the marginal values. This finding underscored the cumulative effect of multiple confounders on the observed relationship between *urban* living and total expenditure. While some variables, like *fee_educ_pc*, exhibited negligible individual effects, their inclusion in the combined adjustment revealed significant interactions with other variables. This highlights the need for comprehensive approaches to confounder adjustment, as neglecting such interactions can lead to biased conclusions.

8.4 Case Studies on Oaxaca-Blinder Decomposition

The Oaxaca-Blinder decomposition is a statistical technique widely used to analyse group-based differences in a selected outcome variable, see Section 3.3. The method is particularly valuable for disentangling the sources of disparities observed between groups, as it decomposes the overall difference into two primary components: the explained component and the unexplained component.

The Oaxaca-Blinder decomposition identified less confounding effects than any other measures used in this study, i.e., 66.43%, see Table 12. This could be attributed to several factors, such as the method's relative complexity or its specific applicability to certain types of data. Unlike more straightforward methods, the Oaxaca-Blinder decomposition requires rigorous model specification and a deep understanding of the relationships between variables, which may make it less accessible or less commonly applied in certain contexts.

In this section, we identify a negative explained effect of confounding variable, which is an interesting phenomenon in the datasets. A negative explained effect indicates that the observed differences in characteristics should, in theory, reduce the gap – but instead, the gap persists or even increases. This implies that these improved characteristics should reduce the observed gap; however, they do not, because the unexplained component dominates.

To further demonstrate negative explained effect, we analyse some examples from our dataset.

8.4.1 K12Education

The dataset comprises comprehensive information about total wages of different persons, capturing different aspects of attributes such as employer name, position, elected , etc. The primary focus of the analysis is to investigate the effect of gender on total wages while carefully accounting for potential confounders.

- Gender (X): Gender, serving as the primary influencing factor, measured as a categorical variable encoded as integers *(Primary Influencing Factor)*.
- TotalWages(Y): Total Wages, serving as the target variable, measured as continuous variable *(Target Variable)*.
- EmployerName (Z1): The name of the employer, measured as a categorical variable *(Potential Confounder)*.
- Position (Z2): Position, measured as a categorical variable *(Potential Confounder)*.

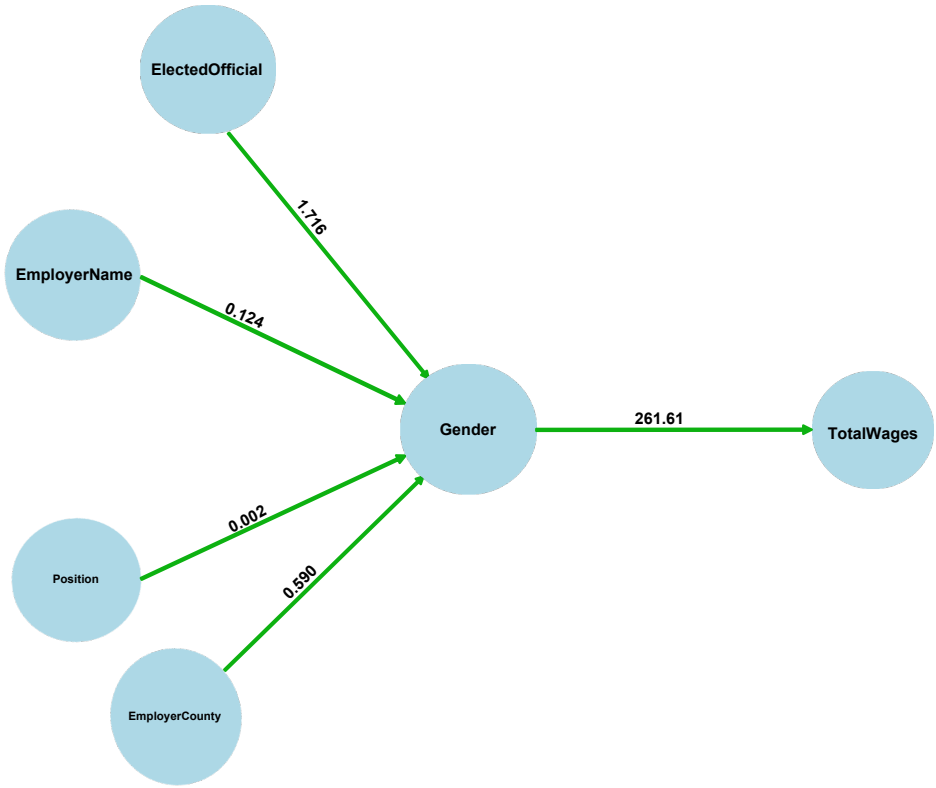


Figure 12: Impact graph of the K12Education Dataset: a green edge indicates a positive slope of the linear regression line between two nodes. Edges are annotated with the concrete slopes as labels. The information in the impact graph is incomplete. It highlights only slopes between potential confounders and the primary influencing factor Gender*, as well as the slope between the primary influencing factor* Gender *and target value* TotalWages*.*

When only *gender* is used as an explanatory variable, the explained effect is 0.00 and the unexplained effect is 262.24. None of the wage gap is attributed to observable differences in gender. The entire wage gap is left unexplained, suggesting that factors such as discrimination or unobservable characteristics contribute significantly to the gap. This serves as the baseline for subsequent models. These values are considered as marginal values for performing a drill-down operation to identify the impacts and patterns within the dataset.

Adding *EmployerName* as an explanatory variable accounts for a portion of the wage gap, as evidenced by the positive explained effect (62.95) and a reduced unexplained effect (206.70). This indicates that differences in employer

representation between groups contribute to the wage disparity. The significant decrease in the unexplained effect compared to the marginal value demonstrates that employer-related factors play a substantial role in reducing the unexplained component of the wage gap. However, despite this reduction, the majority of the gap remains unexplained, implying that additional factors not captured by *EmployerName* may also influence the wage gap.

Including *Position* and *ElectedOfficial* independently as explanatory variables results in negative explained effects of -8.37 and -1.97, respectively, while the unexplained effects are 270.6 and 264.21, respectively. These results indicate that the wage gap increases after accounting for positions. This counterintuitive finding may arise from overlapping effects between variables. Furthermore, the status of being an elected official provides minimal explanatory power, as shown by the small negative explained effect. Despite these adjustments, most of the wage gap remains unexplained, indicating that neither the elected official status nor the position significantly contribute to the wage disparity.

The most comprehensive model, which includes all available potential confounding variables, achieves the highest explained effect (64.05) among the models. This indicates that these variables collectively explain a significant portion of the wage gap. Additionally, the unexplained effect decreases to 198.19, suggesting that this model captures more variation in the data than simpler models. However, even with this more sophisticated model, a substantial portion of the wage gap remains unexplained.

Observable factors such as *EmployerName* and *Position* explain a meaningful portion of the wage gap, with *EmployerName* consistently contributing to a significant portion of the explained effect. This highlights the role of employer-level factors in wage disparities. However, variables such as *Position* and *ElectedOfficial* occasionally result in negative explained effects, suggesting potential biases i.e., how these variables interact with others.

The dominance of the unexplained effect across all models is evident in this dataset. Regardless of the explanatory variables used, the unexplained effect remains the largest contributor to the wage gap. This finding suggests that unmeasured or unobservable factors play a critical role in the observed disparities. This indicates that the gap persists or widens because individuals receive lower returns for these variables. It suggests that improving observable characteristics alone will not be sufficient to close the gap.

8.4.2 Payroll-2014-2015 Dataset

The dataset comprises comprehensive information about payroll for the year of 2014 and 2015, capturing different aspects of attributes such as agency name, hours worked, account description , etc. The primary focus of the analysis is to investigate the effect of *Year* on wages while carefully accounting for potential confounders.

The dataset contains payroll information for 2014 and 2015 from the city of Oklahoma, USA, comprising 903,131 records and six columns.

- Year (X): Year, serving as the primary influencing factor, measured as a categorical variable *(Primary Influencing Factor)*.
- Wages (Y): Wages, serving as the target variable, measured as continuous variable *(Target Variable)*.

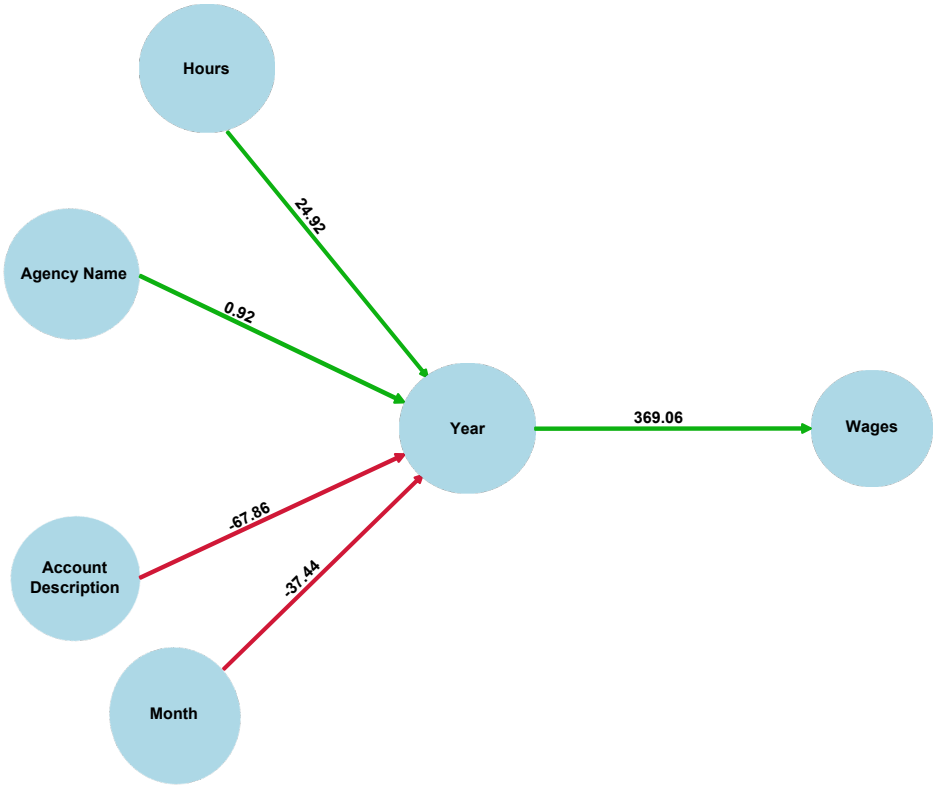


Figure 13: Impact graph of the Payroll-2014-2015 Dataset: a green edge indicates a positive slope and a red edge indicates a negative slope of the linear regression line between two nodes. Edges are annotated with the concrete slopes as labels. The information in the impact graph is incomplete. It highlights only slopes between potential confounders and the primary influencing factor Year*, as well as the slope between the primary influencing factor* Year *and target value* Wages*.*

The initial analysis focused on the decomposition of *Wages* where *Year* was the sole explanatory variable. The approach revealed that the explained effect was negligible, showing a value of 0.00, while the unexplained effect was 2563.04. These results indicate that *Year* alone does not account for any observed *Wages*

disparities. This outcome underscores the necessity of incorporating additional variables to capture the complexities of *Wages* differences and identify their sources.

A significant shift occurred when *Hours* was added alongside *Year* in the analysis. The inclusion of *Hours* dramatically increased the explained effect to a value 2246.48, while the unexplained effect sharply dropped to 316.56. This represents a percentage difference of -156.02% from the marginal value of the unexplained effect. The reduction in the unexplained effect demonstrates that *Hours* captures much of the variation that *Year* alone could not.

The inclusion of *Agency Name* as an additional factor alongside *Year* brought another layer of insight to the analysis. With this combination, the explained effect increased to 461.69, while the unexplained effect decreased from 2563.04 to 2101.35. However, despite the improvement, a significant unexplained portion of the *Wages* difference persists. This indicates that *Agency Name* alone are insufficient to fully account for the observed disparities.

In contrast, when *Account Description* and *Month* were added individually alongside *Year*, the explained effect turned negative (-92.32 and -591.95 , respectively), while the unexplained effect increased to 2655.36 and 3154.99. These negative explained effects indicate that the variations in these variables actually contribute to widening the gap. This counterintuitive outcome suggests that certain confounding factors associated with *Account Description* and *Month* may interact in ways that exacerbate disparities. These findings are particularly noteworthy, as they highlight the complex interplay of factors that can influence outcomes.

Interestingly, all combinations involving *Account Description* and *Month* consistently exhibited this negative explained pattern, except when *Hours* was included. This anomaly further underscores the significant role of *Hours* in mitigating disparities. The findings imply the following: while *Account Description* and *Month* may independently amplify disparities, their combined effect with *Hours* neutralizes, i.e., reduces, these impacts. This highlights the importance of drill-down analysis, as isolated effects may not capture the full picture of their influence on outcomes.

8.5 Case Studies on the Linear-Regression-Based Method

In this section, we investigate the confounding effects of various variables ("Z") on the relationship between primary influencing variable ("X") and target variable ("Y") using multiple regression analysis.

To further demonstrate its effectiveness, we will analyse some examples from our dataset.

8.5.1 Residential Survey Dataset

The dataset comprises comprehensive information about citizens from various states in India, capturing key demographic and socio-economic attributes such as average income, average expenditure, gender, religion, land area, age, and education levels. The primary focus of the analysis is to investigate the causal effect of age on average income while carefully accounting for potential confounders.

Age, as a critical demographic variable, is hypothesized to influence income levels through various channels such as experience and career progression. However, this relationship can be obscured or distorted by confounding factors like

education level, gender, or regional disparities in economic opportunities. For example, individuals with higher education levels may earn more, regardless of their age; or cultural norms around gender roles may impact income levels differently across groups.

The dataset consists of 14,851 observations with the following variables:

- Age (X): Age, serving as the primary influencing factor, measured as continuous variable. *(Primary Influencing Factor)*
- Avgincome (Y): Average income, serving as the target variable, measured as continuous variable. *(Target Variable)*
- Avgexp (Z1): Average expenditure, measured as a continuous variable *(Potential Confounder)*.
- Education (Z2): Education, measured as a categorical variable indicating the education level *(Potential Confounder)*.
- Gender (Z3): Gender, measured as a categorical variable encoded as integers *(Potential Confounder)*.
- Landarea (Z4): Land area, measured as a continuous variable *(Potential Confounder)*.
- Religion (Z5): Religion, measured as a categorical variable encoded as integers *(Potential Confounder)*.
- State (Z6): State, measured as a categorical variable indicating the states *(Potential Confounder)*.

See Figure 14 for an overall impact graph (Def. 3) for the dataset. This specific impact graph indicates the slopes of the regression line between each potential confounder and the primary influencing factor *(Age)* as a green edge (for positive slopes) or red edge (for negative slopes), together with the concrete slope as a label. Furthermore, it indicates the slope of the regression line between the primary influencing factor *(Age)* and the target value *(Avgincome)*.

In this analysis, age was the sole predictor of average income, and the analysis revealed a coefficient of 5.11 for age. This coefficient represents the marginal increase in average income corresponding to a one-unit increase in age. To assess whether other variables act as confounders, we incorporated, step by step, additional potential confounders into the regression model and analyzed their impact on the coefficient for age.

When gender was introduced into the regression model as an additional predictor, the coefficient for age increased slightly from 5.11 to 5.19, reflecting a percentage difference of 1.53%. This marginal change suggests that gender has a minimal confounding effect on the relationship between age and average income. Notably, the coefficient for gender itself was 20.48, but it did not pass our study's 10% cutoff rule (Def. 17), indicating that gender alone does not substantially influence average income in this context. These findings suggest that, while gender may have some effect on income, its role as a confounder in the age-income relationship is limited.

Introducing average expenditure into the model substantially altered the coefficient for age, reducing it from 5.11 to 4.27, a percentage difference of -17.90%.

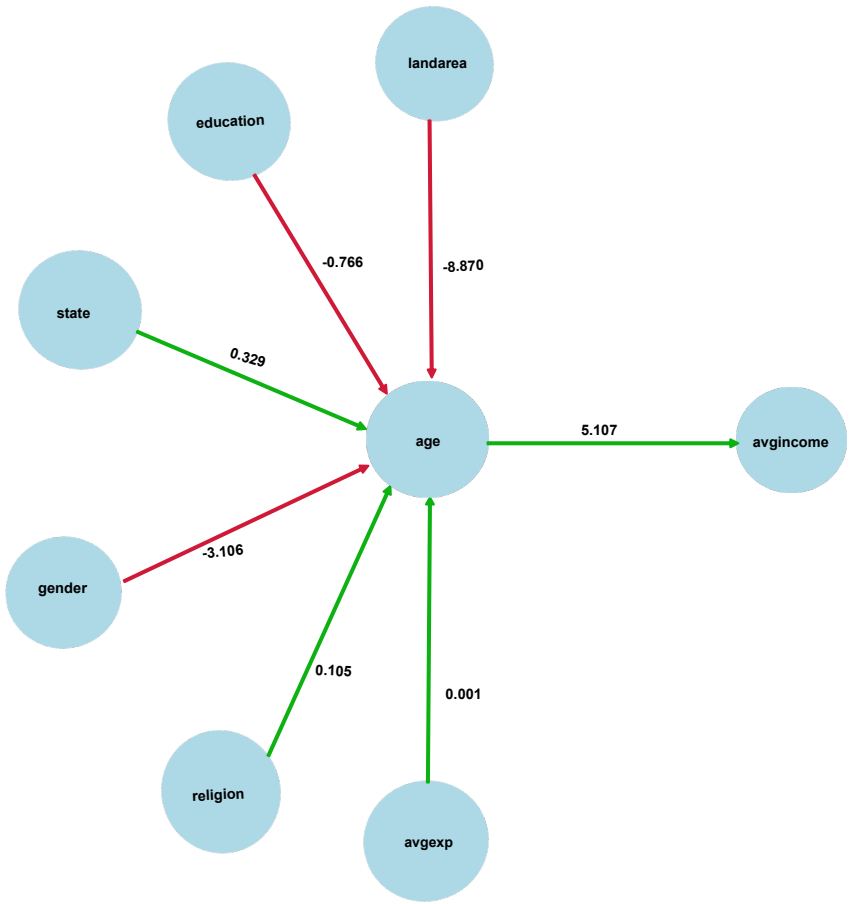


Figure 14: Impact graph of the Residential Survey Dataset: a green edge indicates positive slope and a red edge indicates a negative slope of the linear regression line between two nodes. Edges are annotated with the concrete slopes as labels. The information in the impact graph is incomplete. It highlights only slopes between potential confounders and the primary influencing factor Age*, as well as the slope between the primary influencing factor* Age *and the target value* Avgincome*.*

This substantial decrease in the age coefficient indicates that average expenditure acts as a strong confounder. Its inclusion suggests that part of the income variation previously attributed to age can actually be explained by expenditure patterns. This result highlights the importance of accounting for economic behaviours such as spending when analysing the determinants of income.

Education and state were added to the regression model to investigate their individual and combined effects. When education was included as a predictor, the coefficient for age increased significantly to 14.67, a percentage difference of 96.72% to the marginal coefficient. Similarly, when state was included independently, the coefficient for age increased to 7.99, a percentage difference of 43.90%. When both education and state were included together as predictors, the coefficient for age was 7.23, reflecting a combined percentage difference of 103.62%. These results suggest that both education and state strongly con-

found the age-income relationship, as they substantially influence the coefficient of age. Higher education levels are often associated with better-paying jobs, while state-level variations could reflect regional economic disparities.

When land area and religion were added as predictors, the changes in the age coefficient were negligible. The percentage differences were below our minimum threshold (Def. 17) required to indicate a confounding effect. While these variables might play a role in specific contexts or subgroups, their overall impact appears to be minimal in the current analysis.

When education and state were combined with average expenditure in the model, the confounding effect was effectively cancelled out, with the age coefficient decreasing slightly to 4.96, reflecting a percentage difference of -2.9%. This phenomenon suggests that adding multiple confounders can offset their individual influences, neutralizing the overall confounding effect. Interestingly, the combination of education and average expenditure also failed to meet the confounding criteria based on the cutoff threshold. This highlights the complex interplay between these variables, where their combined inclusion may account for overlapping variance or interactions that mitigate their individual impacts on the age-income relationship.

8.5.2 Cooking Energy Survey Dataset

This dataset contains information related to cooking energy access and household characteristics. The primary focus is on understanding factors that influence cooking expenditure, which serves as the dependent variable. The dataset includes several demographic, socioeconomic, and lifestyle variables, making it suitable for analysing relationships and confounding effects between household characteristics and their cooking-related spending.

The dataset is well-suited for regression and causal analysis, as it includes both continuous variables (e.g., cooking expenditure and household size) and categorical variables (e.g., ration card type, income source, and primary fuel). By examining these variables, we can investigate how household characteristics and resource availability impact cooking expenditures and identify potential confounding effects among the variables.

- hh_num (X): Number of persons in the house, serving as the primary influencing factor, measured as continuous variable *(Primary Influencing Factor)*.
- Cooking_expenditure (Y): The amount of money spent on cooking-related resources, serving as the target variable, measured as continuous variable *(Target Variable)*.
- hh_type (Z1): Type of household (e.g., rural or urban), measured as a categorical variable *(Potential Confounder)*.
- Age (Z2): Age, measured as continuous variable*(Potential Confounder)*.
- Education (Z3): Education, measured as a categorical variable indicating the education level *(Potential Confounder)*.
- Gender (Z4): Gender, measured as a categorical variable encoded as integers *(Potential Confounder)*.

- Religion (Z5): Religion, measured as a categorical variable encoded as integers *(Potential Confounder)*.
- type_rationcard: (Z6): Type of ration card held by the household indicating socio-economic classification, measured as a categorical variable *(Potential Confounder)*.
- monthly_income (Z7): Monthly income, measured as a categorical variable indicating income levels *(Potential Confounder)*.
- primary_cookstove (Z8): The type of primary cookstove used (e.g., LPG stove, traditional stove), measured as a categorical variable *(Potential Confounder)*.
- primary_fuel (Z9): The type of primary fuel used for cooking (e.g., firewood, LPG, kerosene), measured as a categorical variable *(Potential Confounder)*.
- income_source (Z10): Primary income source , measured as a categorical variable *(Potential Confounder)*.

See Figure 15 for an overall impact graph (Def. 3) for the dataset. Figure 15 indicates positive and negative linear regression slopes between potential confounders and the primary influencing factor *hh_num*, plus the slope between the primary influencing factor *hh_num* and the target value *Cooking_expenditure*.

The analysis investigates the presence of confounding effects in the relationship between the number of persons in a household (hh_num) and cooking expenditure (Cooking_expenditure) by incorporating socio-economic variables as potential confounders. Using regression analysis, the aim was to determine whether these variables altered the observed association between household size and cooking expenditure, indicating the presence of confounding effects.

In the initial regression model, the number of persons in a household was the sole predictor of cooking expenditure. The analysis revealed a coefficient of 278.99 for household size, indicating a positive relationship. This coefficient suggests that, on average, for every additional person in a household, cooking expenditure increases by approximately 278.99 money units. This finding is intuitive, as larger households typically consume more resources for cooking due to higher demand.

To explore potential confounding effects, primary fuel type (*primary_fuel*) was introduced as an additional predictor in the regression model. The inclusion of primary fuel type resulted in a slight decrease in the coefficient for household size, from 278.99 to 278.05, a percentage difference of 0.34%. This minimal change suggests that primary fuel type has a negligible confounding effect on the relationship between household size and cooking expenditure. However, the coefficient for primary fuel type itself was 272.30, indicating a significant positive association with cooking expenditure. This finding implies that the choice of fuel type significantly influences cooking expenditure.

When ration card type (*type_rationcard*) was introduced into the regression model, the coefficient for number of persons increased from 278.99 to 317.00, a percentage difference of 12.75%. This change suggests that the ration card type has a moderate confounding effect on the relationship between number of persons in a house and cooking expenditure. The ration card type often reflects a household's socio-economic status, with different card types providing varying levels of subsidies or access to resources [24]. Households with certain types

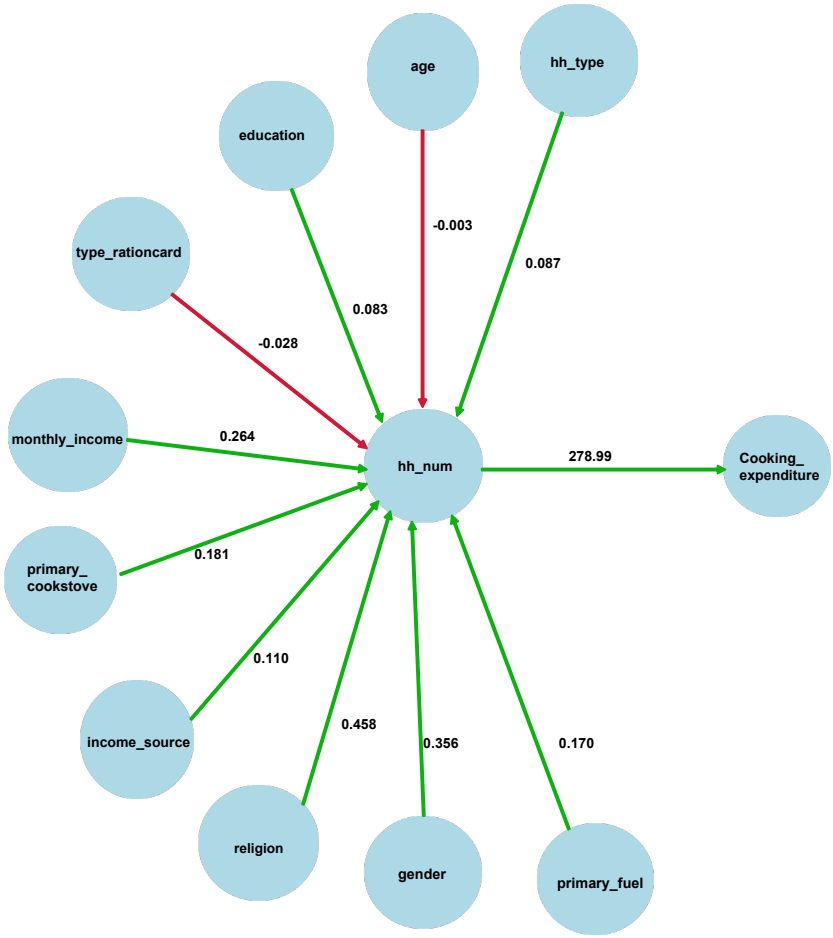


Figure 15: Impact graph of the Residential Cooking Energy Survey Dataset: a green edge indicates a positive slope and a red edge indicates a negative slope of the linear regression line between to nodes. Edges are annotated with the concrete slopes as labels. The information in the impact graph is incomplete. It highlights only slopes between potential confounders and the primary influencing factor hh_num*, as well as the slope between the primary influencing factor* hh_num *and the target value* Cooking_expenditure*.*

of ration cards may receive subsidized food or fuel, which could influence their cooking expenditure. Larger households, often eligible for higher subsidies, may experience reduced cooking costs, leading to this confounding effect.

When monthly income was included in the regression model, the coefficient for household size decreased significantly, from 278.99 to 203.00, reflecting a percentage difference of -31.54%. This substantial change indicates that monthly income acts as a strong confounder in the relationship between household size and cooking expenditure. Households with higher monthly incomes typically spend less on cooking in proportion to their income, as they may have access to more efficient cooking technologies, such as gas stoves or electric cookers, and benefit from subsidies or discounts on modern fuel types. Conversely, lower-income households may rely more heavily on traditional cooking methods and cheaper

fuels, leading to higher relative expenditures. Interestingly, the effect of monthly income was consistent across different combinations of predictors, reinforcing its significant role as a confounder.

When ration card type and monthly income were combined as predictors in the regression model, the confounding effect appeared to be canceled out. The coefficient for household size decreased slightly from 278.99 to 274.92, a percentage difference of 1.47%. This minimal change indicates that the combination of these two variables neutralized their individual confounding effects. The observed cancellation of the confounding effect may result from overlapping influences of ration card type and monthly income. Both variables are closely tied to socio-economic status and access to resources, and their combined inclusion in the model likely captures the majority of variance in cooking expenditure attributable to socio-economic factors. This overlap diminishes their individual impact on the relationship between household size and cooking expenditure.

8.6 Case Studies on the C-IA Method

In these case studies, we investigate the results of utilizing the C-IA method (78) for identifying confounding.

8.6.1 California State University Public Salary Dataset

This dataset provides comprehensive salary and benefit data for employees of the California State University system over multiple years (from 2009 till 2016). The primary goal is to analyze factors influencing *TotalWages*, which serves as the dependent variable. The analysis examines how different factors, including *EmployerCounty* differences and *Position* impact *TotalWages*. By investigating these variables, the study aims to uncover the primary influences on *TotalWages* while identifying potential confounding effects.

The dataset consists of 524,280 observations with the following variables:

- EmployerCounty (X): Employer County, serving as the primary influencing factor, measured as a categorical variable. *(Primary Influencing Factor)*.
- TotalWages (Y): Total Wages, serving as the target variable, measured as continuous variable. *(Target Variable)*.
- Position (Z1): Position, measured as a categorical variable. *(Potential Confounder)*.
- Gender (Z2): Gender, measured as a categorical variable encoded as integers. *(Potential Confounder)*.

For an overview over the California State University dataset and relationships among its factors, see Figure 16.

As an initial step, we have calculated the impact of *EmployerCounty* on *TotalWages* for all availabe years, with the maximum difference observed in 2009 (-193.21) and the minimum in 2012 (-182.47). The year-by-year analysis highlighted minor fluctuations, as depicted in Figure 17. Figure 17 demonstrates that while the values have shown some variation, the changes over time were relatively small, suggesting a stable influence of *EmployerCounty* on *TotalWages*.

When *Position* was added to the analysis, a significant shift in the results has been observed. The impact of *EmployerCounty* on the *TotalWages* for 2009

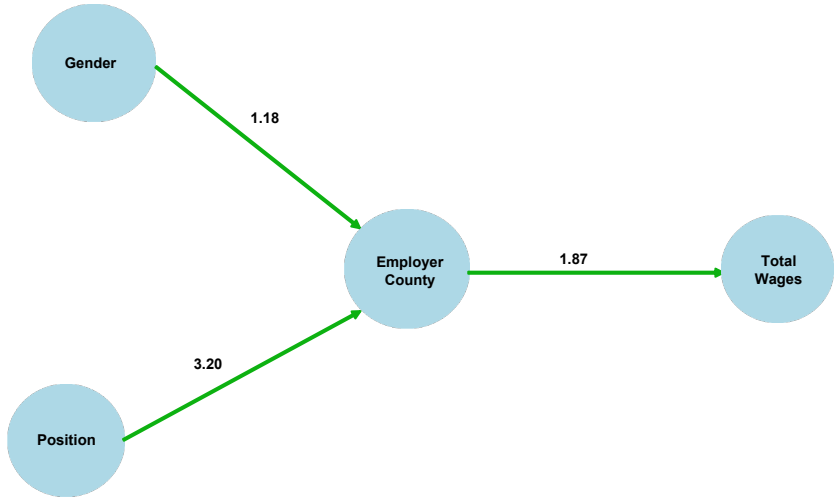


Figure 16: Impact graph of California State University Dataset. It highlights the C-IA impacts of the potential confounders onto the primary influencing factor EmployerCounty*, as well as the C-IA impact of the primary influencing factor* EmployerCounty *onto the target value* TotalWages*.*

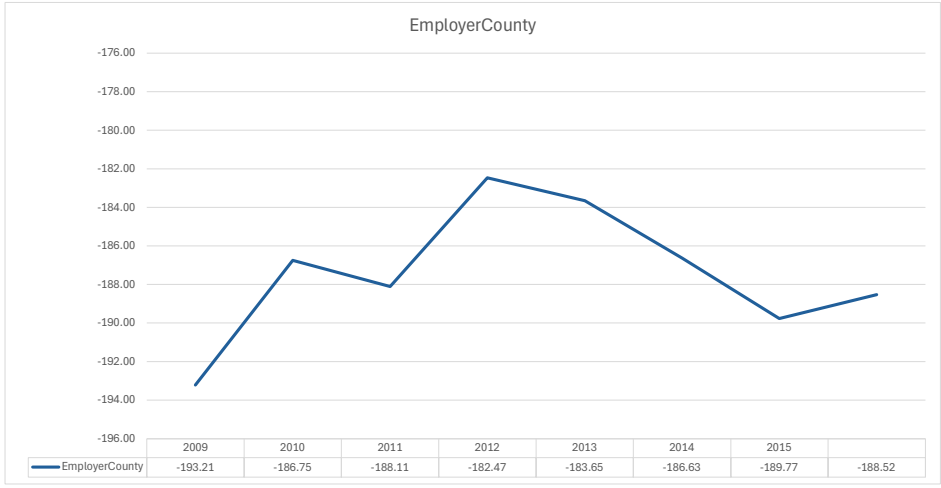


Figure 17: Impact of EmployerCounty *on* TotalWages *over the years.*

changed from -193.21 to -96.26, while for 2012, it changed from -182.47 to 32.10. This stark change signifies the profound influence of occupational roles on *TotalWages*. Figure 18 clearly expresses this transformation, where the graph illustrates a significant divergence as compared to Figure 17. The distance between points in the graph increased substantially, indicating a higher variability in the data when *Position* was included into the consideration.

Then, *Gender* was added to the model to examine its combined effect with *EmployerCounty* and *Position*, see Figure 19. Interestingly, the inclusion of *Gen-*

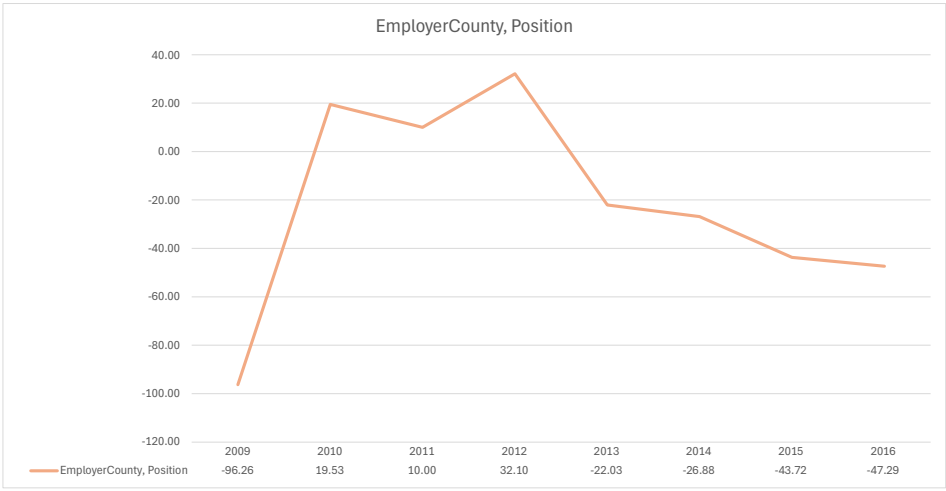


Figure 18: Impact of EmployerCounty *and* Position *on* TotalWages *over the years.*

der did not significantly alter the overall pattern observed in the case of only adding *Position* to the analysis (Figure 18). The impacts remained aligned with those seen in the case of only adding *Position,* suggesting that gender differences did not substantially influence the *TotalWages* beyond the occupational factors already accounted for. These fluctuations are depicted in Figure 19, which shows a pattern similar to Figure 18. This similarity reinforces the idea that *Position* is the dominant factor influencing the observed trends, whereas *Gender* rather plays a secondary role in shaping the outcomes.

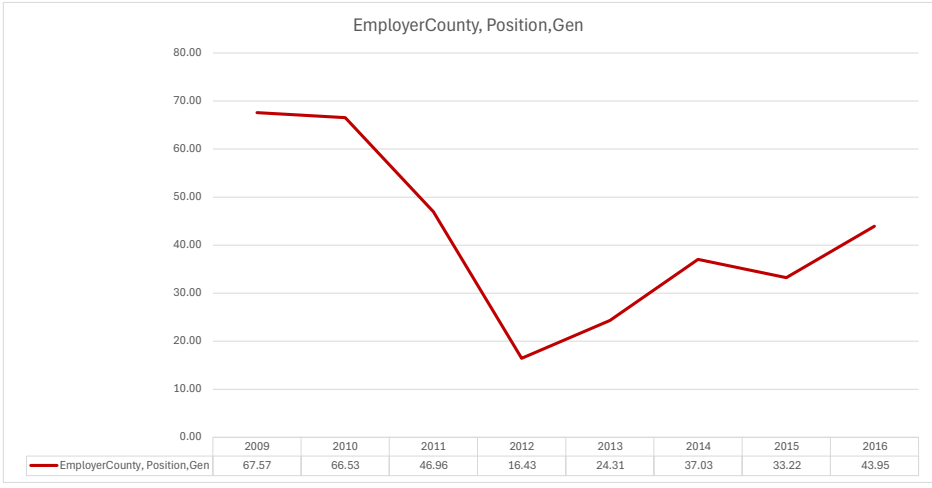


Figure 19: Impact of EmployerCounty, Position *and* Gender *on* TotalWages *over the years.*

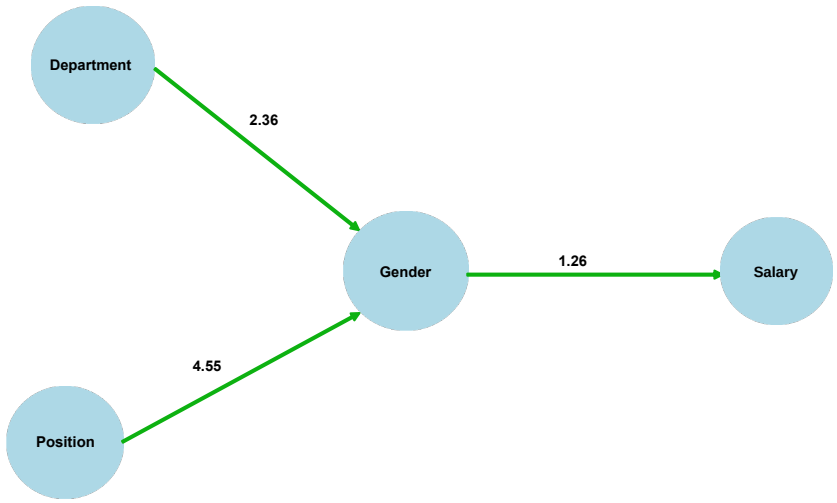


Figure 20: Impact graph of the Old Dominion College dataset. It highlights the C-IA impacts of the potential confounders onto the primary influencing factor Gender*, as well as the C-IA impact of the primary influencing factor* Gender *onto the target value* salary*.*

Initially, the analysis focused on the marginal effect of *Gender* on *Salary* across all available years. The maximum percentage difference was observed in 2016, reaching 116.71%, while the minimum was recorded in 2009 at 128.90%. Despite these year-specific variations, the overall differences across years were relatively small. As illustrated in Figure 21, the yearly percentage differences reflect small changes, suggesting that the influence of *Gender* on *Salary* has been relatively

stable during the analysis period. This stability implies that the disparities associated with *Gender* in terms of *Salary* have not undergone significant shifts.

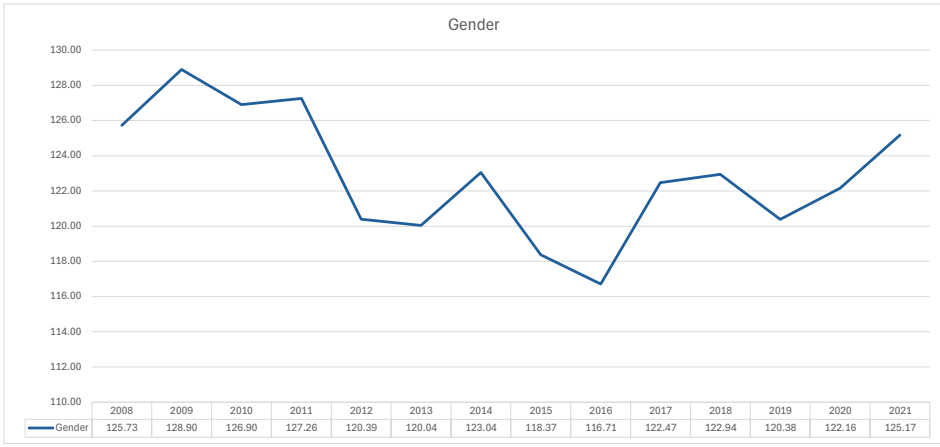


Figure 21: Impact of Gender *on* Salary *over the years.*

When the potential confounding variable *Department* was included in the analysis, the results followed a pattern similar to the marginal analysis of *Gender*. The observed changes remained relatively small, as shown in Figure 22. The maximum impact of *Department* on *Salary* in terms of percentage difference occurred in 2012, with a value of 11.47%. Conversely, the minimum impact was observed in 2017, at 16.18%, which amounts to a difference of 4.71%. For the year 2016, the inclusion of *Department* changed the percentage difference from 116.71% to 11.60%. Similarly, for 2009, the percentage difference changed from 128.90% to 11.74%. These adjustments demonstrate the moderate role of *Department* in the relationship between *Gender* and *Salary*. However, the overall stability of the impact across years, as illustrated in Figure 22, suggests that the influence of *Department* on *Salary* variations is relatively small.

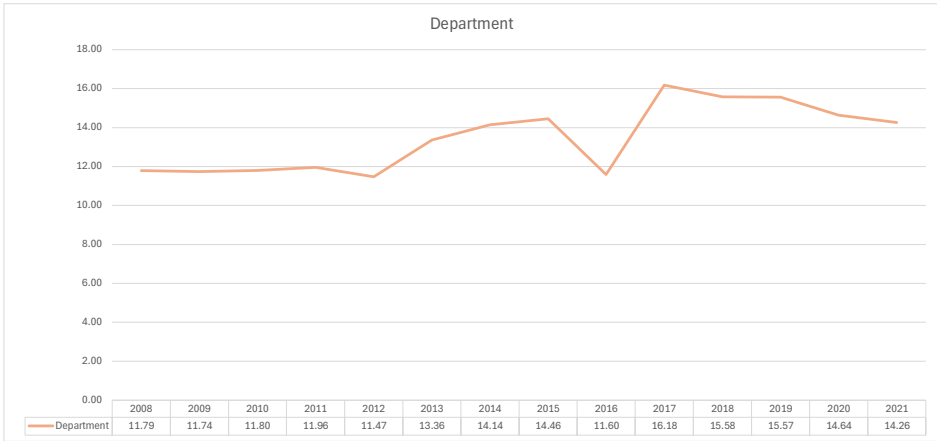


Figure 22: Impact of Department *on* Salary *over the years.*

The inclusion of the variable *Position* introduced significant changes to the observed patterns, indicating a substantial influence on *Salary*. Unlike the relatively stable trends observed in the analyses of *Gender* and *Department*, the percentage differences fluctuated significantly when *Position* was accounted for. The maximum percentage difference was recorded in 2012 at 48.73%, while the minimum was observed in 2021 at 10.35%. This range reflects a difference of 38.38% between the minimum and maximum values. The fluctuations in percentage differences, as shown in Figure 23, highlight the variability introduced by *Position*. The significant divergence in Figure 23, compared to Figures 21 and 22, underscores the critical role of occupational roles in shaping salary outcomes. Such insights emphasize the importance of considering *Position* when analysing salary-related disparities, as it appears to be a primary driver of variation.

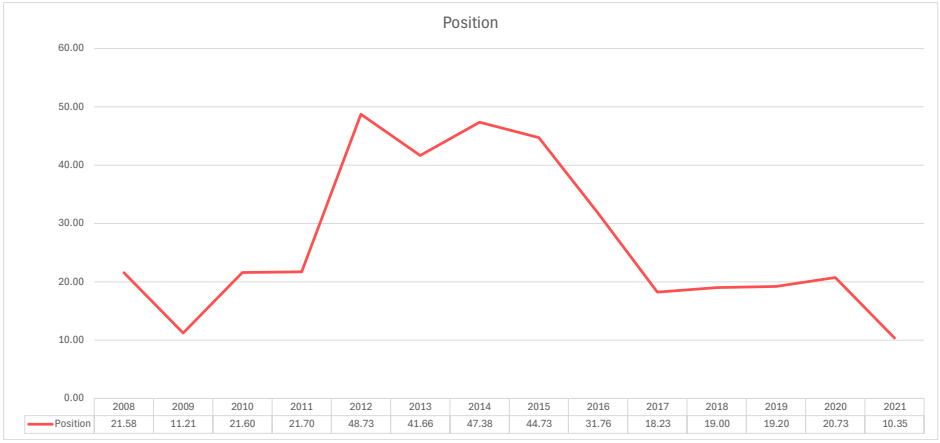


Figure 23: Impact of Position *on* Salary *over the years.*

To further explore the combined effects of these variables, both *Department* and *Position* were added to the model alongside *Gender*. The results of this comprehensive analysis are depicted in Figure 24. Interestingly, the inclusion of these potential confounders showed a pattern that closely resembled the one observed when only *Position* is included (Figure 23). This similarity suggests that *Department* variations did not substantially influence *Salary* beyond the variations already accounted for by *Position*. In the combined model (Figure 24), the maximum percentage difference was observed in 2018 at 64.75%, while the minimum occurred in 2016 at 10.16%, resulting in a difference of 54.59% between the maximum and minimum values. These fluctuations (Figure 24) further highlight the dominant role of *Position* in shaping *Salary* outcomes, with *Department* contributing minimally when considered alongside *Position*.

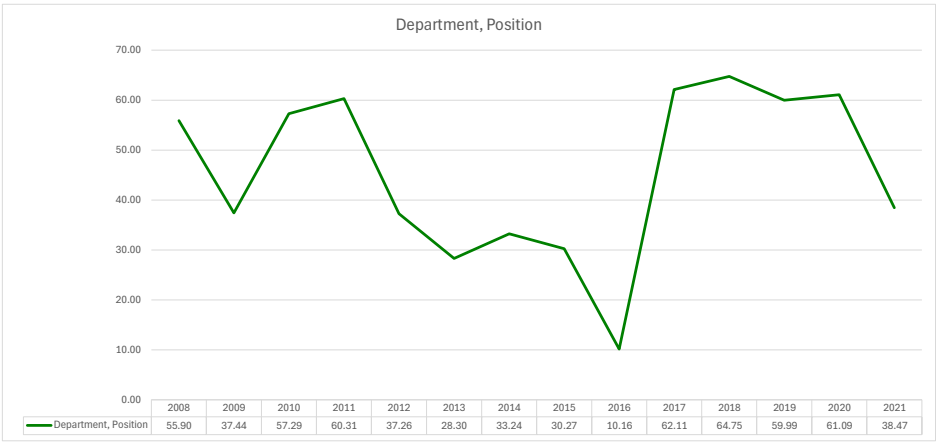


Figure 24: Impact of Department *and* Position *on* Salary *over the years.*

9 Future Directions

9.1 Further Measures for the Impact of Factors in their Entirety

The novelty of the C-IA in Def. 15 is that is allows for assessing the impact of categorical variable as a whole, i.e., in its entirety. However, the concrete C-IA in Def. 15 is not the only possibility to define such measure, and other should be tried out and compared with Def. 15.

Let us have a look at the definition of the C-IA measure (78). Now, let us call the expression

$$\left\| \frac{\mathsf{E}(y \mid X = v_j)}{\mathsf{E}(y \mid X = v_i)} \right\| \tag{85}$$

in (78) the *inner impact* of the C-IA measure. We have chosen the quotient of $\mathsf{E}(y \mid X = v_j)$ and $\mathsf{E}(y \mid X = v_j)$, i.e., their change factor, to define the inner impact of (78). We have chosen the change factor as basic measure, because it is also used in the definition of the *lift* impact measure in ARM [4] and, therefore, rather a familiar choice. Basically, replacing the inner impact of (78) based on any other basic measure[43] to compare $\mathsf{E}(y \mid X = v_j)$ and $\mathsf{E}(y \mid X = v_j)$ is one means to yield other measure for assessing the impact of a categorical variable in its entirety.

Remember that the C-IA measure is only defined for positive real numbers as target values. A concrete goal would be to define an impact measure for categorical values in their entirety that are defined for all real numbers, positive and negative. A first attempt would be to replace the inner impact of (78) by its absolute value yielding the following potential impact measure:

$$\sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n} \mathsf{P}(X = v_i)\mathsf{P}(X = v_j) \left| \left\| \frac{\mathsf{E}(y \mid X = v_j)}{\mathsf{E}(y \mid X = v_i)} \right\| \right| \tag{86}$$

Unfortunately, (86) is not a useful definition of an impact measures, because it works only properly whenever $\mathsf{E}(y \mid X = v_j)$ and $\mathsf{E}(y \mid X = v_i)$ have the same sign, i.e., if they are either both positive or both negative. For example, if $\mathsf{E}(y \mid X = v_j)$ and $\mathsf{E}(y \mid X = v_i)$ have different signs but the same absolute value $a \neq 0$, their inner impact value amounts neutral impact value 1[44], which is counter-intuitive, as the distance of their values on the number line amounts to $2a$ and, therefore, intuitively, the situation represents some impact. Here, utilizing the percentage difference as basic measure to define the inner impact, would be an interesting option.

Replacing the inner impact of (78) by utilizing other basic measures to compare $\mathsf{E}(y \mid X = v_j)$ and $\mathsf{E}(y \mid X = v_i)$ is only one means to create other, potentially interesting, impact measures. With (81), we have introduced an alternative impact measure that compares the effect[45] $\mathsf{E}(y \mid X = v_j)$ of the event $X = v_j$ onto the target variable y with the marginal expectation $\mathsf{E}(y)$ of y instead of comparing it with the effect $\mathsf{E}(y \mid X = v_i)$ of any other event $X = v_i$. We have tried to provide

[43] see Section 2.3, in particular, Table 2.

[44] In case of utilizing relative factors as basic measure, 1 needs to be considered as neutral impact value, representing no impact; whereas, e.g., in case of utilizing the percentage difference as basic measure, 0 needed to be considered as neutral impact value.

[45] or: confidence (when generalizing ARM [4] terminology from probabilities of target values to expectations of target variables, compare with Section 2.1).

some intuition about the difference of (81) and the C-IA measure (see page 59) and we have started a comparison of the two measures in terms of an artificial data example in Table 7. A more systematic investigation of the measure in (81) would be interesting.

Further potentially interesting measures can be created and then investigated by replacing the inner impact of (81)[46] by utilizing all kinds of different basic measures, as we have described for (78) and (85).

9.2 Integration of Confounding Patterns Into Tool Support

As described in Chapter 1, confounder adjustment should be integrated more systematically into data mining tools. For example, in Publication II, we have started this by integrating linear-regression-based confounder adjustment into our tool *GrandReport*, see Publications V and II. It would be interesting to integrate detection of the confounding patterns that we have found during our analyses (see Table 15 on page 72) into data mining tools and to validate the usefulness of such pattern detection. A first step would be to integrate such pattern detection into our tool *GrandReport*.

9.3 High-Performant Implementation of GrandReport

Our tool *GrandReport*[47,48] is a prototypical implementation to evaluate the usefulness [51] of grand reports [65, 62], see Publication V, and linear-regression-based detection of confounders, see Publication II. However, the *GrandReport* tool is a prototypical implementation and no particular effort has been invested into its performance. Yet, in today's instant economics [217, 216], decision support systems are expected to answer as quick as possible, ideally, in real-time. Therefore, it would be interesting to optimize the *GrandReport* tool for high performance, in service to evaluate its usefulness [51] with larger datasets and enabling deeper drill-downs in short response times. In particular, it would be interesting to re-implement the *GrandReport* tool on the basis of an underlying columnar in-memory database [169], as massive real-time analytical reporting is exactly what columnar in-memory databases are designed for. Given its design as columnar in-memory database [95] , its extensibility [66] and its availability[49] for research, we have identified the innovative Hyrise[50] technology as particularly promising platform for these endeavours, see Publication I.

9.4 Utilizing Our Findings for Machine Learning

In [71, 70], Esmaelizadeh et al. suggest to utilize the data science toolkit to analyse model outputs and model-performance metadata. In particular, they suggest to present significant *positive predictive values* of the trained model in the form of explanation tables[51] [50, 82, 81], each row drill-down in regard of one or more

[46] In case of (81), the inner impact has to be defined as $\|\mathrm{E}(y \mid X = v_i) : \mathrm{E}(y)\|$ accordingly.

[47] `http://grandreport.me`

[48] `https://github.com/istaltech/grandreport`

[49] `https://github.com/hyrise/hyrise/wiki`

[50] `https://hpi.de/plattner/projects/hyrise.html`

[51] An explanation table [50, 82, 81] is a particularly intuitive and convenient form of presenting association rules [4] for a fixed target variable together with some impact measure

instances of the training data set's input parameters. The suggested approach [71, 70] is an important into the direction of responsible and trustworthy artificial intelligence.

It would be interesting to integrate such model-performance analyses with detection of confounders and detection of our confounding patterns (see Table 15). A particular challenge would be in finding useful [51] measures for the significance of detected confounding effects and confounding patterns, that should be oriented towards or be consistent with the *Kullback-Leibler* (KL) [127] divergence-based approach of explanation tables [81].

9.5 Utilizing Neural Networks for Our Findings

Simpson's paradox[52] [233, 200], see also [62], is an extreme form of confounding effect, where the stratification towards the confounder seemingly reverts the effect of an exposure. In [195, 192, 194] we have implemented[53] and investigated the detection of Simpson's paradox in data from various perspectives, including fairness [194] and the generalization of Simpson's paradox to continuous variables [195]. Various other implementation of Simpson's paradox exists such as, e.g., [229] and [9, 10]. The implementations in [195, 192, 194, 229, 9, 10] have in common, that they implement the detection of Simpson's paradox in a straightforward manner in terms of the definition[54] of Simpson's paradox. Instead, Wang et al. [222] have implemented the detection of Simpson's paradox with a neural network called *SimNet*, claiming the advantage that their approach *"can discover various Simpson's paradoxes caused by discrete"*[55] *"and continuous variables,"*[56] *"even hidden variables."*[222]. In the same vein, it would be interesting to understand the potential of utilizing neural networks for our findings, i.e., for detecting C-IA impact (see Section 5), detection of confounding based on our C-IA measure (see Chapter 6 and Section 8.6) and detecting the confounding patterns that we have found during our analyses (see Table 15 on 72 and Section 9.2).

(such as ARM confidence [4] and positive predictive value). Explanation tables come with a notion of significance of rules to be included in tables [81], which is defined as informativeness based on *Kullback-Leibler* (KL) divergence [127] between the training and the predicted distribution.

[52] also called Yule-Simpson's paradox

[53] `https://github.com/rahul-sharmaa/SimpsonP`

[54] https://plato.stanford.edu/entries/paradox-simpson/

[55] compare to [192, 194, 229, 9, 10]

[56] compare to [195]

10 Conclusion

This thesis is devoted to the *study*, *improvement*, and *unification* of confounding measures. In any scientific discipline, analysing confounding effects and deconfounding is one of the critical ingredients of any high-quality quantitative scientific study. Confounding, a common problem in experimental and observational studies, occurs when the relationship between an exposure and an outcome is distorted by a third variable. Confounding is pervasive in scientific data and, therefore, is a severe risk for the accuracy of results and conclusions of scientific studies.

Based on the research findings with our data mining tool GrandReport, including behaviour of Pearson correlation during drill-down and linear-regression-based confounder adjustment, we have identified the need for more systematic understanding of confounding measures and confounding patterns. Therefore, we have decided to systematically compare four methods of confounding adjustments using the same datasets on a large scale.

In this thesis, we have contributed as follows. First, we have introduced a novel measure for the impact of categorical variables in their entirety, called Coupled Impact Assessment (C-IA). Furthermore, we have introduced a novel method to detect confounders utilizing the C-IA measure. Then, we have conducted combinatorially designed experiments with 694 datasets from the Harvard Dataverse and the NZ Government Repository to investigate three well-established approaches for detecting confounders, i.e., the Ad-Hoc method, Oaxaca-Blinder decomposition, and the linear-regression-based method, together with our own novel C-IA-based method.

Based on our experiment results, we have discovered, that the four investigated methods for detecting confounders do not show any relevant agreement or disagreement beyond chance (in terms of both Cohen's κ and Yule's ϕ). This is surprising result, as all of the four methods have been specifically designed for exactly the same target: to detect confounders, and, actually, three of the methods have been in widespread use over the decades in a plethora of scientific studies for detecting confounders. Additionally, we argue that the finding is highly relevant for the working data scientist.

Furthermore, based on our experiment results, we have identified four interesting patterns of confounding effects during drill-down into potential confounders, that we have showcased in eight data case studies.

We have elaborated a systematic interpretation of the linear regression model utilizing so-called multiplicative edges diagrams. We have utilize this interpretation to reflect on linear-regression-based confounding, including a critical discussion off cutoff rules for confounding adjustments.

Although confounding effects are ubiquitous in data and scientific studies, confounding is rather neglected in the common data mining tools, be it from Association rule mining (ARM) or Online analytical processing (OLAP). Here is, where this thesis aims at envisioning a paradigm shift, i.e., to design an integrate systematic support for treatment of confounding in data mining tools.

We have identified a series of interesting future research directions. It would be interesting to investigate further measures for the impact of categorical factors in their entirety, including such based on other basic measures and such based on comparing impacted values with marginals. It would be interesting to integrate the identified confounding patterns into our *GrandReport* tools. More-

over, it would be interesting to optimize the *GrandReport* tool for high performance, in service to evaluate its usefulness [51] with highest-volume datasets. In particular, it would be interesting to re-implement the tool on the basis of an underlying columnar in-memory database, as these have been designed for massive real-time analytical reporting. Also, it would be interesting to investigate our findings can be utilized in the machine learning pipeline, as well, in how far neural network model can be utilized for our findings.

Eventually, this thesis aims at extending the limits of confounding measures so that they can be utilized in scenarios of more informed, massive, automatic exploratory analysis to the benefit of today's decision-makers.

A Source Code

```
     # Categorical to numerical Coercion

def ca_coe(dfInput,infVariable,trgtVariable):

    for dtColumn in dfInput :

        # Checking for the categorical columns
        if dfInput[dtColumn].dtypes == 'O' and dtColumn in infVariable:

            dfavg=(
                dfInput.groupby(dtColumn)[trgtVariable].mean()
            ).sort_values(trgtVariable)

            # Avg of target column values according to the distince
            values of influvencing column and sort according to it
            (hot number encoding with groupby)
            dfavg['row_num'] = np.arange(len(dfavg))

            # Removing the hierarchical index of dataframe
            dfavg.reset_index(inplace=True)

            dfInput[dtColumn+"_cat"] =dfInput[dtColumn]

            #Assigning values to the coresponding column values
            for index, dtavgColumn in dfavg.iterrows() :

                dfInput = dfInput.replace(
                    {dtColumn: {dtavgColumn[dtColumn] :
                    dtavgColumn['row_num']}}
                    )

    return dfInput
```

```
    # Calculating the expected value

def expectation_function(y, x_value):

    # Mean value of each variable
    return np.mean(y[X == x_value])



    # Compare expected values with others

def compexpvalue(dfGroup,trgtVariable,infVariable):

    for i in range(len(dfGroup)):
        for j in range(len(dfGroup)):

             # Probability of X = v_i
            prob_X_i = np.mean(dfGroup == dfGroup[i][infVariable])

             # Probability of X = v_j
            prob_X_j = np.mean(dfGroup == dfGroup[j][infVariable])

             # Expected value calculation.
             expectation_ratio = abs(
                expectation_function(
                    dfGroup[trgtVariable],
                    dfGroup[j][infVariable]
                ) /
                expectation_function(
                    dfGroup[trgtVariable],
                    dfGroup[i][infVariable]
                )
            )

            impact += prob_X_i * prob_X_j * expectation_ratio

    print ("GROUP "+[i]+": " +impact)
```

```
    # C-IA Measure
    # Numerical-to-Categorical Coercion
def fciameasure(dfInput, infVariable, trgtVariable):

    # numerical checking
    if 'int64' in list(dfInput.dtypes) :

        for dtColumn in dfInput :

            if (dfInput[dtColumn].dtypes == 'int64' or
                dfInput[dtColumn].dtypes == 'float64')
                and dtColumn in infVariable:

                minMAxDiff=dfInput[dtColumn].max() -
                    dfInput[dtColumn].min() # Min Max Difference

                if(dfInput[dtColumn].min()==0):
                    minMAxDiff=minMAxDiff/5

                else :
                    minMAxDiff=minMAxDiff/4

                for i in range(5):

                    if(i==0):

                        dfInput[dtColumn] = np.where(
                            (dfInput[dtColumn] <= minMAxDiff * (i+1)),
                            i,
                            dfInput[dtColumn]
                        )


                    else :
                        dfInput[dtColumn] = np.where(
                            (dfInput[dtColumn] > minMAxDiff * (i) &
                            dfInput[dtColumn] <=minMAxDiff * (i+1)),
                            i,
                            dfInput[dtColumn]
                        )
    dfGroup = dfInput.groupby(infVariable)
    # Calling expecetd value comparison function
    compexpvalue(dfGroup,trgtVariable,infVariable)

    return 0
```

```
    #Oaxaca-Blinder decomposition

def fobd(dfInput,infVariable,trgtVariable,groupCol):

    # Categorical to numerical Coercion
    dfInput = ca_coe(dfInput,infVariable,trgtVariable)
    model=[]
    xmean = []
    ymean = []
    dfGroup=[]

    if (groupCol !=''):
        for index,i in dfInput.groupby(groupCol) :

            x = sm.add_constant(i[infVariable])
            model.append(
                sm.OLS(i[trgtVariable],
                x.astype(float)).fit()
            )
            xmean.append(x.mean())

            ymean.append((i[trgtVariable]).mean())

        # Calling Oaxaca-Blinder decomposition calcualation function
        fobdCmp(model,xmean,ymean)

    return 0

    # Calculating Oaxaca-Blinder decomposition

def fobdCmp(model,xmean,ymean) :

    # Calculating explanined part
    explained = (xmean[0] - xmean[1])
        .dot((model[0].params + model[1].params) / 2)

    # Calculating unexplained part
    unexplained = (model[0].params - model[1].params)
        .dot((xmean[0] + xmean[1]) / 2)

    # Calculating total difference
    total_difference = ymean[0] - ymean[1]

    print("Difference in coefficients:", total_difference)
    print("explained:", explained)
    print("unexplained:", unexplained)

    return 0
```

```
    # Linear-Regression based adjustment

def fregAdj (dfInput, infVariable, trgtVariable):

    # Categorical to numerical Coercion
    dfInput = ca_coe(dfInput,infVariable,trgtVariable)

    for dtColumn in infVariable :

        x = sm.add_constant(dfInput[dtColumn]) # adding a constant
        model = sm.OLS(dfInput[trgtVariable], x).fit()
        predictions = model.predict(x)

        # Calling print function
        fprint_reg(
            model.params,
            (x.drop(['const'],
            axis=1)).keys(),
            trgtVariable
        )

    return model

    # Printing Linear-Regression results

def fprint_reg(model, infVariable, trgtVariable):

    # Printing the constent
    print(trgtVariable[0] + ' (Const) : ' + str(D(model['const'])))

    # Printing the slops (Coefficients)
    for variableFields in infVariable :
        print(variableFields + ' : ' + str(D(model[variableFields])))

    print("\n")
```

```
    # Ad-Hoc based adjustmet

def fadhmethod(dfInput, infVariable, trgtVariable):

    # numerical checking
    if 'int64' in list(dfInput.dtypes) :

        for dtColumn in dfInput :

            if (dfInput[dtColumn].dtypes == 'int64' or
                dfInput[dtColumn].dtypes == 'float64')
                and dtColumn in infVariable:

                minMAxDiff=dfInput[dtColumn].max() -
                    dfInput[dtColumn].min() # Min Max Difference

                if(dfInput[dtColumn].min()==0):
                    minMAxDiff=minMAxDiff/5

                else :
                    minMAxDiff=minMAxDiff/4

                for i in range(5):

                    if(i==0):

                        dfInput[dtColumn] = np.where(
                            (dfInput[dtColumn] <= minMAxDiff * (i+1)),
                            i,
                            dfInput[dtColumn]
                        )


                    else :
                        dfInput[dtColumn] = np.where(
                            (dfInput[dtColumn] > minMAxDiff * (i) &
                            dfInput[dtColumn] <=minMAxDiff * (i+1)),
                            i,
                            dfInput[dtColumn]
                        )

    grouped = dfInput.groupby(infVariable)

    # Calculate conditional expectation
        except primary influencing factor
    conditional_means = grouped[trgtVariable].mean().
                        reset_index(name='E_Y_given_X_Z')
```

```
# Calculate marginal probability
z_group_counts = dfInput.groupby(infVariable[1:]).size()
z_marginal_probabilities = z_group_counts / len(dfInput)
z_marginal_probabilities = z_marginal_probabilities.
                           reset_index(name='P_Z')

# Merge conditional expectations with marginal probabilities
merged_data = conditional_means.merge(
                z_marginal_probabilities,
                on=trgtVariable
             )

# Calculate the weighted contribution for P(Y | do(X))
merged_data['Weighted_Contribution'] = (
                merged_data['E_Y_given_X_Z']
                * merged_data['P_Z']
        )

# Sum over Z to get P(Y | do(X)) for each level of X
results = merged_data.groupby(X)
            ['Weighted_Contribution'].sum()
            .reset_index(name='P_Y_do_X')

return results
```

```
    # Main function

def main():

    pd.set_option("future.no_silent_downcasting", True)

    # excel file path
    mstrfle = pd.ExcelFile("file path")

    # reading the master sheet
    excelMaster = pd.read_excel(mstrfle, 'MasterSheet')


    # dataframe creation according to master sheet
    dfMaster=pd.DataFrame(excelMaster,
            columns= [
                        'sheetName',
                        'targetVariable',
                        'variableFields',
                        'stratifyVariable',
                        'groupCol'
                    ]
            )

    # removing null values
    dfMaster = dfMaster.fillna({'stratifyVariable': ''})
    dfMaster = dfMaster.fillna({'groupCol': ''})
    dfMaster = dfMaster.reset_index()

    excelReaderold=""

    try:
        for index,dfRows in dfMaster.iterrows():

            #  reading data sheet
            excelReader = pd.read_excel(mstrfle, dfRows['sheetName'])

            # influencing variables
            infVariable=list(dfRows['variableFields'].split(","))

            # target variable
            trgtVariable=list(dfRows['targetVariable'].split(","))

            dfInput= pd.DataFrame(
                        excelReader,
                        columns= (
                            infVariable
                            +
```

```
                        trgtVariable
                    )
                )

            if excelReaderold != dfRows['sheetName']:
                print (dfRows['sheetName'])

            print('Influencing variables : ' + dfRows['variableFields'])
            print('Target variables : ' + dfRows['targetVariable'])
            print('Stratify variables : ' + dfRows['stratifyVariable'])
            print('Grouping variables : ' + dfRows['groupCol'])


            # Calling Linear-Regression based adjustment function
            fregAdj (dfInput,infVariable,trgtVariable)

            # Calling Oaxaca-Blinder decomposition function
            fobd(dfInput,infVariable,trgtVariable,groupCol)

            # Calling Ad-Hoc based adjustmet function
            fadhmethod(dfInput, infVariable, trgtVariable)

            # Calling C-IA Measure function
            fciameasure(dfInput, infVariable, trgtVariable)

    except NameError:
        print(NameError)

main()
```

B Auxiliary Lemma

Lemma 3 (Estimator Coefficients in Terms of Pearson Corrletation)**.** *Given two numerical factors $x : \Omega \longrightarrow \mathbb{R}$ and $y : \Omega \longrightarrow \mathbb{R}$ and the linear regression problem*

$$\langle y_i = \beta_0 + \beta_1 x_{i_1} + \epsilon_i \rangle_{1 \leqslant i \leqslant N} \tag{87}$$

we have that the following holds for their fitted regression line:

$$\hat{\beta}_1 = \rho_{x,y} \frac{\sigma_y}{\sigma_x} \tag{88}$$

Proof. We have that (see [227] Equation 2.19, page 29):

$$\hat{\beta}_1 = \frac{\sum_{i=1} \Big(x_i - \mathsf{E}(X) \Big) \Big(y_i - \mathsf{E}(Y) \Big)}{\sum_{i=1} \Big((x_i - \mathsf{E}(Y) \Big)^2} \tag{89}$$

Due to (89) and the definitions of covariance, variance and the Pearson correlation coefficient (see, e.g., [149], p.14), we have the following:

$$\hat{\beta}_1 = \frac{N \mathrm{cov}(x,y)}{N \sigma_x^2} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \rho_{x,y} \frac{\sigma_y}{\sigma_x} \tag{90}$$

□

List of Figures

List of Tables

References

[1] A. Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005. doi:10.1111/0034-6527.00321.

[2] A. Abdullin and O. Nasraoui. Clustering heterogeneous data sets. In *Proceedings of La-WEB'2012 – the 8th Latin American Web Congress*, pages 1–8. IEEE, 2012. doi:10.1109/LA-WEB.2012.27.

[3] A. Abelló and O. Romero. On-line analytical processing. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 1–7. Springer New York, 2016. doi:10.1007/978-1-4899-7993-3_252-2.

[4] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993. doi:10.1145/170036.170072.

[5] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of VLDB'1994 – the 20th International Conference Very Large Data Bases*, volume 1215, pages 487–499. Morgan Kaufmann Publishers, 1994.

[6] M. Aickin. Conceptualization and analysis of mechanistic studies. *Journal of Alternative and Complementary Medicine*, 13(1):151 – 158, 2007. doi:10.1089/acm.2006.6341.

[7] A. N. Albatineh, H. M. Khan, and M. Niewiadomska-Bugaj. On the equivalence of some indices of similarity: Implication for binary presence/absence data. *Australian & New Zealand Journal of Statistics*, 54(2):189–198, 2012. doi : 10.1111/j.1467-842X.2012.00674.x.

[8] N. Ali, D. Neagu, and P. Trundle. Classification of heterogeneous data based on data type impact on similarity. In A. Lotfi, H. Bouchachia, A. Gegov, C. Langensiepen, and M. McGinnity, editors, *Advances in Computational Intelligence Systems*, volume 840 of *Advances in Intelligent Systems and Computing*, pages 252–263, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-319-97982-3_21.

[9] N. Alipourfard, P. G. Fennell, and K. Lerman. Can you trust the trend?: Discovering simpson's paradoxes in social data. In *Proceedings of WSDM'2018 – the 11th ACM International Conference on Web Search and Data Mining*, pages 19–27. Association for Computing Machinery, 2018. doi:10.1145/3159652.315968.

[10] N. Alipourfard, P. G. Fennell, and K. Lerman. Using simpson's paradox to discover interesting patterns in behavioral data. In *In Proceedings of ICWSM 2018 – the 12th International Conference on Web and Social Media*, pages 2–11. AAAI Press, 2018. https://ojs.aaai.org/index.php/ICWSM/article/view/15017/14867.

[11] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81, 2013. doi:10.1109/MIC.2013.20.

[12] C. Andrade. Confounding. *Indian Journal of Psychiatry*, 49(2):129–131, 2007. 10.4103/0019-5545.33263.

[13] C. Andrade. Confounding by indication, confounding variables, covariates, and independent variables: Knowing what these terms mean and when to use which term. *Indian Journal of Psychological Medicine*, 46(1):78–80, 2024. doi:10.1177/02537176241227586.

[14] H.-J. Andreß, K. Golsch, and A. W. Schmidt. *Applied Panel Data Analysis for Economic and Social Surveys*. Springer, Berlin, Heidelberg, 2013. doi10.1007/978-3-642-32914-2.

[15] S. Arakkal Peious, M. Kaushik, S. A. Shah, R. Sharma, S. Suran, and D. Draheim. On measuring confounding bias in mixed multidimensional data. In J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, editors, *Proceedings of ICTIS'2024 – the 8th International Conference on ICT for Intelligent Systems, Volume 6*, volume 1112 of *Lecture Notes in Network and Systems*, pages 329–342, Singapore, 2024. Springer Nature Singapore. doi:10.1007/978-981-97-6684-0_27.

[16] S. Arakkal Peious, M. Kaushik, M. Shahin, R. Sharma, and D. Draheim. On choosing the columnar in-memory database Hyrise as high-performant implementation platform for the GrandReport tool. In *Proceedings of NGISE'2025 – the 1st International Conference in Next Generation Information Systems Engineering*, pages 1–7. IEEE, 2025. to appear, preprint available at IEEE TechRxiv, doi: 10.36227/techrxiv.174015861.15410981/v1.

[17] S. Arakkal Peious, R. Sharma, M. Kaushik, S. Mahtab, and D. Draheim. On observing patterns of correlations during drill-down. In P. D. Haghighi, E. Pardede, G. Dobbie, V. Yogarajan, N. A. Sanjaya, G. Kotsis, and I. Khalil, editors, *Proceedings of iiWAS'2023 – the 25th International Conference on Information Integration and Web-based Applications and Services*, volume 14416 of *Lecture Notes in Computer Science*, pages 134–143, Cham, 2023. Springer. doi:10.1007/978-3-031-48316-5_16.

[18] S. Arakkal Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia. Grand reports: A tool for generalizing association rule mining to numeric target values. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery*, volume 12393 of *Lecture Notes in Computer Science*, pages 28–37, Cham, 2020. Springer. doi:10.1007/978-3-030-59065-9_3.

[19] S. Arakkal Peious, S. Suran, V. Pattanaik, and D. Draheim. Enabling sensemaking and trust in communities: An organizational perspective. In E. Pardede, M. Indrawan-Santiago, P. D. Haghighi, M. Steinbauer, I. Khalil, and G. Kotsis, editors, *Proceedings of iiWAS'2021 – the*

23rd International Conference on Information Integration and Web Intelligence, pages 95–103. Association for Computing Machinery, 2021. doi:10.1145/3487664.3487678.

[20] J. K. Aronson, A. La Caze, M. P. Kelly, V.-P. Parkkinen, and J. Williamson. The use of mechanistic evidence in drug approval. *Journal of Evaluation in Clinical Practice*, 24(5):1166 – 1176, 2018. doi:10.1111/jep.12960.

[21] S. Arslanturk, S. Draghici, and T. Nguyen. Integrated cancer subtyping using heterogeneous genome-scale molecular datasets. *Pacific Symposium on Biocomputing*, 25(2020):551–562, 2020. https://pubmed.ncbi.nlm.nih.gov/31797627/.

[22] J. Z. Ayanian, B. E. Landon, J. P. Newhouse, and A. M. Zaslavsky. Racial and ethnic disparities among enrollees in medicare advantage plans. *New England Journal of Medicine*, 371(24):2288–2297, 2014. doi:10.1056/NEJMsa1407273.

[23] B. Azvine, Z. Cui, D. D. Nauck, and B. Majeed. Real time business intelligence for the adaptive enterprise. In *Proceedings of CEC/EEE'06 – the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services*, pages 29–29. IEEE, 2006. doi:10.1109/CEC-EEE.2006.73.

[24] S. Balani. Functioning of the public distribution system: An analytical report. Working Papers id:5628, eSocialSciences, 2014.

[25] R. Barillec, B. Ingram, D. Cornford, and L. Csató. Projected sequential Gaussian processes: A C++ tool for interpolation of large datasets with heterogeneous noise. *Computers and Geosciences*, 37(3):295 – 309, 2011. doi:10.1016/j.cageo.2010.05.008.

[26] S. Bazen. *Econometric Methods for Labour Economics*. Oxford University Press, 2011.

[27] M. Bertl, S. Mahtab, P. Ross, and D. Draheim. Finding indicator diseases of psychiatric disorders in big data using clustered association rule mining. In *Proceedings SAC'2022 – the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 826–833, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3555776.3577594.

[28] J. H. Bjørngaard, A. T. Nordestgaard, A. E. Taylor, J. L. Treur, M. E. Gabrielsen, M. R. Munafó, B. G. Nordestgaard, B. O. Åsvold, P. Romundstad, and G. D. Smith. Heavier smoking increases coffee consumption: Findings from a mendelian randomization analysis. *International Journal of Epidemiology*, 46(6):1958 – 1967, 2017. doi:10.1093/ije/dyx147.

[29] A. S. Blinder. Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, 1973. doi:10.2307/144855.

[30] R. J. Boland. Decision making and sensemaking. In *Handbook on Decision Support Systems 1: Basic Themes*, pages 55–63. Springer Berlin Heidelberg, Berlin,Heidelberg, 2008. doi:10.1007/978-3-540-48713-5_3.

[31] E. Bothos, D. Apostolou, and G. Mentzas. Collective intelligence with web-based information aggregation markets: The role of market facilitation in idea management. *Expert Systems with Applications*, 39(1):1333–1345, 2012. doi:10.1016/j.eswa.2011.08.014.

[32] L. Bottini. The role of neighborhood quality in predicting place attachment: Results from ITA.LI, a newly established nationwide Italian panel survey. *Cities*, 143(104632), 2023. doi:10.1016/j.cities.2023.104632.

[33] D. C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

[34] D. C. Brabham. *Crowdsourcing*. MIT Press, 2013.

[35] P. Braveman. Health disparities and health equity: concepts and measurement. *Annual Review of Public Health*, 27:167–194, 2006. doi:10.1056/NEJMsa1407273.

[36] M. A. Brookhart, T. Stürmer, R. J. Glynn, J. Rassen, and S. Schneeweiss. Confounding control in healthcare database research: Challenges and potential approaches. *Medical Care*, 48(6):114–120, 2010. doi:10.1097/MLR.0b013e3181dbebe3.

[37] E. Budtz-Jørgensen, N. Keiding, P. Grandjean, and P. Weihe. Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. *Annals of Epidemiology*, 17(1):27–35, 2007. doi:10.1016/j.annepidem.2006.05.007.

[38] T. Buecheler, J. H. Sieg, R. M. Füchslin, and R. Pfeifer. Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. In *Proceeding of Alife XII – the 12th International Conference on the Synthesis and Simulation of Living Systems*, pages 679–686. MIT Press, 2010. https://digitalcollection.zhaw.ch/handle/11475/2725.

[39] J. Carifio and R. Perla. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, 3(3):106–116, 2007. doi:10.3844/jssp.2007.106.116.

[40] R. Carnap. On inductive logic. *Philosophy of Science*, 12(2):72–97, 1945. doi:10.1086/286851.

[41] R. Carnap. The two concepts of probability – the problem of probability. *Philosophy and Phenomenological Research*, 5(4):513–532, 1945. doi:10.2307/2102817.

[42] A. D. Chavez-Rivera, Y. Inostroza-Nieves, K. Hemal, and W. Chen. *Handbook for Designing and Conducting Clinical and Translational Surgery*, chapter Longitudinal Study: Design, Measures, and Classic Example, pages 223–226. Academic Press, 2023. doi:10.1016/B978-0-323-90300-4.00074-4.

[43] H. Chen, R. H. L. Chiang, and V. C. Storey. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems*, 36(4):1165 – 1188, 2012. doi:10.2307/41703503.

[44] D. Chicco, M. J. Warrens, and G. Jurman. The Matthews Correlation Coefficient (MCC) is more informative than Cohen's kappa and Brier score in binary classification assessment. *IEEE Access*, 9:78368–78381, 2021. doi:10.1109/ACCESS.2021.3084050.

[45] China Data Lab. *Meteorological Data*. Harvard Dataverse, 2020. doi:10.7910/DVN/TU0JDP.

[46] E. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970. doi:10.1145/362384.362685.

[47] E. Codd, S. Codd, and C. Salley. *Providing OLAP to User-Analysts: An IT Mandate*. E. F. Codd and Associates, 1993.

[48] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

[49] A. Cuzzocrea, I.-Y. Song, and K. C. Davis. Analytics over large-scale multidimensional data: the big data revolution! In *Proceedings of DOLAP'11 – the 14th International ACM Workshop on Data Warehousing and OLAP*, pages 101–104. Association for Computing Machinery, 2011. doi:10.1145/2064676.2064695.

[50] V. Dadvar, L. Golab, and D. Srivastava. Exploring data using patterns: A survey. *Information Systems*, 108(101985), 2022. doi:10.1016/j.is.2022.101985.

[51] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989. doi:10.2307/249008.

[52] A. Dawson and J. Sim. The nature and ethics of natural experiments. *Journal of Medical Ethics*, 41(10):848–853, 2015. doi:10.1136/medethics-2014-102254.

[53] R. Delgado and X.-A. Tibau. Why cohen's kappa should be avoided as performance measure in classification. *PLOS ONE*, 14(9):1–26, 09 2019. doi.org/10.1371/journal.pone.0222916.

[54] Department of the Environment. *National Environment Protection (Ambient Air Quality) Measure*. Federal Register of Legislative Instruments F2016C00215, 2016.

[55] G. A. Diamond, L. Bax, and S. Kaul. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Annals of Internal Medicine*, 147(8):578 – 581, 2007. doi:10.7326/0003-4819-147-8-200710160-00182.

[56] R. Doll and A. B. Hill. Smoking and carcinoma of the lung; preliminary report. *British Medical Journal*, 2(4682):739–748, 1950. doi:10.1136/bmj.2.4682.739.

[57] R. Doll and A. B. Hill. The mortality of doctors in relation to their smoking habits; a preliminary report. *British Medical Journal*, 1(4877):1451–1455, 1954. doi:10.1136/bmj.328.7455.1529.

[58] R. Doll and A. B. Hill. Mortality in relation to smoking: ten years' observations of british doctors. *British Medical Journal*, 1(5396):1460–1467, 1964. doi:10.1136/bmj.1.5396.1460.

[59] R. Doll and R. Peto. Mortality in relation to smoking: 20 years' observations on male british doctors. *British Medical Journal*, 2(6051):1525–1536, 1976. doi:10.1136/bmj.2.6051.1525.

[60] R. Doll, R. Peto, J. Boreham, and I. Sutherland. Mortality in relation to smoking: 50 years' observations on male british doctors. *British Medical Journal*, 328(7455):1519", 2004. doi:10.1136/bmj.38142.554479.AE.

[61] R. Doll, R. Peto, K. Wheatley, R. Gray, and I. Sutherland. Mortality in relation to smoking: 40 years' observations on male british doctors. *British Medical Journal*, 309(6959):901–911, 1994. doi:10.1136/bmj.309.6959.901.

[62] D. Draheim. DEXA'2019 Keynote presentation: Future perspectives of association rule mining based on partial conditionalization (2019). doi:10.13140/RG.2.2.17763.48163.

[63] D. Draheim. *Generalized Jeffrey Conditionalization – A Frequentist Semantics of Partial Conditionalization*. Springer, Berlin, Heidelberg, 2017.

[64] D. Draheim. Collective intelligence systems from an organizational perspective – iiWAS'2019 Keynote. In *Proceedings of iiWAS'2019 – the 21st International Conference on Information Integration and Web-based Applications and Services*, pages 3–4. Association for Computing Machinery, 2019. doi:10.1145/3366030.3368457.

[65] D. Draheim. Future perspectives of association rule mining based on partial conditionalization (DEXA'2019 keynote). In S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA'2019 – the 30th International Conference on Database and Expert Systems Applications*, number 11706 in Lecture Notes in Computer Science, page xvi, Berlin, Heidelberg, 2019. Springer. doi:10.1007/978-3-030-27615-7.

[66] M. Dreseler, J. Kossmann, M. Boissier, S. Klauck, M. Uflacker, and H. Plattner. Hyrise re-engineered: An extensible database system for research in relational in-memory data management. In *Proceedings of EDBT'2019 - the 22nd International Conference on Extending Database Technology*, pages 313–324. OpenProceedings, 2019. doi:10.5441/002/edbt.2019.28.

[67] T. Dunning. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press, 2012.

[68] T. Eden and R. A. Fisher. Studies in crop variation, VI. Experiments on the response of the potato to potash and nitrogen. *The Journal of Agricultural Science*, 19(2):201–213, 1929. doi:10.1017/S0021859600011254.

[69] I. Ekerete, M. Garcia-Constantino, C. Nugent, P. McCullagh, and J. McLaughlin. Data mining and fusion framework for in-home monitoring applications. *Sensors*, 23(21), 2023. doi:10.3390/s23218661.

[70] A. Esmaeilzadeh, L. Golab, and K. Taghva. InfoMoD: Information-theoretic model diagnostics. In *Proceedings of SSDBM'23 – the 35th International Conference on Scientific and Statistical Database Management*. Association for Computing Machinery, 2023. doi:10.1145/3603719.3603725.

[71] A. Esmaelizadeh, J. Rorseth, A. Yu, P. Godfrey, L. Golab, D. Srivastava, J. Szlichta, and K. Taghva. On integrating the data-science and machine-learning pipelines for responsible ai. In *Proceedings of GUIDE-AI '24 – the 1st Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI*, page 50–53. Association for Computing Machinery, 2024. doi:10.1145/3665601.3669849.

[72] European Comission, Directorate F: Social statistics. *Item 4.3 – Adjusted Gender Pay Gap. Doc. DSS/2018/March/4.3*. European Comssion, Eurostat, 2018.

[73] H. Ewald, J. P. Ioannidis, A. Ladanie, K. Mc Cord, H. C. Bucher, and L. G. Hemkens. Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. *Journal of Clinical Epidemiology*, 118:29–41, 2020. doi:10.1016/j.jclinepi.2019.10.012.

[74] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of IJCAI'93 – the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993. https://api.semanticscholar.org/CorpusID:18718011.

[75] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinborough, London, 1925.

[76] J. Fleiss, J. Cohen, and B. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969. doi:10.1037/h0028106.

[77] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions, 3rd edition*. Wiley & Sons, Hoboken, New Jersey, 2003.

[78] N. Fortin, T. Lemieux, and S. Firpo. Decomposition methods in economics. In *Handbook of Labor Economics, Volume 4, Part A*, pages 1–102. Elsevier, 2011. doi:10.1016/S0169-7218(11)00407-2.

[79] S. Ganesh. *International Encyclopedia of Education (Third Edition)*, chapter Multivariate Linear Regression, pages 324–331. Elsevier, Oxford, 2010. doi:10.1016/B978-0-08-044894-7.01350-6.

[80] S. Garcia, J. Luengo, J. A. Sáez, V. Lopez, and F. Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2012. doi:10.1109/TKDE.2012.35.

[81] K. E. Gebaly, P. Agrawal, F. K. Lukasz Golab, and D. Srivastava. Interpretable and informative explanations of outcomes. *Proceedings of the VLDB Endowment*, 8(1):61–72, 2014. doi:10.14778/2735461.2735467.

[82] K. E. Gebaly, G. Feng, L. Golab, F. Korn, and D. Srivastava. Explanation tables. *IEEE Data Engineering Bulletin*, 41(1):43–51, 2018.

[83] J. B. Gelbach. When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2):509–541, 2016. 10.1086/683668.

[84] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9–40, 2006. doi:10.1145/1132960.1132963.

[85] J. M. Genkinger and V. Gebara. Coffee intake and pancreatic cancer risk. In V. R. Preedy, editor, *Coffee in Health and Disease Prevention*, pages 367–374. Academic Press, 2015. doi:10.1016/B978-0-12-409517-5.00040-1.

[86] J. Ghorpade-Aher and B. Sonkamble. A machine learning algorithm for multi-source heterogeneous data with block-wise missing information. *Indian Journal of Computer Science and Engineering*, 13(6):1893 – 1904, 2022. doi:10.21817/indjcse/2022/v13i6/221306103.

[87] A. S. Goldberger. *Econometric Theory*. John Wiley & Sons, 1964.

[88] A. S. Goldberger. *A Course in Econometrics*. Harvard University Press, Cambridge, London, 1991.

[89] M. J. Grant, C. M. Button, and B. Snook. An evaluation of interrater reliability measures on binary tasks using d-prime. *Applied Psychological Measurement*, 41(4):264 – 276, 2017. doi : 10.1177/0146621616684584.

[90] P. Green and V. Srinivasan. Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5:103–123, 1978.

[91] S. Greenland. Modeling and variable selection in epidemiologic analysis. *American Journal of Epidemiology*, 79(3):340–349, 1989. doi:10.2105/ajph.79.3.340.

[92] S. Greenland and H. Morgenstern. Confounding in health research. *Annual Review of Public Health*, 22:189–212, 2001. doi:10.1146/annurev.publhealth.22.1.189.

[93] D. A. Grimes and K. F. Schulz. Bias and causal associations in observational research. *The Lancet*, 359(9302):248–252, 2002. doi:10.1016/S0140-6736(02)07451-2.

[94] G. Grosso, A. Micek, J. Godos, S. Sciacca, A. Pajak, M. A. Martínez-González, E. L. Giovannucci, and F. Galvano. Coffee consumption and risk of all-cause, cardiovascular, and cancer mortality in smokers and non-smokers: a dose-response meta-analysis. *European Journal of Epidemiology*, 31(12):1191–1205, 2016. doi:10.1007/s10654-016-0202-2.

[95] M. Grund, J. Krüger, H. Plattner, A. Zeier, P. Cudre-Mauroux, and S. Madden. Hyrise-a main memory hybrid storage engine. *Proceedings of the VLDB Endowment*, 4(2):105–116, 2010. doi:10.14778/1921071.1921077.

[96] K. A. Guertin, N. D. Freedman, E. Loftfield, B. I. Graubard, N. E. Caporaso, and R. Sinha. Coffee consumption and incidence of lung cancer in the NIH-AARP diet and health study. *International Journal of Epidemiology*, 45(3):929 – 939, 2016. doi:10.1093/ije/dyv104.

[97] K. Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters (Fith Edition), vol. 1 – Analysis of Categorical Ratings*, chapter Agreement Coefficients for Nominal Ratings: A Review. AgreeStat Analytics, 2021.

[98] K. Haerian, D. Varn, S. Vaidya, L. Ena, H. S. Chase, and C. Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology and Therapeutics*, 92(2):228–234, 2012. doi:10.1038/clpt.2012.54.

[99] J. Han, M. Kamber, and J. Pei. Data preprocessing. In J. Han, M. Kamber, and J. Pei, editors, *Data Mining Concepts and Techniques, 3. edition*, pages 83–124. Morgan Kaufmann Publishers, San Francisco, CA, 2012.

[100] N. M. Hanssen, J. Westerink, J. L. Scheijen, Y. van der Graaf, C. D. Stehouwer, C. G. Schalkwijk, and S. S. Group. Higher plasma Methylglyoxal levels are associated with incident cardiovascular disease and mortality in individuals with Type 2 Diabetes. *Diabetes Care*, 41(8):1689–1695, 2018. doi:10.2337/dc18-0159.

[101] J. Hardin and J. Hilbe. *Generalized Estimating Equations*. Chapman and Hall/CRC, London, 2003.

[102] A. B. Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society in Medicine*, 58(5):295–300, 1965. doi:10.1177/003591576505800503.

[103] M. Hojo and W. Senoh. Do the disadvantaged benefit more from small classes? evidence from a large-scale survey in japan. *Japan and the World Economy*, 52, 2019. doi:10.1016/j.japwor.2019.100965.

[104] P. P. Howards. An overview of confounding. Part 2: How to identify it and special situations. *Acta Obstetricia et Gynecologica Scandinavica*, 97(4):400–406, 2018. doi:10.1111/aogs.13293.

[105] D. Hrubá. Epigenetic effects of cigarette smoke in carcinogenesis; [epigenetické úcinky cigaretového koure v procesu karcinogeneze]. *Hygiena*, 58(4):167 – 170, 2013.

[106] W. Hämäläinen and G. I. Webb. Specious rules: An efficient and effective unifying method for removing misleading and uninformative patterns in association rule mining. In *Proceedings of SDM'2017 – the 17th SIAM International Conference on Data Mining*, pages 309–317. SIAM, 2017.

[107] B. A. Israel, E. Eng, A. J. Schulz, and E. A. Parker. *Methods in Community-Based Participatory Research for Health*. Wiley, 2005.

[108] B. A. Israel, A. J. Schulz, E. A. Parker, and A. B. Becker. Review of comminity-based research: Assessing partnership approaches to improve public health. *Annual Review of Public Health*, 19:173–202, 1998. doi:10.1146/annurev.publhealth.19.1.173.

[109] K. Jager, C. Zoccali, A. MacLeod, and F. Dekker. Confounding: What it is and how to deal with it. *Kidney International*, 73(3):256–260, 2008. doi:10.1038/sj.ki.5002650.

[110] E. Jaynes. *Papers on Probability, Statistics and Statistical Physics*. Kluwer Academic Publishers, Dodrecht Boston London, 1989.

[111] B. Jaynthi, P. Parameshwar, M. Tajuddin Baba, K. P. Patil, and G. Cheranjeevi. Effect of insulin sensitizer, rosiglitazone in streptozotocine induced diabetic db/db mice model. *Research Journal of Pharmacy and Technology*, 5(5):619–623, 2012.

[112] R. A. Johnson and D. W. Winchern. *Applied Multivariate Statistics, 6xt edition*. Pearson Prentice Hall, 2007.

[113] F. L. Jones. On decomposing the wage gap: a critical comment on Blinder's method. *The Journal of Human esources*, 18(1):126–130, 1983. doi:10.2307/145660.

[114] M. Jullum and N. L. Hjort. What price semiparametric cox regression? *Lifetime Data Analysis*, 25(3):406 – 438, 2019. doi:10.1007/s10985-018-9450-7.

[115] M. Kalra, N. Lal, and S. Qamar. K-mean clustering algorithm approach for data mining of heterogeneous data. In D. K. Mishra, M. K. Nayak, and A. Joshi, editors, *Proceedings of ICT4SD'2016 – Information and Communication Technology for Sustainable Development, Volume 2*, volume 10 of *Lecture Notes in Networks and Systems*, pages 61–70, Singapore, 2018. Springer Singapore. doi:10.1007/978-981-10-3920-1_7.

[116] M. Kaushik, R. Sharma, S. Arakkal Peious, and D. Draheim. Impact-driven discretization of numerical factors: Case of two- and three-partitioning. In S. N. Srirama, J. C.-W. Lin, R. Bhatnagar, S. Agarwal, and P. K. Reddy, editors, *Proceedings of BDA'21 – the 9th International Conference on Big Data Analytics*, pages 244–260, Cham, 2021. Springer International Publishing. doi:10.1007/978-3-030-93620-4_18.

[117] M. Kaushik, R. Sharma, S. Arakkal Peious, S. Mahtab, S. B. Yahia, and D. Draheim. A systematic assessment of numerical association rule mining methods. *SN Computer Science*, 2(5):1–13, 2021. doi:10.1007/s42979-021-00725-2.

[118] M. Kaushik, R. Sharma, S. Arakkal Peious, M. Shahin, S. B. Yahia, and D. Draheim. On the potential of numerical association rule mining. In T. Dang, J. Küng, M. Takizawa, and T. Chung, editors, *Proceedings of FDSE'20 – the*

7th International Conference on Future Data and Security Engineering, volume 1306 of *Communications in Computer and Information Science*, pages 3–20. Springer, 2020. doi:10.1007/978-981-33-4370-2_1.

[119] M. Kaushik, R. Sharma, M. Shahin, S. Arakkal Peious, and D. Draheim. An analysis of human perception of partitions of numerical factor domains. In M. Indrawan-Santiago, E. Pardede, I. L. Salvadori, M. Steinbauer, I. Khalil, and G. Kotsis, editors, *Proceedings of iiWAS 2022 – the 24th International Conference on Information Integration and Web Intelligence*, volume 13635 of *Lecture Notes in Computer Science*, pages 137–144, Cham, 2022. Springer Nature Switzerland. doi:10.1007/978-3-031-21047-1_13.

[120] M. Kaushik, R. Sharma, A. Vidyarthi, and D. Draheim. Discretizing numerical attributes: An analysis of human perceptions. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *New Trends in Database and Information Systems*, volume 1652 of *Communications in Computer and Information Science*, pages 188–197, Cham, 2022. Springer International Publishing.

[121] A. A. Khan, D. Hussain, K. Ali, G. Khan, M. Ali, and A. Jamil. Time series assessment of the relationship between land surface temperature due to change in elevation: a case study from Hindukush-Himalayan Region (HKH). *Arabian Journal of Geosciences*, 13(13), 2020. doi:10.1007/s12517-020-05530-4.

[122] M. J. Khoury and J. P. A. Ioannidis. Big data meets public health. *Science*, 346(6213):1054–1055, 2014. doi:10.1126/science.aaa2709.

[123] H. A. Kissinger, E. Schmidt, and D. Huttenlocher. *The Age of AI: And Our Human Future*. John Murray, London, 2021.

[124] D. G. Kleinbaum, K. M. Sullivan, and N. D. Barker. *A Pocket Guide to Epidemiology*. Springer, New York, 2006. doi:10.1007/978-0-387-45966-0.

[125] R. Kohavi, D. Tang, and Y. Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020. doi:10.1017/9781108653985.

[126] A. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin Heidelberg, 1933. doi:10.1007/978-3-642-49888-6.

[127] S. Kullback. Letter to the editor: The Kullback-Leibler distance. *The American Statistician*, 41(4):340–341, 1987. doi:10.1080/00031305.1987.10475510.

[128] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi:10.2307/2529310.

[129] T. L. Lash, T. J. VanderWeele, S. Haneuse, and K. J. Rothman. *Modern Epidemiology, 4th edition*. Lippincott Williams and Wilkins, 2021.

[130] P. H. Lee. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *Journal of Epidemiology*, 24(2):161–167, 2014. doi:10.2188/jea.JE20130062.

[131] P. Lévy and R. Bononno. *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Perseus Books, 1997.

[132] C. Li, Y. Peng, X. Zhu, Y. Liu, J. Zou, H. Zhu, X. Li, H. Yi, J. Guan, X. Zhang, H. Xu, , and S. Yin. Independent relationship between sleep apnea-specific hypoxic burden and glucolipid metabolism disorder: a cross-sectional study. *Respiratory Research*, 25(214):1–11, 2024. doi:0.1186/s12931-024-02846-7.

[133] K. T. Li. Frontiers: A simple forward difference-in-differences method. *Marketing Science*, 43(2):267 – 279, 2024. doi:10.1287/mksc.2022.0212.

[134] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 2007.

[135] K.-Y. Lin, P. S. Chen, and C.-F. Lin. Physical function as a predictor of chemotherapy-induced peripheral neuropathy in patients with pancreatic cancer. *BMC Gastroenterology*, 24(154):1–8, 2024. doi:10.1186/s12876-024-03227-6.

[136] Y. Liu, T. Lin, J. Zhang, F. Wang, Y. Huang, X. Wu, H. Ye, G. Zhang, X. Cao, and G. de Leeuw. Opposite effects of aerosols and meteorological parameters on warm clouds in two contrasting regions over eastern China. *Atmospheric Chemistry and Physics*, 24:4651–4673, 2024. doi:10.5194/acp-24-4651-2024.

[137] D. Madden. Towards a broader explanation of male-female wage differences. *Applied Economics Letters*, 7(12):765–770, 2000.

[138] G. Maldonado and S. Greenland. Simulation study of confounder-selection strategies. *American Journal of Epidemiology*, 138(11):923–963, 2014. doi:10.1093/oxfordjournals.aje.a116813.

[139] T. W. Malone and M. S. Bernstein. *Handbook of Collective Intelligence*. MIT Press, 2015.

[140] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta – Protein Structure*, 405(2):442–451, 1975. doi:10.1016/0005-2795(75)90109-9.

[141] A. Maydeu-Olivares, D. Shi, and A. J. Fairchild. Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, 25(2):243 – 258, 2020. doi:10.1037/met0000226.

[142] M. L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012. doi:10.11613/bm.2012.031.

[143] R. McNamee. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine*, 62:500–506, 2005. doi:10.1136/oem.2002.001115.

[144] M. F. Mendoza, R. M. Sulague, T. Posas-Mendoza, and C. J. Lavie. Impact of coffee consumption on cardiovascular health. *Ochsner Journal*, 23(2):152–158, 2023. doi:10.31486/toj.22.0073.

[145] R. M. Mickey and S. Greenland. The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, 129(1):125–37, 1989. doi:10.1093/oxfordjournals.aje.a115101.

[146] R. M. Mickey and S. Greenland. The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, 129(1):125–37, 1989. doi:10.1093/oxfordjournals.aje.a115101.

[147] G. A. Miller and J. P. Chapman. Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110(1):40–48, 2001. doi:10.1037//0021-843x.110.1.40.

[148] R. E. Miller. *Optimization: Foundations and Applications*. John Wiley & Sons, New York, Hoboken, 2012.

[149] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis, 5th edition*. John Wiley & Sons, 2012. doi:10.1111/biom.12129.

[150] A. Morabia. History of the modern epidemiological concept of confounding. *Journal of Epidemiology and Community Health*, 65(4):297–300, 2011. doi:10.1136/jech.2010.112565.

[151] N. Morar, C. Baber, F. McCabe, S. D. Starke, I. Skarbovsky, A. Artikis, and I. Correai. Drilling into dashboards: Responding to computer recommendation in fraud analysis. *IEEE Transactions on Human-Machine Systems*, 49(6):633–641, 2019. doi:10.1109/THMS.2019.2925619.

[152] K. Moreland and K. Truemper. Discretization of target attributes for subgroup discovery. In P. Perner, editor, *Proceedings of MLDM'2009 – the 15th International Workshop on Machine Learning and Data Mining in Pattern Recognition*, volume 5632 of *Lecture Notes in Artificial Intelligence*, pages 44–52. Springer, 2009. doi:978-3-642-03070-3_4.

[153] K.-W. Nam, H.-M. Kwon, H.-Y. Jeong, J.-H. Park, and K. Min. Blood urea nitrogen to albumin ratio is associated with cerebral small vessel diseases. *Scientific Reports*, 14(1), 2024. doi:10.1038/s41598-024-54919-8.

[154] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, 236(767):333–380, 1937. doi:10.1098/rsta.1937.0005.

[155] J. Neyman. Frequentist probability and frequentist statistics. *Synthese*, 36:97–131, 1977. 10.1007/BF00485695.

[156] R. Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973. doi:10.2307/2525981.

[157] R. L. Oaxaca and M. R. Ransom. Identification in detailed wage decompositions. *Review of Economics and Statistics*, 81(1):154–157, 1999. doi:10.1162/003465399767923908.

[158] S. Park, H.-L. Kim, K.-T. Park, H. S. Joh, W.-H. Lim, J.-B. Seo, S.-H. Kim, and M.-A. Kim. Association between arterial stiffness and autonomic dysfunction in participants underwent treadmill exercise testing: a cross-sectional analysis. *Scientific Reports*, 14(1), 2024. doi:10.1038/s41598-024-53681-1.

[159] J. Pearl. *Probabilistic Reasoning in Intelligent Systems – Networks of Plausible Inference, 2nd edition*. Morgan Kaufmann, San Francisco, 1988.

[160] J. Pearl. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Science Institute*, pages 391–401, 1993.

[161] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. doi:10.1093/biomet/82.4.694.

[162] J. Pearl. Causal inference in statistics – an overview. *Statistics Surveys*, 3:96–146, 2009.

[163] J. Pearl. *Causality – Models, Reasoning, and Inference, 2nd edition*. Cambridge University Press, 2009.

[164] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[165] J. Pearl and D. McKenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

[166] K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi:10.1098/rspl.1895.0041.

[167] G. Peevely, L. Hedges, and B. A. Nye. The relationship of class size effects and teacher salary. *Journal of Education Finance*, 31(1):101 – 109, 2005. http://www.jstor.org/stable/40704252.

[168] K. Pforr. Femlogit—implementation of the multinomial logit model with fixed effects. *The Stata Journal*, 14(4):847–862, 2014. doi:10.1177/1536867X1401400409.

[169] H. Plattner. The impact of columnar in-memory databases on enterprise systems. *Proceedings of the VLDB Endowment*, 7(13):1722–1729, 2014. doi:10.14778/2733004.2733074.

[170] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi. How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from Bed to Bench*, 5(2):79–83, 2012. doi:10.22037/ghfbb.v5i2.246.

[171] K. J. Preacher, P. J. Curran, and D. J. Bauer. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31(4):437–448, 2006. doi:10.3102/10769986031004437.

[172] F. Quin, D. Weyns, M. Galster, and C. C. Silva. A/B testing: A systematic literature review. *Journal of Systems and Software*, 211, 2024. doi:10.1016/j.jss.2024.112011.

[173] A. Rajasekar and V. Kumar. Randomized controlled trials: Gold standard of evidence. *Drug Invention Today*, 12(6):1215 – 1217, 2019.

[174] H. Reichenbach. *The Direction of Time*. Dover Books on Physics. Dover Publications, 2012.

[175] A. E. Rhodes, E. Lin, and D. L. Streiner. Confronting the confounders: The meaning, detection, and treatment of confounders in research. *The Canadian Journal of Psychiatry*, 44(2):175–179, 1999. doi:10.1177/070674379904400209.

[176] P. A. Riach and J. Rich. Field experiments of discrimination in the market place. *The Economic Journal*, 112(483):F480–F518, 2002. doi.org:10.1111/1468-0297.00080.

[177] J. M. Robins and S. Greenland. Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, 95(450):431–435, 2000. doi:10.2307/2669381.

[178] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. doi:10.1093/biomet/70.1.41.

[179] M. R. Rosenzweig and J. Morgan. Wage discrimination: A comment. *The Journal of Human Resources*, 11(1):3–7, 1976. doi:10.2307/145069.

[180] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology, 3rd edition*. Lippincott Williams & Wilkins, Philadelphia, 2008.

[181] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons, 2002. doi:10.1002/0470870958.

[182] D. A. Savitz and G. A. Wellenius. *Interpreting Epidemiologic Evidence – Connecting Research to Applications, 2nd ed.*, chapter Confounding I – Theoretical Considerations. Oxford University Press, 2016. doi:10.1093/acprof:oso/9780190243777.001.0001.

[183] D. A. Savitz and G. A. Wellenius. *Interpreting Epidemiologic Evidence – Connecting Research to Applications, 2nd ed.*, chapter Confounding II – Practical Considerations. Oxford University Press, 2016. doi:10.1093/acprof:oso/9780190243777.001.0001.

[184] A. J. Schulz, B. A. Israel, A. G. Reyes, D. Wilkins, and S. Batterman. *Promoting health equity with community-based participatory research: The community action to promote healthy environments (CAPHE) partnership*, volume 3. Springer International Publishing, 2023. doi:10.1007/978-3-031-20401-2_20.

[185] N. A. Schuster, J. W. Twisk, G. Ter Riet, M. W. Heymans, and J. J. Rijnhart. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC Medical Research Methodology*, 21:1–9, 2021. doi:10.1186/s12874-021-01316-8.

[186] S. Shabaninejad, H. Khosravi, M. Indulska, A. Bakharia, and P. Isaias. Automated insightful drill-down recommendations for learning analytics dashboards. In *Proceedings of LAK'20 – the 10th International Conference on Learning Analytics and Knowledge*, pages 41–46, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3375462.3375539.

[187] S. A. Shah, D. Z. Seker, S. Hameed, and D. Draheim. The rising role of big data analytics and IoT in disaster management: Recent advances, taxonomy and prospects. *IEEE Access*, 7:54595–54614, 2019. doi:10.1109/ACCESS.2019.2913340.

[188] M. Shahin, S. Arakkal Peious, R. Sharma, M. Kaushik, S. Ben Yahia, S. A. Shah, and D. Draheim. Big data analytics in association rule mining: A systematic literature review. In *Proceedings of BDET'23 – the 3rd International Conference on Big Data Engineering and Technology*, pages 40–49. Association for Computing Machinery, 2021. doi:10.1145/3474944.347495.

[189] M. Shahin, W. Inoubli, S. A. Shah, S. B. Yahia, and D. Draheim. Distributed scalable association rule mining over COVID-19 data. In T. Dang, J. Küng, T. Chung, and M. Takizawa, editors, *Proceedings of FDSE'2021 – the 8th International Conference on Future Data and Security Engineering*, volume 1307 of *Lecture Notes in Computer Science*, pages 39–52. Springer, 2021. 10.1007/978-3-030-91387-8_3.

[190] M. Shahin, S. Saeidi, S. A. Shah, M. Kaushik, R. Sharma, S. Arakkal Peious, and D. Draheim. Cluster-based association rule mining for an intersection accident dataset. In *Proceedings ICE Cube 2021 – the 1st International Conference on Computing, Electronic and Electrical Engineering*, pages 1–6. IEEE, 2021. doi:10.1109/ICECube53880.2021.9628206.

[191] M. Shahin, S. A. Shah, R. Sharma, T. Ghasempouri, J. A. Poveda, T. Fahringer, and D. Draheim. Performance of a distributed apriori algorithm using the serverless functions of the apollo framework. In R. Silhavy and P. Silhavy, editors, *Proceedings of CSOC'24 – 13th Computer Science Online Conference, vol. 4 (Machine Learning Methods in Systems)*, volume 1126 of *Lecture Notes in Network and Systems*, pages 363–374. Springer, 2024. doi:10.1007/978-3-031-70595-3_37.

[192] R. Sharma, H. Garayev, M. Kaushik, S. Arakkal Peious, P. Tiwari, and D. Draheim. Detecting Simpson's paradox: A machine learning perspective. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA 2022 – the 33rd International Conference on Database and Expert Systems Applications*, volume 13426 of *Lecture Notes in Computer Science*, pages 323–335, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-12423-5_25.

[193] R. Sharma, M. Kaushik, S. Arakkal Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, and D. Draheim. A novel framework for unification of association

rule mining, online analytical processing and statistical reasoning. *IEEE Access*, 10:12792–12813, 2022. doi:10.1109/ACCESS.2022.3142537.

[194] R. Sharma, M. Kaushik, S. Arakkal Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *Proceedings of ADBIS 2022 – the 26th International Conference on New Trends in Database and Information Systems*, volume 1652 of *Communications in Computer and Informations Science*, pages 67–76, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-15743-1_7.

[195] R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. Vidyarthi, and D. Draheim. Existence of the Yule-Simpson effect: An experiment with continuous data. In *Proceedings of Confluence 2022 – the 12th International Conference on Cloud Computing, Data Science and Engineering*, pages 351–355. IEEE, 2022. doi:10.1109/Confluence52989.2022.9734211.

[196] R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim. Why not to trust big data: Discussing statistical paradoxes. In U. K. Rage, V. Goyal, and P. K. Reddy, editors, *Proceedings of DASFAA 2022 International Workshops – the 27th International Conference on Database Systems for Advanced Applications*, volume 13248 of *Lecture Notes in Computer Science*, pages 50–63, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-11217-1_4.

[197] R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. S. Yadav, and D. Draheim. Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, and R. Uday Kiran, editors, *Proceedings of DASFAA 2022 – the 27th International Conference on Database Systems for Advanced Applications*, volume 13245 of *Lecture Notes in Computer Science*, pages 596–603, Cham, 2022. Springer International Publishing. 10.1007/978-3-031-00123-9_48.

[198] R. Sharma, M. Kaushik, S. Arakkal Peious, S. B. Yahia, and D. Draheim. Expected vs. unexpected: Selecting right measures of interestingness. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK 2020 – the 22nd International Conference on Big Data Analytics and Knowledge Discovery*, volume 13428 of *Lecture Notes in Computer Science*, pages 38–47, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-59065-9_4.

[199] T. Shen. Small-class effects on science achievement in secondary education. *Studies in Educational Evaluation*, 82, 2024. doi:10.1016/j.stueduc.2024.101368.

[200] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. doi:10.1111/j.2517-6161.1951.tb00088.x.

[201] F. A. Sloan, P. Ayyagari, M. Salm, and D. Grossman. The longevity gap between Black and White men in the United States at the beginning and end

of the 20th century. *American Journal of Public Health*, 100(2):357–363, 2009. doi:10.2105/AJPH.2008.158188.

[202] J. M. Snowden, S. Rose, and K. M. Mortimer. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738, 2011. doi:10.1093/aje/kwq472.

[203] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi:10.2307/1412159.

[204] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. *ACM SIGMOD Record*, 25(2):1–12, 1996. doi:10.1145/235968.233311.

[205] D. Stanley, Z. Marshall, L. Lazarus, S. Leblanc, T. Heighton, B. Preater, and M. Tyndall. Harnessing the power of community-based participatory research: Examining knowledge, action, and consciousness in the PROUD study. *Social Work in Public Health*, 30(3):312 – 323, 2015. doi:10.1080/19371918.2014.1001935.

[206] M. Stead, A. MacKintosh, A. Findlay, L. Sparks, A. Anderson, K. Barton, and D. Eadie. Impact of a targeted direct marketing price promotion intervention (Buywell) on food-purchasing behaviour by low income consumers: a randomised controlled trial. *Journal of Human Nutrition and Dietetics*, 30(4):524 – 533, 2017. doi:10.1111/jhn.12441.

[207] C. M. Stein, N. J. Morris, N. B. Hall, and N. L. Nock. Structural equation modeling. *Methods in Molecular Biology*, 1666:557 – 580, 2017. doi:10.1007/978-1-4939-7274-6_28.

[208] J. H. Stock. *Instrumental Variables in Statistics and Econometrics*. Elsevier Inc., 2015. doi:10.1016/B978-0-08-097086-8.42037-4.

[209] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 12 2018. doi:10.1111/j.2517-6161.1974.tb00994.x.

[210] D. Suarez, R. Borràs, and X. Basagaña. Differences between marginal structural models and conventional models in their exposure effect estimates: a systematic review. *Epidemiology*, 22(4):586–588, 2011. doi:10.1097/EDE.0b013e31821d0507.

[211] Y. Suk and H. Kang. Tuning random forests for causal inference under cluster-level unmeasured confounding. *Multivariate Behavioral Research*, 58(2):408 – 440, 2023. doi:10.1080/00273171.2021.1994364.

[212] S. Suran, V. Pattanaik, and D. Draheim. Frameworks for collective intelligence: A systematic literature review. *ACM Computing Surveys*, 52(1):1–36, 2020. doi:10.1145/3368986.

[213] E. J. Tchetgen Tchetgen, C. Park, and D. B. Richardson. Universal difference-in-differences for causal inference in epidemiology. *Epidemiology*, 35(1):16 – 22, 2024. doi:10.1097/EDE.0000000000001676.

[214] T. Teichert. Confounding of effects in rank-based conjoint-analysis. Technical Report 409, Institute for Research in Innovation Management, Christian-Albrechts-University zu Kiel, Kiel, Germany, September 1996.

[215] T. Teichert. *Conjoint Measurement: Methods and Applications*, chapter Confounding of Effects in Rank-Based Conjoint-Analysis, pages 225–250. Springer, Berlin, Heidelberg, 2001. doi:10.1007/978-3-662-06392-7_10.

[216] The Economist. Instant economics. The Economist October 23rd-29th, page 13, 2021.

[217] The Economist. The real-time revolution. The Economist October 23rd-29th, pages 22–24, 2021.

[218] J. Thomas and L. Sael. Overview of integrative analysis methods for heterogeneous data. In *Proceedings of BIGCOMP 2015 – the 2nd International Conference on Big Data and Smart Computing*, pages 266–270. IEEE, 2015. doi:10.1109/35021BIGCOMP.2015.7072811.

[219] T. L. Towner. Class size and academic achievement in introductory political science courses. *Journal of Political Science Education*, 12(4):420 – 436, 2016. doi:10.1080/15512169.2016.1154470.

[220] J. P. Vandenbroucke, A. Broadbent, and N. Pearce. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45(6):1776–1786, 2016. doi:10.1093/ije/dyv341.

[221] H. Wainer. Eelworms, bullet holes, and geraldine ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics*, 14(2):121–140, 1989.

[222] J. Wang, J. He, W. Xu, R. Li, and W. Chu. Learning to discover various Simpson's paradoxes. In *Proceedings of KDD'23 – the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5092–5103, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3580305.3599859.

[223] X. Wang, G. Hripcsak, M. Markatou, and C. Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337, 2009. doi:10.1197/jamia.M3028.

[224] J. J. Ware, J.-A. Tanner, A. E. Taylor, Z. Bin, P. Haycock, J. Bowden, P. J. Rogers, G. Davey Smith, R. F. Tyndale, and M. R. Munafò. Does coffee consumption impact on heaviness of smoking? *Addiction*, 112(10):1842 – 1853, 2017. doi:10.1111/add.13888.

[225] R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi:10.1080/00031305.2016.1154108.

[226] N. S. Weiss. *Clinical Epidemiology: The Study of the Outcome of Illness*, volume 36 of *Monographs in Epidemiology and Biostatistics*. Oxford University Press, 2006. doi:10.1093/oso/9780195305234.001.0001.

[227] J. M. Wooldridge. *Introductory Econometrics: A Modern Approach, 6th edition*. Cengage Learning, 2003.

[228] E. L. Wynder and E. A. Graham. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. *Journal of the American Medical Association*, 143(4):329–336, 1950. doi:10.1001/jama.1950.02910390001001.

[229] C. Xu, S. M. Brown, and C. Grant. Detecting simpson's paradox. In K. Brawner and V. Rus, editors, *Proceedings of FLAIRS 2018 – the 31st International Florida Artificial Intelligence Research Society Conference*, pages 221–224. AAAI Press, 2018. https://cdn.aaai.org/ocs/17641/17641-77733-1-PB.pdf.

[230] B. Yan, G. Mai, Y. Hu, and K. Janowicz. Harnessing heterogeneous big geospatial data. In M. Werner and Y.-Y. Chiang, editors, *Handbook of Big Geospatial Data*, pages 459–473. Springer, Berlin, Heidelberg, 2021. doi:10.1007/978-3-030-55462-0_17.

[231] F. Yates and K. Mather. Ronald Aylmer Fisher 1890–1962. *Biographical Memoirs of Fellows of the Royal Society*, 9:91–129, 1963.

[232] D. L. Yu and S. W. Buchanan. *An Integrated Approach to Environmental Management*, chapter Geographic Information Systems in Environmental Management, pages 423–439. Wiley, 2016. doi:10.1002/9781118744406.ch18.

[233] G. U. Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903. doi:10.1093/biomet/2.2.121.

[234] G. U. Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912. doi:10.2307/2340126.

[235] J. Zhang, P. Wang, Q. Pang, S. Wang, and A. Zhang. Handgrip strength is associated with cognitive function in older patients with stage 3-5 chronic kidney disease: results from the NHANES. *Scientific Reports*, 14(10329):1–8, 2024. doi:10.1038/s41598-024-60869-y.

[236] Q. Zhao, E. Adeli, and K. M. Pohl. Training confounder-free deep learning models for medical applications. *Nature Communications*, 11(1), 2020. doi:10.1038/s41467-020-19784-9.

[237] Y. Zhong, S. G. Carmella, P. Upadhyaya, J. B. Hochalter, D. Rauch, A. Oliver, J. Jensen, D. Hatsukami, J. Wang, C. Zimmerman, and S. S. Hecht. Immediate consequences of cigarette smoking: Rapid formation of polycyclic aromatic hydrocarbon diol epoxides. *Chemical Research in Toxicology*, 24(2):246 – 252, 2011. doi:10.1021/tx100345x.

Index

Technical Aids

Grammarly, Writefull and ChatGPT has been used to improve the grammar and sentences.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dirk Draheim, for his unwavering support and patience throughout my PhD journey. His academic excellence, profound knowledge, and strong ethical principles have been a constant source of guidance and inspiration in my research. I deeply appreciate the countless hours he spent sharing invaluable insights and simplifying complex concepts for me. In particular, I am grateful to Dirk Draheim for suggesting the C-IA measure.

Special thanks to Ahto Buldas for the valuable discussions on the theory of linear regression.

Special thanks to Shweta Suran and Vishwajeet Pattanaik for the deep discussions about the collective intelligence systems and reputation models.

I would also like to extend my gratitude to my colleagues Rahul Sharma, Minakskshi Kaushik, Mahtab Shahin, Sidra Butt, Silvia Lips, Richard III Dreyling and the staff at Tallinn University of Technology for their constant support and assistance throughout my PhD journey.

My deepest gratitude goes to my parents, who have unconditionally believed in me from the very beginning. I feel truly fortunate to share this personal achievement with them, whose love and support I have always felt, despite the physical distance between us.

A special and warm thank-you goes to my dear partner, friend, and soulmate, who has stood by me through all my highs and lows. It is one of the greatest comforts to know that someone believes in you, is confident you will succeed, and is ready to share in both your joyous and challenging moments.

Lastly, I want to thank my friends who have stood by my side. In fact, I am grateful to anyone who has ever offered a kind word of encouragement. Each supportive gesture, no matter how small, has contributed to my determination, perseverance, patience, and self-confidence.

This work has been conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

This work was co-funded by the European Union and Estonian Research Council via project TEM-TA141.

Abstract

Measures of Impact and Confounding – An Analysis and Experimental Comparison of Novel and Established Measures

Confounding is pervasive in data, both in experimental and observational studies, and is a severe risk for the accuracy of results and conclusions of scientific studies. Based on the research findings with our data mining tool GrandReport, including behaviour of Pearson correlation during drill-down and linear-regression-based confounder adjustment, we have identified the need for more systematic understanding of confounding measures and confounding patterns. Therefore, in this thesis, we systematically compare four methods of confounding adjustments using the same datasets on a large scale.

In this thesis, we contribute as follows. First, we introduce a novel measure for the impact of categorical variables in their entirety, called Coupled Impact Assessment (C-IA). Furthermore, we introduce a novel method to detect confounders utilizing the C-IA measure. Then, we conduct combinatorially designed experiments with 694 datasets from the Harvard Dataverse and the NZ Government Repository to investigate three well-established approaches for detecting confounders, i.e., the Ad-Hoc method, Oaxaca-Blinder decomposition, and the linear-regression-based method, together with our own novel C-IA-based method.

Based on our experiment results, we discover that the four investigated methods for detecting confounders do not show any relevant agreement or disagreement beyond chance (in terms of both Cohen's κ and Yule's ϕ). This is surprising, as all of the four methods have been specifically designed for exactly the same target: to detect confounders, and, actually, three of the methods have been in widespread use over the decades in a plethora of scientific studies for detecting confounders. Additionally, we argue that the finding is highly relevant for the working data scientist.

Furthermore, based on our experiment results, we identify four interesting patterns of confounding effects during drill-down into potential confounders, that we showcase in eight data case studies.

Also, we elaborate a systematic interpretation of the linear regression model utilizing so-called multiplicative edges diagrams. We utilize this interpretation to reflect on linear-regression-based confounding, including a critical discussion of cutoff rules for confounding adjustments.

Although confounding effects are ubiquitous in data and scientific studies, confounding is rather neglected in the common data mining tools, be it from Association rule mining (ARM) or Online analytical processing (OLAP). Here is, where this thesis aims at envisioning a paradigm shift, i.e., to design an integrate systematic support for treatment of confounding in data mining tools.

Kokkuvõte

Mõju ja segajate mõõdikud – uute ja väljakujunenud meetmete analüüs ja eksperimentaalne võrdlus

Segasus on levinud nii eksperimentaalsetes kui ka vaatlusuuringutes ning kujutab endast tõsist ohtu teaduslike uuringute tulemuste ja järelduste täpsusele. Meie andmekaevetööriista GrandReport abil tehtud uurimistulemuste põhjal, sealhulgas Pearsoni korrelatsiooni käitumise analüüsil andmete süvitsiminekul ning lineaarregressioonipõhisel segajate korrigeerimisel, oleme tuvastanud vajaduse süsteemsema arusaama järele segasuse mõõtmismeetoditest ja segasuse mustritest. Seetõttu võrdleme selles doktoritöös süsteemselt nelja erinevat segasuse korrigeerimise meetodit, kasutades samu andmestikke suures mahus.

Selle doktoritöö peamised panused on järgmised. Esiteks, tutvustame uudset kategooriliste muutujate mõju hindamise mõõdikut, mida nimetame Coupled Impact Assessment (C-IA) ehk sidusmõju hindamiseks. Lisaks esitleme uut meetodit segajate tuvastamiseks, mis põhineb C-IA mõõdikul. Seejärel viime läbi kombineeritud eksperimendid 694 andmestikuga, mis on pärit Harvard Dataverse ja Uus-Meremaa valitsuse andmehoidlast, et uurida kolme laialdaselt kasutatud segajate tuvastamise meetodit – Ad-Hoc meetodit, Oaxaca-Blinderi dekompositsiooni ja lineaarregressioonil põhinevat meetodit – koos meie enda uudse C-IA meetodil põhineva lähenemisega.

Eksperimentaalsete tulemuste põhjal avastasime, et uuritud neljal segajate tuvastamise meetodil ei esine olulist kokkulangevust ega lahknevust juhuslikkuse tasemest (hinnatud nii Coheni κ kui ka Yule'i ϕ abil). See on üllatav, kuna kõik neli meetodit on loodud täpselt sama eesmärgi jaoks – segajate tuvastamiseks – ning kolm neist on olnud laialdaselt kasutusel aastakümneid paljudes teadusuuringutes. Samuti väidame, et see tulemus on äärmiselt oluline andmeteadlaste jaoks.

Lisaks tuvastasime oma eksperimentaalsete tulemuste põhjal neli huvitavat segasuse mustrit, mis ilmnevad võimalike segajate analüüsil andmete süvitsiminekul. Neid mustreid illustreerime kaheksas juhtumiuuringus.

Samuti töötame välja süsteemse tõlgenduse lineaarse regressioonimudeli kohta, kasutades niinimetatud multiplikatiivsete servade diagramme. Kasutame seda tõlgendust, et analüüsida lineaarregressioonil põhinevat segasust ning kriitiliselt arutleda segasuse korrigeerimise läviväärtuste reeglite üle.

Kuigi segasuse mõju on andmetes ja teadusuuringutes kõikjal esinev, on seda andmekaevetööriistades, nagu assotsiatsioonireeglite kaevandamine (ARM) või veebipõhine analüütiline töötlemine (OLAP), seni suuresti eiratud. Käesoleva doktoritöö eesmärk on algatada paradigmanihe kavandada ja integreerida andmekaevetööriistadesse süsteemne tugi segasuse käsitlemiseks.

Appendix I

I

S. Arakkal Peious, M. Kaushik, M. Shahin, R. Sharma, and D. Draheim. On choosing the columnar in-memory database Hyrise as high-performant implementation platform for the GrandReport tool. In *Proceedings of NGISE'2025 – the 1st International Conference in Next Genenation Information Systems Engineering*, pages 1–7. IEEE, 2025. to appear, preprint available at IEEE TechRxiv, doi: 10.36227/techrxiv.174015861.15410981/v1

On Choosing the Columnar In-Memory Database Hyrise as High-Performant Implementation Platform for the GrandReport Tool

1st Sijo Arakkal Peious
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
sijo.arakkal@taltech.ee

2nd Minakshi Kaushik
Dependability of Software-Intensive Systems Group
Karlsruhe Institute of Technology
Karlsruhe, Germany
minakshi.kaushik@kit.edu

3rd Mahtab Shahin
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
mahtab.shahin@taltech.ee

4th Rahul Sharma
Department of Information Technology
AKGEC College
Ghaziabad, India
sharmarahul@akgec.ac.in

5th Dirk Draheim
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
dirk.draheim@taltech.ee

***Abstract*—In this paper, we provide informed arguments for using columnar in-memory database technology, in particular the Hyrise database, for the high-performant implementation of the highly combinatorial data mining tool GrandReport. In service of that we provide a targeted review of columnar databases, a targeted review of columnar in-memory databases, and a comparison of nine established columnar in-memory databases. On the basis of this, we discuss advantages of using the Hyrise database as the future implementation platform for the GrandReport tool.**

***Index Terms*—Online Analytical Processing, OLAP, data mining, numerical association rule mining, grand reports, GrandReport, columnar databases, in-memory databases, Hyrise**

I. Introduction

Our tool *GrandReport*[1,2] [1] is a prototypical implementation to evaluate the usefulness [2] of grand reports [3], [4], and linear-regression-based detection of confounders [5]. However, the *GrandReport* tool is a prototypical implementation and no particular effort has been invested into its performance. Yet, in today's instant economics [6], [7], decision support systems are expected to answer as quick as possible, ideally, in real-time. Therefore, it would be interesting to optimize the *GrandReport* tool for high performance, in service to evaluate its usefulness [2] with larger datasets and enabling deeper drill-ins in short response times.

In this paper, we provide informed arguments for re-implementing the GrandReport tool on the basis of an underlying columnar in-memory database [8], as massive real-time analytical reporting is exactly what columnar in-memory databases are designed for. Given its design as columnar in-memory database [9] , its extensibility [10] and its availability[3] for research, we identify the innovative Hyrise[4] technology as particularly promising platform for these endeavors.

We proceed as follows. In Section II, we provide a targeted review of columnar databases. In Section III, we provide a targeted review of columnar in-memory databases and a comparison of nine established columnar in-memory databases, including SAP HANA, ClickHouse, Google BigQuery, Amazon Redshift, MonetDB, kdb+, Apache Arrow, HyPer, and Hyrise. In Section IV, we provide a targeted review of the Hyrise database technology. In section IV, we provide a brief overview of the GrandReport tool. In Section V, we discuss advantages of using the Hyrise as implementation platform for GrandReport. We finish the paper with a conclusion in Section VI.

II. Columnar Databases

Databases play a crucial role in storing, organizing, and managing data efficiently across various applications. A database is a structured collection of data stored in a computer system, enabling efficient storage, retrieval, and management [11]–[14]. The advent of database management systems (DBMS) revolutionized data handling, enhancing performance, scalability, and security. Today, databases remain fundamental to modern computing, powering a wide range of digital applications.

Columnar databases [15], also known as column-oriented database management systems (DBMSs) [15], are designed to store data in a column-oriented format rather than in

[1] http://grandreport.me
[2] https://github.com/istaltech/grandreport
[3] https://github.com/hyrise/hyrise/wiki
[4] https://hpi.de/plattner/projects/hyrise.html

the traditional row-based structure [16]. This architectural shift offers significant advantages for analytical and read-heavy workloads, making them particularly well-suited for large-scale data processing and business intelligence applications [17]. By organizing data column by column instead of row by row, these databases allow for efficient data compression [18], optimized storage utilization [19], and improved query performance [20]. Since analytical queries often require accessing only a subset of columns, column-based databases reduce I/O overhead, leading to faster processing speeds and better resource efficiency [21].

The primary motivation behind columnar databases is to accelerate analytical queries by selectively retrieving only the necessary columns. This selective retrieval reduces data transfer and computation time, making them highly efficient for data warehousing, business analytics, and real-time reporting. The ability to compress data more effectively than row-based databases also contributes to lower storage costs and faster query execution. Compression techniques such as run-length encoding, dictionary encoding, and delta encoding further enhance performance, allowing for high-speed data retrieval and improved analytical capabilities.

Modern columnar databases, including Google BigQuery, Amazon Redshift, and ClickHouse, have gained widespread adoption due to their ability to handle massive datasets efficiently. These databases support complex analytical queries, making them essential for enterprises that require scalable and high-performance solutions. Additionally, columnar storage formats such as Apache Parquet and Apache ORC enhance processing efficiency through optimized encoding and vectorized query execution [22]. The continued evolution of columnar storage techniques and hybrid database models indicates a promising future for this technology in the realm of big data and analytics [23], further revolutionizing data-driven decision-making.

A. Advantages of Columnar Databases

a) Improved Query Performance: The retrieval of specific (fewer) columns of data is faster in columnar databases [64]. Their structure is particularly advantageous for analytical queries that involve aggregations and computations on large datasets across a limited number of attributes. Since only the necessary columns are retrieved, query performance improves significantly, reducing I/O operations [65], [66]. Additionally, columnar databases leverage parallel processing by distributing workloads across multiple processors, which further accelerates data retrieval and query execution [67]. This faster retrieval make columnar databases highly effective for analytical workloads, enabling faster insights and improved performance in data-intensive environments.

b) Enhanced Compression: The data can be efficiently compressed because the column formats exhibit a high degree of homogeneity, meaning they follow consistent structures and patterns. This homogeneity allows advanced compression algorithms to achieve significant data reduction without loss of information [68], [69]. Effective compression minimizes storage requirements [28], [70], making data management more cost-effective and scalable. Additionally, reduced data size enhances query execution speed, as smaller datasets require less processing power and memory, leading to faster retrieval and improved overall system performance.

c) Scalability: Columnar databases offer scalability by enabling horizontal expansion, where additional nodes can be added to a cluster to accommodate growing datasets efficiently [66], [71]. This distributed architecture enhances performance, allowing faster query execution and improved fault tolerance. Columnar databases also support cross-platform data sharing and are optimized for modern analytical requirements [22]. Unlike traditional row-based databases, columnar storage is well-suited for denormalized schemas, reducing the need for complex joins and optimizing read-heavy workloads [66], [72].

III. Columnar In-Memory Databases

Dictionary-encoded column-oriented in-memory databases [8], [73] (henceforth called columnar in-memory databases [8] for short) have transformed data processing by leveraging the advantages of both columnar storage and in-memory computing, resulting in exceptional performance for analytical workloads [20]. By storing data in a columnar format, they optimize compression and retrieval, significantly enhancing query performance and real-time analytics [74]. The in-memory approach eliminates disk I/O bottlenecks, enabling faster data access and efficient data management. As organizations increasingly rely on data-driven decision-making, the adoption of these databases has become essential for modern analytics and business intelligence applications. Their ability to handle large-scale data with speed and efficiency makes them a vital component in contemporary data ecosystems.

Understanding columnar in-memory databases requires a deep dive into two fundamental concepts: in-memory computing and columnar databases. In-memory computing refers to the technique of storing and processing data directly in main memory, i.e., in random-access memory (RAM). instead of relying on traditional disk-based storage. This approach dramatically reduces latency, as data retrieval and processing occur at much faster speeds as compared to disk I/O operations. A columnar database, on the other hand, structures data in a way that stores each column separately, rather than following the conventional row-based storage model, see Section II. When these two concepts are combined, columnar in-memory databases emerge as a powerful solution for handling large-scale data processing efficiently. A key to the performance is dictionary encoding as follows. In case of assuming a row-based in-memory database, column scans could be speed-up significantly via striding (picking only the scanned column cell from each row) [20]. However, stride access provokes significantly more L2-cache misses than dictionary-encoded columns scans, making dictionary-encoded columns scans significantly faster [20]. Columnar in-memory databases enable real-time aggregations, removing the need for pre-built aggregate tables [8]. This streamlines data models

TABLE I
COMPARISON OF COLUMNAR IN-MEMORY DATABASES

Database	o.s.*	Primary Use Case	Pros	Cons
SAP HANA by SAP	**No**	SAP HANA is designed to handle both Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) workloads on the same data representation [24], [25].	SAP HANA allows extremely fast data processing and real-time analytics [26]–[28]. SAP HANA is designed to handle large-scale data analytics and complex queries efficiently [27], [29]. SAP HANA supports the complete data lifecycle, including modeling, provisioning, and consumption, making it a holistic data management solution [26].	Integration and utilization of latest features of SAP HANA into existing systems for the effective use of its features can be time-consuming and might require significant effort [30], [31]
ClickHouse by Yandex	**Yes**	ClickHouse is optimized for performing complex analytical queries on petabyte-scale data sets with high ingestion rates [32]. The storage layer of ClickHouse combines a data format based on traditional log-structured merge (LSM) trees with novel techniques for continuous transformation of historical data.	ClickHouse uses a state-of-the-art vectorized query execution engine with optional code compilation, which enhances query processing speed [32]. ClickHouse can handle large volumes of data efficiently, making it suitable for big data applications [32], [33]. ClickHouse uses aggressive pruning techniques to avoid evaluating irrelevant data in queries [32].	The implementation and maintenance are complex [33]. Ensuring data consistency can be challenging, particularly in distributed environments. ClickHouse does not support transactional operations (OLTP) as efficiently as other databases [33].
Google BigQuery by Google	**No**	BigQuery is extensively used to design and provision data warehouses, enabling organizations to store and manage large volumes of data efficiently [34], [35]	BigQuery supports real-time insights and fast query processing, making it suitable for large-scale data analytics [36]. It offers multi-cloud support, allowing deployment on non-GCP clouds, which is beneficial for organizations with a multi-cloud strategy [37]	Preparing and de-normalizing data for import into BigQuery can be challenging and additional frameworks for preprocessing might be needed to complete it [38]. Integration and maintenance of data in multi-cloud deployment setup is complex and it has the security issues like data privacy and compliance [37], [39].
Amazon Redshift by AWS	**No**	Amazon Redshift is primarily used as a cloud-based data warehousing solution designed to efficiently analyze large volumes of data [40].	For large-scale data analysis, Redshift shows highest cost-effectiveness [41]. The integration with AWS services are very easy [40], [42].	Sometimes Redshift requires manual integration to adjust the aspects of workload scaling and optimization [43]. Cold start issues can impact the predicted exicution time and performance [43], [44].
MonetDB by CWI Amsterdam	**Yes**	MonetDB is mostly used in read-heavy scenarios of web applications [45], [46].	A physical design strategy in MonetDB improves query execution times [47]. MonetDB supports database cracking for self-organization and informative query summaries [48].	Due to the complex design of MonetDB, it require a steep learning curve to achieve full system capabilities [48]. Performance improvements of MonetDB depend on the availability and integration of high-performance hardware [48], [49]
kdb+ by KX Systems	**No**†	kdb+ is mostly used in the financial industry for real-time data analysis and time series databases [50], [51]	The Q programming language is optimized for querying time series data [52]. kdb+ can handle large-scale data environments and its architecture efficently handles data storage and retrieval [51], [52].	The Q programming language requires a steep learning curve [51]. kdb+ applications cannot run directly on SQL databases [52], creating integration issues.
Apache Arrow by Apache Foundation	**Yes**	Apache Arrow is mostly used for cross-language development platforms [53]	Apache Arrow uses DataFusion, which makes it more adaptive for OLAP engines and other data-intensive systems [54]. The utilization of data is less expensive with Arrow APIs [53], [55].	To utilize the full potential of Arrow's capabilities, sometimes the existing systems needs modifications to adapt Arrow's protocols [56]. Understanding the architecture and APIs of Apache Arrow is challenging for those new to the platform [54].
HyPer by TU Munich	**No**‡	HyPer supports hybrid transaction/analytical processing workloads, allowing it to handle real-time business analytics [57], [58]	HyPer achieves high transaction rates (up to 100,000 transactions per second) [59], [60]. The horizontal scale-up of HyPer is easy with the ScyPer extension [57].	HyPer's performance is highly dependent on the availability of sufficient main memory [61]. Needs large amounts of main memory and additional servers for scaling out [57].
Hyrise by Hasso Plattner Institute	**Yes**	Hyrise has evolved to significantly support a wide range of research areas and projects [9], [10].	The hybrid row- and column-format data storage allows Hyrise to optimize for different types of queries and workloads, [9], [62], [63]. Hyrise aims at continously integrating latest research innovations.	Replication improves the performance of Hyrise, but managing and maintaining a replicated system adds complexity [63].

*o.s.=open source †Free for personal usage. ‡Acquired by Tableau.

and simplifies applications by reducing complexity and storage overhead.

A key advantage of columnar in-memory databases is their ability to achieve high performance through vector processing over main memory-resident columns. Systems such as SAP HANA leverage this capability to execute analytical queries at unprecedented speeds [75]. Vector processing involves performing operations on entire sets of values (vectors) at once, rather than processing individual values sequentially. This approach takes full advantage of modern CPU architectures, where SIMD (Single Instruction, Multiple Data) vectorization techniques play a crucial role in optimizing query execution [76]. SIMD enables the parallel processing of multiple data points in a single CPU instruction, significantly accelerating query performance.

To enhance efficiency, columnar in-memory databases make extensive use of data compression techniques. Since columnar storage often contains repetitive values within a column, compression algorithms can significantly reduce the data footprint, leading to faster processing and reduced memory usage [68], [77]. Techniques such as run-length encoding, dictionary encoding, and delta encoding are commonly employed to compress columnar data effectively. Compression not only reduces memory requirements but also speeds up query execution by minimizing the amount of data that needs to be scanned and processed.

In Table I, we provide a comparison of nine established columnar in-memory databases in terms of their primary use case, advantages and disadvantages (pros and cons) and their public availability (whether open source or not). The investigated databaes comprise SAP HANA[5], ClickHouse[6], Google BigQuery[7], Amazon Redshift[8], MonetDB[9], kdb+[10], Apache Arrow[11], HyPer[12], and Hyrise[13].

IV. Hyrise Database

Hyrise is an advanced in-memory database system designed to deliver high-performance analytical and transactional workloads. It serves as a research database, developed to explore, analyze, and optimize the performance of database systems in handling large-scale data processing tasks. Hyrise is particularly focused on leveraging modern hardware capabilities, such as multi-core processors and large main memory capacities, to achieve superior data processing efficiency [10], [78]. By integrating cutting-edge database management techniques with modern hardware optimizations, Hyrise enables real-time data processing and hybrid transactional and analytical processing (HTAP), making it suitable for both research and enterprise applications [62].

[5]https://www.sap.com/estonia/products/data-cloud/hana.html
[6]https://clickhouse.com/
[7]https://cloud.google.com/bigquery/
[8]https://aws.amazon.com/redshift/
[9]https://www.monetdb.org/
[10]https://kx.com/products/kdb/
[11]https://arrow.apache.org/
[12]https://hyper-db.de/
[13]https://hpi.de/plattner/projects/hyrise.html

The architecture of Hyrise is centered around the concept of dictionary-encoded columnar storage, which enhances in-memory performance through efficient data compression and faster query execution. Columnar databases are particularly well-suited for analytical queries, as they allow for improved cache efficiency and reduced data retrieval times compared to traditional row-based storage systems. By implementing a hybrid row- and column-format storage mechanism, Hyrise effectively balances the needs of transactional and analytical workloads, ensuring optimal data access patterns for different types of queries [10]. This hybrid storage model facilitates seamless transitions between transactional and analytical operations, minimizing the need for data duplication or movement.

Initially developed at the Hasso Plattner Institute (HPI) in Germany, Hyrise [9], [10] has evolved into a sophisticated database system that integrates state-of-the-art database management strategies[14]. The system is designed to optimize database performance by leveraging multi-core processors, vectorized query execution, and modern storage techniques. Vectorized execution enables Hyrise to process multiple data points simultaneously by utilizing SIMD (Single Instruction, Multiple Data) instructions, significantly accelerating data processing tasks [80], [81]. By taking advantage of parallelism at the hardware level, Hyrise achieves remarkable performance improvements in executing complex analytical queries.

A key differentiator of Hyrise from traditional database systems is its ability to support HTAP workloads, effectively bridging the gap between online transactional processing (OLTP) and online analytical processing (OLAP). Traditional databases often require separate systems for handling transactional and analytical queries, leading to increased complexity and latency. Hyrise eliminates this limitation by enabling real-time analytics alongside transaction processing, making it an attractive solution for applications requiring instant insights from operational data. This capability is particularly beneficial for industries such as finance, healthcare, and e-commerce, where real-time decision-making is critical.

Hyrise incorporates advanced optimization techniques, such as adaptive indexing and partitioning, to enhance query performance based on the workload characteristics [62]. Adaptive indexing dynamically adjusts index structures to optimize query execution, reducing the overhead associated with maintaining traditional indexes. Partitioning strategies help distribute data efficiently across processing units, ensuring balanced workload execution and minimizing bottlenecks. These optimizations contribute to the system's ability to scale effectively and maintain high performance under varying workloads.

Furthermore, Hyrise continuously evolves through ongoing research and development efforts, incorporating new advancements in database technology and hardware capabilities. Researchers and developers actively explore novel techniques to further enhance its efficiency, adaptability, and scalability. As modern hardware architectures continue to advance, Hyrise

[14]another important research prototype of HPI is SanssouciDB [73], [79], which shares concepts with Hyrise and SAP Hana.

remains at the forefront of database innovation, pushing the boundaries of high-performance data processing.

Standard association rule mining (ARM) [82] often necessitates discretizing numeric target variables [83], [84], a process that can result in information loss and yield less precise insights. To overcome this limitation, we introduced the novel tool *GrandReport* [1], which enhances ARM by calculating and reporting the mean values of a selected numeric target column across all possible combinations of influencing factors. The tool enables decision-makers to interpret associations based on aggregate values as in *online analytical processing* (OLAP) [85], however, automatically as in ARM instead of interactively as in OLAP and at the scale of ARM, providing results that align more closely with real-world analytical practices.

While conducting numerous analyses using the GrandReport platform [5], we observed significant patterns of variation in results between marginal values and drill-down analyses. The observed behavioral differences suggest the presence of statistical paradoxes [86] and other data-related fallacies such as selecting suboptimal measures [87], and, last but not least, confounding effects [88], [89]. To address these challenges, we have expanded the capabilities of the GrandReport tool by incorporating additional data analytical methods [5]. This enhancement aims to generate diverse perspectives and identify potential confounding effects within datasets. Specifically, we integrated multiple linear regression adjustment into the platform to mitigate confounding effects in mixed multidimensional data [5]. This enhancement allows for more accurate analyses by adjusting for confounders without requiring the segregation of numerical and categorical data.

V. On Utilizing Hyrise for GrandReport

The GrandReport is designed for generating comprehensive reports, analytics, and insights from diverse datasets, see Section IV. Integrating the Hyrise database [10], [62] into GrandReport could significantly enhance its capabilities, particularly in terms of performance, scalability, and efficiency, see Table I. Furthermore, the primary use case of Hyrise are research projects, as it has evolved significantly to support a wide range of research areas, see Table I.

One of the primary advantages of using Hyrise for GrandReport is enabling real-time analytics and reporting. Given that GrandReport processes large volumes of data, users often require instant insights and dynamically updating dashboards. Hyrise's in-memory architecture and columnar storage format are optimized for fast data retrieval and processing, making it ideal for handling complex analytical queries in real time.

Hyrise's in-memory design and efficient data compression techniques enhance scalability, ensuring that GrandReport can manage growing data volumes without compromising performance. By storing data in memory, Hyrise eliminates the latency associated with disk-based storage, will allow GrandReport to process large datasets quickly, even as data grows over time.

GrandReport frequently performs complex, multi-dimensional queries. Hyrise's adaptive query processing capabilities dynamically optimize execution plans based on data distribution and workload. This ability to adjust query plans on the fly will ensure that GrandReport maintains optimal performance even when query patterns or underlying data change.

Additionally, GrandReport requires support for advanced analytical operations, including machine learning integration. Hyrise's architecture efficiently handles complex joins, window functions, and other analytical operations, making it well-suited for such tasks. Furthermore, Hyrise supports concurrent read and write operations, will allow GrandReport to serve multiple users simultaneously without performance degradation.

Hyrise's efficient memory usage and data compression techniques will help to reduce GrandReport's operational costs. Moreover, Hyrise seamlessly integrates with other data processing and visualization tools, which can enable GrandReport to function as a comprehensive analytics platform. Researchers will be able to leverage GrandReport with Hyrise to analyze large datasets from experiments, simulations, and observations, facilitating faster discovery and decision-making.

VI. Conclusion

In this paper, we have provided informed arguments for using columnar in-memory database technology for the high-performant implementation of the GrandReport tool. In service of that, we contribute as follows:

- We have provided a targeted review of columnar databases.
- We have provided a targeted review of columnar in-memory databases.
- We have provided a comparison of nine established columnar in-memory databases, including SAP HANA, ClickHouse, Google BigQuery, Amazon Redshift, MonetDB, kdb+, Apache Arrow, HyPer, and Hyrise.

On the basis of these investigations, we have discussed the advantages of utilizing the Hyrise database as the future implementation platform for the GrandReport tool.

Acknowledgements

This work has been conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

This study was co-funded by the European Union and Estonian Research Council via project TEM-TA141.

References

[1] S. Arakkal Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia, "Grand reports: A tool for generalizing association rule mining to numeric target values," in *Proc. of DaWaK'2020 – the 22nd Intl. Conf. on Data Warehousing and Knowledge Discovery*, ser. LNCS, vol. 12393. Cham: Springer, 2020, pp. 28–37, doi:10.1007/978-3-030-59065-9_3.

[2] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[3] D. Draheim, "Future perspectives of association rule mining based on partial conditionalization (DEXA'2019 keynote)," in *Proc. of DEXA'2019 – the 30th Intl. Conf. on Database and Expert Systems Applications*, ser. LNCS, no. 11706, 2019, p. xvi.

[4] ——, "DEXA'2019 Keynote presentation: Future perspectives of association rule mining based on partial conditionalization (2019)," doi:10.13140/RG.2.2.17763.48163.

[5] S. Arakkal Peious, M. Kaushik, S. A. Shah, R. Sharma, S. Suran, and D. Draheim, "On measuring confounding bias in mixed multidimensional data," in *Proc. of ICTIS'2024 – the 8th Intl. Conf. on ICT for Intelligent Systems, Volume 6*, ser. Lecture Notes in Network and Systems, vol. 1112. Springer, 2024, pp. 329–342.

[6] The Economist, "The real-time revolution," The Economist October 23rd-29th, pages 22–24, 2021.

[7] ——, "Instant economics," The Economist October 23rd-29th, page 13, 2021.

[8] H. Plattner, "The impact of columnar in-memory databases on enterprise systems," *Proc. VLDB Endow.*, vol. 7, no. 13, pp. 1722–1729, 2014.

[9] M. Grund, J. Krüger, H. Plattner, A. Zeier, P. Cudre-Mauroux, and S. Madden, "HYRISE - a main memory hybrid storage engine," *Proc. VLDB Endow.*, vol. 4, no. 2, pp. 105–116, 2010.

[10] M. Dreseler, J. Kossmann, M. Boissier, S. Klauck, M. Uflacker, and H. Plattner, "Hyrise re-engineered: An extensible database system for research in relational in-memory data management," in *Proc. of EDBT'2019 - the 22nd Intl. Conf. on Extending Database Technology*. OpenProceedings, 2019, pp. 313–324.

[11] E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, vol. 13, no. 6, p. 377–387, Jun. 1970.

[12] R. Elmasri and S. Navathe, *Fundamentals of Database Systems*. Addison-Wesley, 2011, ch. 2 Database System Concepts and Architecture.

[13] A. Silberschatz, H. F. Korth, and S. Sudarshan, *Database System Concepts*. McGraw-Hill Education, 2011.

[14] X. Zhao and T. Nakagawa, *Database Maintenance Models*. Cham: Springer International Publishing, 2018, pp. 183–215.

[15] M. Stonebraker et al., "C-Store: A column-oriented DBMS," in *Proc. of VLDB'05 – the 31st Intl. Conf. on Very Large Data Bases*. VLDB Endowment, 2005, pp. 553–564.

[16] W. Qiyue, "Research on column-store databases optimization techniques," in *Proc. of LISS'2015 – the 5th Intl. Conf. on Logistics, Informatics and Service*. IEEE, 2015, pp. 1–7.

[17] D. J. Abadi, P. B. Boncz, and S. Harizopoulos, "Column-oriented database systems," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1664–1665, 2009.

[18] B. Jadhawar and N. Sharma, "An intelligent optimized compression framework for columnar database," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 33, no. 1, p. 29 – 53, 2025.

[19] D. Lemire and L. Boytsov, "Decoding billions of integers per second through vectorization," *Software: Practice and Experience*, vol. 45, no. 1, pp. 1–29, 2015.

[20] H. Plattner, *A Course in In-Memory Data Management – The Inner Mechanics of In-Memory Databases*. Springer, 2014, ch. 14 Scan Performance, pp. 97–102.

[21] D. J. Abadi, S. R. Madden, and N. Hachem, "Column-stores vs. row-stores: how different are they really?" in *Proc. of SIGMOD'08 – the 2008 ACM SIGMOD Intl. Conf. on Management of Data*. ACM, 2008, pp. 967–980.

[22] X. Zeng, Y. Hui, J. Shen, A. Pavlo, W. McKinney, and H. Zhang, "An empirical evaluation of columnar storage formats," *Proc. VLDB Endow.*, vol. 17, no. 2, p. 148–161, 2023.

[23] A. Pavlo, C. Curino, and S. Zdonik, "Skew-aware automatic database partitioning in shared-nothing, parallel OLTP systems," in *Proc. of SIGMOD '12 – the 2012 ACM SIGMOD Intl. Conf. on Management of Data*. ACM, 2012, pp. 61–72.

[24] V. Sikka, F. Färber, W. Lehner, S. K. Cha, T. Peh, and C. Bornhövd, "Efficient transaction processing in SAP HANA database - The end of a column store myth," in *Proc. of SIGMOD'12 – the 2012 ACM SIGMOD Intl. Conf. on Management of Data*. ACM, 2012, pp. 731–741.

[25] N. May, A. Böhm, and W. Lehner, "SAP HANA – the evolution of an in-memory DBMS from pure OLAP processing towards mixed workloads," ser. Lecture Notes in Informatics (LNI), Proceedings, vol. 265. Gesellschaft für Informatik, 2017, p. 545 – 563.

[26] V. Sikka, F. Färber, A. Goel, and W. Lehner, "SAP HANA: The evolution from a modern main-memory data platform to an enterprise application platform," *Proc. VLDB Endow.*, vol. 6, no. 11, p. 1184 – 1185, 2013.

[27] A. Boehm, "In-memory for the masses: Enabling cost-efficient deployments of in-memory data management platforms for business applications," in *Proc. VLDB Endow.*, vol. 12, no. 12, 2018, p. 2273 – 2274.

[28] P. Rösch, L. Dannecker, G. Hackenbroich, and F. Färber, "A storage advisor for hybrid-store databases," *Proc. VLDB Endow.*, vol. 5, no. 12, p. 1748 – 1758, 2012.

[29] V. Sikka, "Re-thinking the performance of information processing systems," in *Proc. of ICDE'2013 – the 29th IEEE Intl. Conf. on Data Engineering*. IEEE, 2013, pp. 9–13.

[30] M. Wang, "Teaching data warehousing with SAP HANA," *Issues in Information Systems*, vol. 23, no. 4, pp. 254–264, 2022.

[31] M. Figueiredo, *SAP HANA Cloud in a Nutshell: Design, Develop, and Deploy Data Models using SAP HANA Cloud*. Apress Berkeley, 2022.

[32] R. Schulze, T. Schreiber, I. Yatsishin, R. Dahimene, and A. Milovidov, "Clickhouse - lightning fast analytics for everyone," in *Proc. VLDB Endow.*, vol. 17, no. 12, 2024, p. 3731 – 3744.

[33] B. Imasheva, A. Nakispekov, A. Sidelkovskaya, and A. Sidelkovskiy, "The practice of moving to big data on the case of the NoSQL database, Clickhouse," in *Optimization of Complex Systems: Theory, Models, Algorithms and Applications*, ser. Advances in Intelligent Systems and Computing, vol. 991, 2020, pp. 820–828.

[34] M. Mucchetti, *BigQuery for Data Warehousing: Managed Data Analysis in the Google Cloud*. Apress Berkeley, 2020.

[35] P. Edara and M. Pasumansky, "Big metadata: When metadata is big data," in *Proc. VLDB Endow.*, vol. 14, no. 12, 2021, p. 3083 – 3095.

[36] S. Fernandes and J. Bernardino, "What is BigQuery?" in *Proc. of IDEAS'15: the 19th International Database Engineering and Applications Symposium*. ACM, 2015, pp. 202–203.

[37] J. Levandoski et al., "BigLake: BigQuery's evolution toward a multi-cloud lakehouse," in *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, 2024, pp. 334–346.

[38] O. Dawelbeit and R. McCrindle, "A novel cloud based elastic framework for big data preprocessing," in *2014 6th Computer Science and Electronic Engineering Conference, CEEC 2014 - Conference Proceedings*, 2014, p. 23 – 28.

[39] H. D. Singh, Y. Hooda, and J. Mehta, *Deep Learning in Internet of Things for Next Generation Healthcare*. Chapman and Hall, 2024, ch. Disaster and Emergency Healthcare, pp. 120–127.

[40] A. Gupta et al., "Amazon Redshift re-invented," in *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*. ACM, 2022, pp. 2205–2217.

[41] A. Gupta, D. Agarwal, D. Tan, J. Kulesza, R. Pathak, S. Stefani, and V. Srinivasan, "Amazon Redshift and the case for simpler data warehouses," in *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, vol. 2015-May, 2015, pp. 1917–1923.

[42] N. Borić, H. Gildhoff, M. Karavelas, I. Pandis, and I. Tsalouchidou, "Unified spatial analytics from heterogeneous sources with Amazon Redshift," in *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, 2020, pp. 2781–2784.

[43] V. Nathan et al., "Intelligent scaling in Amazon Redshift," in *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, 2024, pp. 269–279.

[44] Z. Wu et al., "Stage: Query execution time prediction in Amazon Redshift," in *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, 2024, pp. 280–294.

[45] P. Boncz, M. Zukowski, and N. Nes, "MonetDB/X100: Hyper-pipelining query execution," in *Proc. of CIDR'05 – the 2nd Biennial Conf. on Innovative Data Systems Research*, 2005, pp. 225–237.

[46] A. Supriano, G. M. D. Vieira, and L. E. Buzato, "Evaluation of a read-optimized database for dynamic web applications," in *WEBIST 2008 - 4th Intl. Conf. on Web Information Systems and Technologies, Proceedings*, vol. 1. ScitePress, 2008, pp. 73–80.

[47] M. Goncalves and J. N. Mendoza, *A physical design strategy for datasets with multiple dimensions*. IGI Global, 2016.

[48] M. L. Kersten, "The database architecture jigsaw puzzle," in *Proc. of ICDE'08 - the 24th Intl. Conf. on Data Engineering*. IEEE, 2008, pp. 3–4.

[49] H. Kong, W. Lu, Y. Chen, J. Wu, Y. Zhang, G. Yan, and X. Li, "Doe: database offloading engine for accelerating SQL processing," *Distributed and Parallel Databases*, vol. 41, no. 3, pp. 273–297, 2023.

[50] R. Salama, J. McGuire, and M. K. Rosenberg, "A methodology for managing database and code changes in a regression testing framework," in *Proc. of SPLASH'12 – the 2012 ACM Conf. on Systems, Programming, and Applications: Software for Humanity*, 2012, pp. 117–120.

[51] J. Novotny, P. A. Bilokon, A. Galiotos, and F. Délèze, *Machine Learning and Big Data with kdb+/q*. Wiley-Blackwell, 2019.

[52] L. Antova et al., "Datometry Hyper-Q: Bridging the gap between real-time and historical analytics," in *Proc. of SIGMOD'16 – the 2015 ACM SIGMOD Intl. Conf. on Management of Data*, vol. 26-June-2016, 2016, pp. 1405–1416.

[53] S. Rodriguez et al., "Zero-cost, arrow-enabled data interface for Apache Spark," in *Proc. of Big Data 2021 - the 2021 IEEE Intl. Conf. on Big Data, Big Data 2021*, 2021, p. 2400 – 2405.

[54] A. Lamb et al., "Apache Arrow DataFusion: A fast, embeddable, modular analytic query engine," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2024, p. 5 – 17.

[55] J. Hildebrandt, D. Habich, and W. Lehner, "Integrating lightweight compression capabilities into Apache Arrow," in *DATA 2020 - Proceedings of the 9th International Conference on Data Science, Technology and Applications*, 2020, p. 55 – 66.

[56] J. Cabral, K. Dorofeev, and P. Varga, "Native OPC UA handling and IEC 61499 PLC integration within the Arrowhead framework," in *Proceedings - 2020 IEEE Conference on Industrial Cyberphysical Systems, ICPS 2020*, 2020, p. 596 – 601.

[57] T. Mühlbauer, W. Rödiger, A. Reiser, A. Kemper, and T. Neumann, "ScyPer: Elastic OLAP throughput on transactional data," in *Proc. of DanaC'2013 – the 2nd Workshop on Data Analytics in the Cloud*. ACM, 2013, p. 11–15.

[58] A. Kemper and T. Neumann, "The database group at TUM," *SIGMOD Record*, vol. 43, no. 3, p. 55 – 60, 2014.

[59] F. Funke, A. Kemper, and T. Neumann, "Hypersonic combined transaction and query processing," in *Proc. VLDB Endow.*, vol. 4, no. 12, 2011, p. 1367 – 1370.

[60] A. Kemper and T. Neumann, "Hyper: A hybrid OLTP & OLAP main memory database system based on virtual memory snapshots," in *Proceedings - International Conference on Data Engineering*, 2011, p. 195 – 206.

[61] L. Ma et al., "Larger-than-memory data management on modern storage hardware for in-memory OLTP database systems," in *Proc. of DaMoN'16 – the 12th Intl. Workshop on Data Management on New Hardware*. ACM, 2016, pp. 1–7.

[62] D. Schwalb, M. Faust, J. Wust, M. Grund, and H. Plattner, "Efficient transaction processing for Hyrise in mixed workload environments," in *Proc. of IMDM'2013 – the 1st and 2nd International Workshops on In Memory Data Management and Analysis*, ser. Lecture Notes in Computer Science, vol. 8921. Springer, 2015, pp. 112–125.

[63] D. Schwalb, J. Kossmann, M. Faust, S. Klauck, M. Uflacker, and H. Plattner, "Hyrise-R: Scale-out and hot-standby through lazy master replication for enterprise applications," in *Proc. of IMDM'15 – the 3rd VLDB Workshop on In-Memory Data Mangement and Analytics*. ACM, 2015, pp. 1–7.

[64] P. Bandagale and V. A. Bharadi, "Data analytics on columnar databases in big data environment," in *Intelligent Computing and Networking*, ser. Lecture Notes in Networks and Systems, vol. 146, 2021, pp. 125–135.

[65] D. Abadi, P. Boncz, S. Harizopoulos, S. Idreos, and S. Madden, "The design and implementation of modern column-oriented database systems," *Foundations and Trends in Databases*, vol. 5, no. 3, p. 197 – 280, 2012.

[66] F. Elotmani, R. Esbai, and M. Atounti, "Creation of column-oriented NoSQL databases automatically in big data environments and its impact on energy consumption," *E3S Web of Conferences*, vol. 412, no. 01108, 2023, doi:10.1051/e3sconf/202341201108.

[67] E. Ivanova and L. Sokolinsky, "Parallel processing of very large databases using distributed column indexes," *Programming and Computer Software*, vol. 43, no. 3, p. 131 – 144, 2017.

[68] H. Jiang, C. Liu, J. Paparrizos, A. A. Chien, J. Ma, and A. J. Elmore, "Good to the last bit: Data-driven encoding with CodecDB," in *Proc. of Sigmod'21 – the 2021 ACM SIGMOD Intl. Conf. on Management of Data*. ACM, 2021, pp. 843–856.

[69] T. Mladenova, Y. Kalmukov, M. Marinov, and I. Valova, "Impact of data compression on the performance of column-oriented data stores," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, p. 416 – 421, 2021.

[70] A. Hall, O. Bachmann, R. Büssow, S. Gănceanu, and M. Nunkesser, "Processing a trillion cells per mouse click," in *Proc. VLDB Endow.*, vol. 5, no. 11, 2012, pp. 1436 – 1446.

[71] N. K. Seera and S. Taruna, "Leveraging mapreduce with column-oriented stores: Study of solutions and benefits," *Advances in Intelligent Systems and Computing*, vol. 654, pp. 39–46, 2018.

[72] M. Boussahoua, O. Boussaid, and F. Bentayeb, "Logical schema for data warehouse on column-oriented NoSQL databases," in *Proc. of DEXA 2017 – the 28th Intl. Conf. on Database and Expert Systems Applications*, ser. LNCS, vol. 10439, 2017, p. 247 – 256.

[73] H. Plattner, *A Course in In-Memory Data Management – The Inner Mechanics of In-Memory Databases*. Springer, 2014.

[74] R. Freeze and S. Bristow, "In-memory and column storage changes IS curriculum," in *Proc. of HICCS'23 – the 56th Hawaii Intl. Conf. on System Sciences*. AIS, 2023, pp. 6359–6366.

[75] R. Sherkat, C. Florendo, M. Andrei, A. K. Goel, A. Nica, P. Bumbulis, I. Schreter, G. Radestock, C. Bensberg, D. Booss, and H. Gerwens, "Page as you go: Piecewise columnar access in SAP HANA," in *Proc. of SIGMOD'16 – the 2016 ACM SIGMOD Intl. Conf. on Management of Data*, vol. 26-June-2016. ACM, 2016, p. 1295 – 1306.

[76] S. Chavan, A. Hopeman, S. Lee, D. Lui, A. Mylavarapu, and E. Soylemez, "Accelerating joins and aggregations on the Oracle in-memory database," in *Proc. of ICDE'2018 - the 34th IEEE Intl. Conf. on Data Engineering*. IEEE, 2018, pp. 1453–1464.

[77] J. Wust, J.-H. Boese, F. Renkes, S. Blessing, J. Krueger, and H. Plattner, "Efficient logging for enterprise workloads on column-oriented in-memory databases," in *Proc. of CIKM '12 – the 21st ACM Intl. Conf. on Information and Knowledge Management*. ACM, 2012, pp. 2085–2089.

[78] N. Shekhar and A. V. Pawar, "Big data analytics based on in-memory infrastructure on traditional HPC: A survey," in *Proc. of ICTCS'16 – the 2nd Intl. Conf. on Information and Communication Technology for Competitive Strategies*. ACM, 2016.

[79] H. Plattner, "SanssouciDB: An in-memory database for processing enterprise workloads," ser. Lecture Notes in Informatics (LNI), vol. 180. Gesellschaft für Informatik (GI), 2011, pp. 2–21.

[80] D. Vuta-Popescu, I. C. Antofi, C. B. Ciobanu, and C. Z. Kertesz, "SIMD extensions - A historical perspective," in *Proc. of SIITME'24 – the 2024 IEEE International Symposium for Design and Technology of Electronics Packages*. IEEE, 2024, pp. 108–115.

[81] D. Mustafa, R. Alkhasawneh, F. Obeidat, and A. S. Shatnawi, "MIMD programs execution support on SIMD machines: A holistic survey," *IEEE Access*, vol. 12, pp. 34 354–34 377, 2024.

[82] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. of VLDB'1994 – the 20th Intl. Conf. Very Large Data Bases*, vol. 1215. Morgan Kaufmann Publishers, 1994, pp. 487–499.

[83] M. Kaushik, R. Sharma, S. Arakkal Peious, M. Shahin, S. B. Yahia, and D. Draheim, "On the potential of numerical association rule mining," in *Proc. of FDSE'20 – the 7th Intl. Conf. on Future Data and Security Engineering*, ser. Communications in Computer and Information Science, vol. 1306. Springer, 2020, pp. 3–20.

[84] M. Kaushik, R. Sharma, A. Vidyarthi, and D. Draheim, "Discretizing numerical attributes: An analysis of human perceptions," in *New Trends in Database and Information Systems*, ser. CCIS, vol. 1652. Springer, 2022, pp. 188–197.

[85] R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. S. Yadav, and D. Draheim, "Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics," in *Proc. of DASFAA'2022 – the 27th Intl. Conf. on Database Systems for Advanced Applications*, ser. LNCS, vol. 13245. Springer, 2022, pp. 596–603.

[86] R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim, "Why not to trust big data: Discussing statistical paradoxes," in *Proc. of DASFAA'2022 Intl. Workshops – the 27th Intl. Conf. on Database Systems for Advanced Applications*, ser. LNCS, vol. 13248. Springer, 2022, pp. 50–63.

[87] R. Sharma, M. Kaushik, S. Arakkal Peious, S. B. Yahia, and D. Draheim, "Expected vs. unexpected: Selecting right measures of interestingness," in *Proc. of DaWaK'2020 – the 22nd Intl. Conf. on Big Data Analytics and Knowledge Discovery*, ser. LNCS, vol. 13428. Springer, 2020, pp. 38–47.

[88] K. Jager, C. Zoccali, A. MacLeod, and F. Dekker, "Confounding: What it is and how to deal with it," *Kidney International*, vol. 73, no. 3, pp. 256–260, 2008.

[89] J. Pearl and D. McKenzie, *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books, 2018.

Appendix II

II

S. Arakkal Peious, M. Kaushik, S. A. Shah, R. Sharma, S. Suran, and D. Draheim. On measuring confounding bias in mixed multidimensional data. In J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, editors, *Proceedings of ICTIS'2024 – the 8th International Conference on ICT for Intelligent Systems, Volume 6*, volume 1112 of *Lecture Notes in Network and Systems*, pages 329–342, Singapore, 2024. Springer Nature Singapore. doi:10.1007/978-981-97-6684-0_27

On Measuring Confounding Bias in Mixed Multidimensional Data

Case of Coerced Multiple Linear Regression Adjustment

Sijo Arakkal Peious[1 [0000−0002−7858−9463]],
Minakshi Kaushik[1 [0000−0002−6658−1712]],
Syed Attique Shah[2 [0000−0003−2949−7391]], Rahul Sharma[1,4 [0000−0002−9024−8768]],
Shweta Suran[1,3 [0000−0001−7180−731X]], and Dirk Draheim[1 [0000−0003−3376−7489]]

[1] Information Systems Group, Tallinn University of Technology, Estonia
{sijo.arakkal,minakshi.kaushik,rahul.sharma,dirk.draheim}@taltech.ee
[2] School of Computing and Digital Technology - CDT, Birmingham City University, Birmingham
syed.shah@bcu.ac.uk
[3] The Open University, Walton Hall, Milton Keynes, United Kingdom
shweta.suran@open.ac.uk
[4] Ajay Kumar Garg Engineering College, Ghaziabad, India
sharmarahul@akgec.ac.in

Abstract. Currently, we witness a significantly increased understanding, that disaggregation of data is the key to better-informed decision-making and enactment. The UN 2030 Agenda for Sustainable Development prominently stresses the importance of disaggregated data for achieving its goals. We argue that the key benefit of data disaggregation is the increased potential to understand confounding effects. Unfortunately, in today's tool landscape, it lacks systematic support for understanding confounders. Therefore, in this paper, we contribute as follows. We integrate a means of confounder adjustment, i.e., multiple linear regression adjustment, as a new feature into the data analysis tool GrandReport, which works in mixed mode, i.e., for both numerical and categorical data in parallel. Next, we conduct experiments on an extended air pollutants data set from several cities. We utilize the tool GrandReportl to investigate the correlation between standardized air quality indices and CO2 levels before and after confounder adjustment in regard of additional city data. We argue that the experiments provide evidence for the usefulness and usability of the suggested approach.

Keywords: Decision support systems, association rule mining, linear regression, confounding, confounder adjustment, environmental sciences, smart cities

1 Introduction

Currently, we witness a significantly increased understanding among top-level policymakers and decision makers, that disaggregation of data is the key to better

informed decision making and enactment. The United Nations 2030 Agenda for Sustainable Development stresses the importance of disaggregated data for monitoring the achievement of the UN sustainability goals[5]:

> "Our Governments have the primary responsibility for follow-up and review, at the national, regional and global levels, in relation to the progress made in implementing the Goals and targets over the coming fifteen years. [...] Quality, accessible, timely and reliable *disaggregated data* will be needed to help with the measurement of progress and to ensure that no one is left behind." [30]

In the same vein, the Statistics Division of the UN Department of Economic and Social Affairs explains the fundamental role of disaggregated data:

> "Improving *data disaggregation* is fundamental for the full implementation of the SDG indicator framework to fulfil the ambition of the 2030 Agenda for Sustainable of leaving no one behind."[6]

Higher the data disaggregation immediately allows for more precise, more fine-grained monitoring and control, albeit in terms of more complexity and therefore more efforts. When it comes to decision making, data disaggregation allows for more sophisticated reasoning as follows. The higher the disaggregation of data, the more latent variables become explicit. Therefore, the higher the disaggregation, the better we can potentially understand confounding effects. We argue that the benefits of disaggregation for decision-making can even be characterized as the essentially improved potential to understand confounders. However, unfortunately, in today's tool landscape, it lacks a systematic, mature support for understanding confoundig effects.

With the current study, we aim at contributing to the development of tool-integrated support for understanding confounders. With this paper, we contribute as follows:

- *Tool Integration.* The new data analysis feature of GrandReport[7] [10, 4, 2, 24] tool, i.e., *multiple linear regression adjustment*, we utilize this feature to find the confounder. The adjustment feature works in mixed multidimensional data scenarios, i.e., it allows for the incorporation of both numerical and categorical variables in the same scenario.
- *Experiments.* With GrandReport, we conduct experiments on an extended air pollutants data set. In service of this, we first join air pollutants data of several cities with CO2 level data and, furthermore, additional disaggregated data from cities from various sources. Then, we utilize the tool to investigate the correlation between standardized air quality indices[8] [8] and CO2 levels

[5] https://sdgs.un.org/goals
[6] https://unstats.un.org/sdgs/iaeg-sdgs/disaggregation/
[7] http://grandreport.me/
[8] https://www.legislation.gov.au/Details/F2016C00215

(i) before confounder adjustment in regard of the additional city data, and
(ii) after confounder adjustment in regard of the additional city data.

– *Results.* We argue that the experiments confirm the need of systematic tool-integrated support for confounder adjustment. We argue that the experiments provide evidence for the usability of the suggested tool-integration approach.

We proceed as follows. In Sect. 2, we discuss and define fundamental concepts that are relevant to this study. In Sect. 3, describes the integration of a novel feature for confounder analysis into the tool GrandReport. In Sect. 4, we explain the experimental setup of consisting experiments with standardized air quality indices and CO2 levels before and after confounder adjustment in regard of additional city data. In Sect. 5, we evaluate the experimental results. In Sect. 6, we briefly discuss potential future directions and finish the paper with a conclusion in Sect. 7.

2 Background and Definitions

2.1 Related Work

Confounding arises as a causal misconception when an additional variable is connected to both the subject of interest and the ultimate result, and this connection is not adequately addressed in the design or analysis of the study. Consequently, the apparent link between the subject of interest and the outcome might be attributed to the confounding variable rather than the subject of interest itself.

In their work, Greenland and Robins [21] highlight the significance of confounding in causal inference. Confounding occurs when a third variable is correlated with both the studied exposure and the eventual outcome, potentially leading to a distortion in estimating the causal effect. The authors delve into diverse approaches for recognizing and managing confounding, such as stratification, matching, and propensity score methods. Furthermore, they explore the relevance of addressing selection bias in the realm of causal inference.

Austin (2011) elucidates a collection of statistical methodologies designed to mitigate the impacts of confounding in observational studies. This exploration extends beyond the mere presentation of statistical formulas and equations, providing practical insights into the application of these techniques in real-world research. The study posits that the proposed method is particularly advantageous in identifying confounding variables compared to regression methods.

Judea Pearl and his team provide a layman-friendly introduction to causal inference, as outlined in their works [19, 20]. Emphasizing comprehension of diverse causal relationships and discerning them from misleading associations, the authors delve into various approaches for recognizing causal connections. These include randomization, controlled experiments, and structural equation modelling.

These instances represent a handful of numerous studies that have explored or introduced techniques for detecting confounding effects or variables within

datasets. For instance, in works such as [31, 11, 29, 21, 27], various authors elucidate methods for recognizing and managing confounding. Additionally, [22] discusses diverse approaches to identifying and controlling for confounders as explained by different researchers.

2.2 The GrandReport Tool

The continuous development of information technology created a massive amount of data [13]. Association rule mining(ARM) [1, 28], introduced as early as in the 1990s, has become a standard technique in data mining. Association rule mining technique identifies the interesting patterns or the combinations of the data items with the help of classical measures such as support, confidence and lift [26, **?**]. As such it is well-established in today's data mining tool landscape as well as the subject of exhaustive research endeavours. Unfortunately, the inclusion of numerical target column is achieved only by discretization [25, 15] in today's association rule mining tools. This is a severe limitation, because, when it comes to numeric target values, decision-makers and domain experts want to base their arguments on mean values. The GrandReport have addressed this problem. The tool reports mean values of chosen numerical target columns against all combinations of influencing factors. A *grand report* [10] can be characterized as the complete print-out of all possible generalized association rules [10, 2]. The potential of the approach has been investigated, among others, in [10, 2, 4, 3, 24, 5].

2.3 Confounding

Confounding poses a significant challenge in scientific research, particularly in understanding the impact of an influencing variable on a given outcome or target variable [3]. When confounding occurs, it hides the actual effect of the variable being studied. Thus, controlling or eliminating this confounding influence becomes crucial to obtaining accurate and meaningful results.

Confounding complicates the accurate assessment of how an exposure truly affects an outcome. A variable is considered as a potential confounder if it meets certain criteria like 1; it must connected to the target variable. 2; it must be connected with the influencing variable and distributed unevenly among influencing variable groups. 3; it must not be an outcome of the influencing variable [14]. The confounding bias is often characterized as a mixing of effects [32, 12]. It happens during the impact analysis of an influencing variable on the occurrence of a target variable but unintentionally calculates the impact of another variable. The variable whose effect is calculated is known as the confounding variable [14].

Addressing confounding begins with careful consideration during the study design phase, a pivotal point where researchers can implement strategies to minimize its potential impact. One such strategy is randomization, where participants are randomly assigned to different groups, ensuring an even distribution of

potential confounding variables across these groups [14]. Restriction involves limiting the study population to specific characteristics minimizing the variability of confounding factors.

However, the battle against confounding continues after the study design. Even after completing the study, researchers have options to control for confounding and refine their findings. Stratification allows for the analysis of subgroups based on potential confounders, enabling a clearer understanding of the variable's effect within each subgroup. Multivariate analysis, a more complex technique, involves considering multiple variables simultaneously to assess their independent effects, effectively controlling for confounding.

Importantly, successful adjustment for confounding in post-study analysis depends on the availability and accuracy of information about the potential confounding factors collected during the study. Accurate and comprehensive data are fundamental in addressing confounding and obtaining reliable conclusions from the research. In summary, recognizing and managing confounding at various stages of the research process are essential steps to ensure the integrity and validity of the study's results.

2.4 Coerced Multiple Linear Regression Adjustment

We assume that a *statistical data set* D of n data points is given as an indexed set as follows:

$$D = \big(\langle y_i, x_{i_1}, \underbrace{x_{i_2}, \ldots, x_{i_p}}_{confounders}\rangle\big)_{1\leq i\leq n} \tag{1}$$

The data set D is said to consist of one y-column and p x-columns. The several columns of the data points in (1) have different roles in regard of the intended linear regression analysis as follows. We call y_i the *target variable*. We call x_{i_1} the *primary influencing factor*. We call $x_{i_2}, \ldots, x_{i_p}$ *potential confounders* and also *further influencing factor*. (When clear from the context, the *primary influencing factor* is sometimes simply called *the influencing factor* as opposed to the further influencing factors, which are then also simply called *the confounders*.)

On the basis of that, we define the *marginal regression line* in terms of D as follows:

$$y_i = \beta_0^m + \beta_1^m x_{i_1} + \epsilon_i^m \tag{2}$$

As usual, (2) is considered to be the solution to the respective optimization problem of linear regression with ϵ_i^m represents a normally distributed error term.

Next, we define the *adjusted regression line* in terms of D as an adjustment measure as follows:

$$y_i = \beta_0^a + \beta_1^a x_{i_1} + \beta_2^a x_{i_2} + \cdots + \beta_p^a x_{i_p} + \epsilon_i^a \tag{3}$$

In the context of confounder analysis, the coefficient β_1 is the fundamental regression analysis parameter. It represents the effect of the primary influencing variable x_{i_1} onto the target variable y_i. When this coefficient changes significantly (e.g., more than 10%) after a further influencing variable x' has been added to the regression analysis, it becomes important to consider this x' as a confounding variable [17, 6, 18, 23]. Addressing confounders in analysis gives a comprehensive understanding of the correlation between the exposure and the outcome. It helps in decision-making based on the analysis and generates valid interpretations.

Against this background, we now define the *adjustment lift* α in terms of D as follows:

$$\alpha = \frac{\beta_1^a}{\beta_1^m} \tag{4}$$

Furthermore, we use the following adjustment measure ψ, because it is used in practice and the literature [17, 6, 18, 23], and we call it *adjustment measure in terms of percentage difference*:

$$\psi = \frac{\beta_1^a - \beta_1^m}{\left(\frac{\beta_1^a + \beta_1^m}{2}\right)} \tag{5}$$

In practice and literature, the measure (5) is often utilized to define a threshold for the identifying confounders, where a threshold of 10% is a usual, arbitrary threshold.

3 Tool-Integrated Confounder Adjustment

For this study, the techniques specified in Section 2.4 has been included in the GrandReport tool. The integrated method allows us to effectively apply these techniques within the context of the tool, enhancing its analytical capabilities and providing valuable insights through the implementation of multiple linear regression. Figure 2.4 shows a snippet of a report generated by GrandReport.

In order to generate a comprehensive report using the GrandReport tool, the user initiates the process by choosing the input source – either an Oracle database or an Excel file. Subsequently, the users can select the specific table or sheet, depending on the chosen source, that they intend to utilize for the report generation. The tool then displays the column names, presenting the user with two essential options.

- The first option allows the user to designate the target variable, pinpointing the central focus of the analysis. However, at present, the tool permits the selection of only one column as the target variable.
- Subsequently, the second option enables the user to specify the influencing factors which play a crucial role in the analysis process.

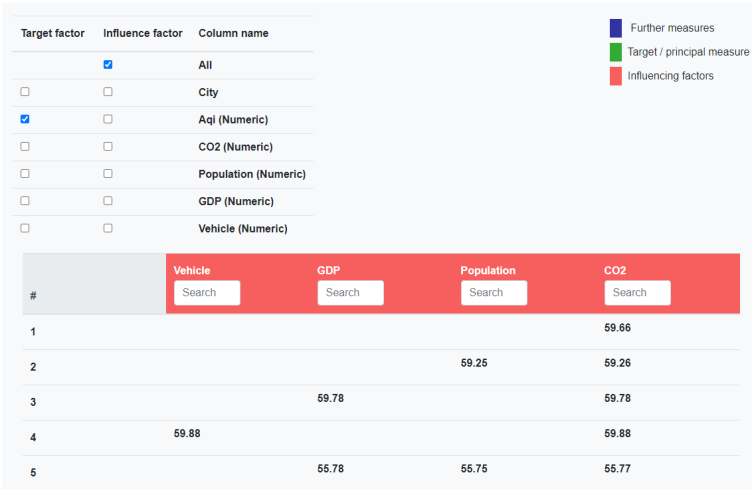


Fig. 1. The report generated by GrandReport.

The GrandReport tool achieves an integration of *association rule mining* [1, 28] and Pearson correlation reports [3]. With the novel feature, the GrandReport tool steps further by integrating and facilitating multiple linear regression. An illustrative representation of the tool's interface is shown in Figure 2.4, where distinct colours have been strategically employed for various columns. These colours help the user to identify each column's behaviour in the course of the analysis.

Through this systematic approach, GrandReport facilitates a seamless and efficient process of generating insightful and visually distinguishable reports for effective decision-making and analysis.

GrandReport is an ASP.NET framework web application with Ajax and JSON functions for seamless data transfer. This integration involved the utilization of the MathNet package, specifically employing LinearRegression. We called the *MultipleRegression.NormalEquations* function to execute the multiple linear regression by passing our influential and target variables.

4 Experimental Setup

For our experiments, we have used meteorological data from China, i.e., the CMDC (China Meteorological Data Service Centre) data set[9] [7]. The CMDC portal[10] was developed with the primary objective of simplifying the sharing

[9] https://www.legislation.gov.au/Details/F2016C00215
[10] http://data.cma.cn/en

and dissemination of daily meteorological data. By doing so, it facilitates advancements in science and technology by promoting seamless collaboration and knowledge exchange within the scientific community.

The CMDC dataset plays an essential part in the research of meteorological data in China. This data contains comprehensive and crucial environmental parameters, including pollutants such as PM2.5, PM10, SO2, NO2, O3, and CO. Additionally, the CMDC also provides data on the average Air Quality Index (AQI) for each city across China. The availability of the mean Air Quality Index (AQI) for every city in China within the CMDC dataset is used for studying air quality across different areas. This information is crucial for decision-makers and researchers.

In our experiments, the AQI from CMDC dataset has been used as the target variable to be investigated. To conduct a useful confounder analysis, we have extended the CDMC dataset with additional parameters, resulting into the following data per city:

- *Target Variable:* Air Quality Index (AQI)
- *Primary Influencing Factor:* CO2 level
- *Potential Confounders:* population density, gross domestic product, numbers of vehicles.

We argue that the usage of the CMDC dataset for our experiments is a particularly good choice, given its vast and diverse availability of meteorological and pollutant-related information. With the vast set of parameters, the CMDC dataset empowers researchers to delve into crucial environmental problems and generate sustainable solutions, in alignment with the CMDC's primary objectives of promoting scientific community and collaboration in meteorology and environmental science.

4.1 Compilation of the Data Set

AQI is a critical measurement for assessing air quality in different cities, it also provides the implications of residing in cities with different AQI levels. The AQI index plays a critical role in making sensible decisions by the policymakers about the health and well-being of individuals.

The Australian Capital Territory[11] has defined a specific method for calculating AQI. According to it, the reading of a particular pollutant is divided by the corresponding pollutant standard value, and the result is then multiplied by 100. This method produces the AQI value for a specific pollutant, delivering an explicit indication of the pollution levels in the air[11]. Australian national standard values[12] for each pollutant are given in Table 1.

The pollutant standard value used in this calculation may vary according to the national guidelines or national air quality standards. The variations in air

[11] https://www.health.act.gov.au/about-our-health-system/population-health/environmental-monitoring/air-quality/measuring-air

[12] https://www.legislation.gov.au/Details/F2016C00215

quality standards are essential to the analysis of air quality to the circumstances of each nations [13]. The air quality standards for different countries are available in the Air Quality Standard portal[14], this enriching the awareness of AQI calculations and their importance across the globe.

The Air Quality Index (AQI) is a comprehensive calculation that takes into account different pollutants present in the air. The average AQI is the product of all pollutants. In order to improve the accuracy and profoundness of our research, we expanded our investigation beyond just the pollutants. We undertook a comprehensive process by integrating further influencing factors which play an important role in air quality. This extended approach involved integrating variables such as CO2 levels, population density, gross domestic product (GDP) data, and the number of vehicles in each city. A sample of the extended data set is given in table 2 The extended influencing factors have been incorporated from the China Urban Carbon Dioxide Emission Dataset, accessible at [15]. This extended integration enriches the deepness of research, providing a more complete perspective on air quality determinants.

The inclusion of additional influencing factors to the air quality helps to understand the impact and correlations of human density, carbon dioxide emissions, economic activities of a region and traffic-related pollution of each city. By adding these influencing factors in to AQI analysis, we are trying to identify the confounding effects that influence the air quality. This method facilitates a better and vast understanding of the AQI, enabling sensible decision-making [4] and improving air quality management in cities.

Table 1. Standards for pollutants according to [8].

Pollutant	**Avg. Period**	**Maximum**
Carbon monoxide	8 hours	9.0 ppm
Nitrogen dioxide	1 hour	0.12 ppm
	1 year	0.03 ppm
Ozone	1 hour	0.10 ppm
Sulfur dioxide	4 hours	0.08 ppm
	1 hour	0.20 ppm
	1 day	0.08 ppm
	1 year	0.02 ppm
Lead	1 day	50 μg/m^3
Particles as PM_{10}	1 day	50 μg/m^3
	1 year	25 μg/m^3
Particles as PM_{25}	1 day	25 μg/m^3
	1 year	8 μg/m^3

[13] https://www.transportpolicy.net/topic/air-quality-standards
[14] https://www.transportpolicy.net/topic/air-quality-standards/
[15] http://www.cityghg.com/toArticleDetail?id=203

Table 2. A Sample table of standardized air quality indices and CO2 levels with additional city data.

City	Aqi	CO2(t)	Population/10^4	GDP/10^8	Vehicles
Ankang	78.71	446	249	1089	242,86
Anqing	46.7	2462	528	2468	650,957
Anshan	54.90	8128	333	1739	382,184
Anshun	62.25	1505	247	967	133,069
Anyang	67.4	6368	548	2301	950,443
Baicheng	47.38	1454	155	492	305,672
Baishan	36.5	861	95	509	156,612
Baiyin	69.33	2342	151	497	268,633
Baoding	71	6445	924	3353	964,879
Baoji	112	2439	332	2277	492,2

4.2 Conducted Experiments

In this experiment, AQI represents the target variable (dependent variable). The remaining columns (CO2, Population density, GDP, Number of vehicles) represent the influencing factors (independent variables).To identify the adjusted regression line Multiple linear regression method is used, with influencing factors as the input and AQI as the target factor. Then, We performed linear regression for each influencing factor with a target factor to identify the marginal regression line.

Subsequently, we created the different combinations of primary influencing factor CO2, by comparing it with potential confounding variables such as population density, GDP, and the number of vehicles. By doing so, we aimed to perform Multiple linear regression to identify the effect of these potential confounding variables on AQI when considering CO2 as the primary influencing factor. To quantify the influence of confounding, we calculated the percentage difference between the adjusted and marginal regression lines. This calculation is crucial for identifying the difference in the relationship between AQI and CO2 while considering the confounding variables.

It is important to realise that defining a cutoff point for identifying an influencing variable as a confounding variable can differ based on several factors. These include sample size, influencing factors correlation, standard deviation error, etc [18]. The experimented method delivers a strong framework to explain the detailed association between influencing factors, AQI and CO2 level.

Table 3. Marginal regression line for influencing factors

Influencing factors	Regression line
CO2	59.6598
Population density	59.8219
GDP	61.9739
Number of Vehicle	61.3835

Table 4. Adjusted regression line for the Air Quality Index (AQI) as target variable dependent on CO2 (primary influencing factor) and various combinations of other influencing factors: the slopes of adjusted regression lines are shown for CO2, population density (=Population/Pop.), gross domestic product (=GDP), and number of vehicles (=Vecicles/Veh.).

Combinations of Further Influencing Factors	**Regression Slope**				**Adjustment Measure in Terms of Percentage Difference**	**Adjustment lift**
(Potential Confounders)	CO2 β_1	Pop. β_2	GDP β_3	Veh. β_4	ψ-Measure (5)	α-Measure (4)
unadjusted/marginal	*59.66*				*0.00*	*1.00*
Population	59.26	59.25	–	–	0.68	0.99
GDP	59.78	–	59.78	–	-0.21	1.00
Vehicles	59.88	–	–	59.88	-0.37	1.00
Population, GDP	55.77	58.75	60.78	–	6.73	0.93
Population, Vehicles	58.70	59.70	–	59.71	1.62	0.98
GDP, Vehicles	59.45	–	59.45	59.45	0.34	1.00
Population, GDP, Vehicles	55.49	58.47	59.49	59.49	7.24	0.93

5 Experimental Results

In the process of coerced multiple linear regression adjustment, two types of regression lines, namely the marginal and the adjusted regression lines, are generated and then compared. The marginal regression line involves the computation of the slopes for each individual influencing factor and the target factor. These slopes are presented in Table 3. On the other hand, the adjusted regression line is created by performing multi-linear regression, using combinations of influencing factors as input and the target factor as the output. The resulting slopes of the adjusted regression for various combinations of CO2 with other influencing factors are detailed in Table 4. Additionally, the table includes the adjustment measure in terms of percentage difference and adjustment lift for CO2 with potential confounders.

The adjusted regression delivers an in-depth knowledge of the effects of various combination factors, it is creating a more useful interpretation of the correlation between the primary influencing factor (CO2) and the target variable (AQI) variable in comparison to the marginal approach.

The presented results shed light on the presence of a confounding effect related to the Gross Domestic Product (GDP) and Population Density when considering CO2 as the primary influencing factor in the context of the Air Quality Index (AQI). This effect manifests distinctly only when both GDP and Population Density are considered in combination; independently, these variables do

not exhibit any confounding effect. Moreover, the third potential confounding variable, Number of Vehicles, does not demonstrate a confounding effect on AQI either when analyzed in isolation or in various combinations.

GDP and Population Density show the characteristics of confounding variables when analysing their correlation with CO2 and AQI. The experiment does not define these are the only confounding variables of AQI and CO2, but somewhat emphasises their possible role as such founded on the monitored effects on the correlation between standardized air quality indices and CO2 levels.

6 Future Directions

There are two major strands of future work. The first is about more generalized linear regression adjustment and its "grand" reporting in terms of arbitrary groups of influencing factors and arbitrary groups of confounders. The report in this paper is limited to a single primary influencing factor that is taken from the first column of influencing factors, see Table 4. The conducted experiment was chosen to provide evidence for the usefulness of the feature, yet, the feature needs to be generalized significantly to realize the very notion of the GrandReport tool to serve as a tool for wide, exploratory analysis. Therefore, in future work, we will generalize the method for multiple primary influencing factors. Hand-in-hand with that, the generalization will allow for primary influencing factors to be taken from arbitrary positions.

The second strand of future work is about a systematic comparison of linear regression adjustment with categorical adjustment as an alternative confounder-adjustment measure. In our work, we want to utilize a generalized form of categorical adjustment that allows for the inclusion of numerical target values [9, 2, 24, 16] as well as numerical influencing factors [9, 2, 24]. The inclusion of numerical influencing factors is achieved by *coercion*, i.e., by splitting numerical columns into partitions. We will use standard methods for that such as splitting column data along the median or other quantiles.

7 Conclusion

The UN 2030 Agenda for Sustainable Development strongly emphasizes the necessity of using disaggregated data to effectively meet its objectives. We contend that the primary advantage of disaggregating data lies in its potential to better comprehend confounding effects within intricate systems. Unfortunately, obtaining such disaggregated data is often challenging due to limitations in available analytical tools or the inherent complexity of performing these processes. The Integrated multiple linear regression adjustment in GrandReport introduces a potent mechanism to identify and account for confounding effects when analyzing complex data sets.

In this research, we utilize the new feature of Grandreport tool, i.e., newly integrated multiple linear regression adjustment as a novel feature to investigate the correlation between standardized air quality indices and CO2 levels. This

investigation was conducted both before and after adjusting for potential confounders, using additional city data. As shown in the result, this approach can identify confounding effects of influencing variables. The experiments conducted with GrandReport provided compelling evidence of the viability and efficacy of the suggested approach. This research highlights the critical importance of identifying and addressing confounding variables in data analysis, underscoring the potential for significant advancements in our understanding of complex systems and contributing to informed decision-making, aligning with the UN 2030 Agenda for Sustainable Development.

Acknowledgements

This work has been conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. pp. 207–216 (1993)
2. Arakkal Peious, S., Sharma, R., Kaushik, M., Attique Shah, S., Ben Yahia, S.: Grand reports: A tool for generalizing association rule mining to numeric target values. In: Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery. Lecture Notes in Computer Science, vol. 12393, pp. 28–37. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59065-9_3
3. Arakkal Peious, S., Sharma, R., Kaushik, M., Shahin, M., Draheim, D.: On observing patterns of correlations during drill-down. In: Proceedings of iiWAS'2023 – the 25th International Conference on Information Integration and Web-based Applications & Services. LNCS (2023)
4. Arakkal Peious, S., Suran, S., Pattanaik, V., Draheim, D.: Enabling Sensemaking and Trust in Communities: An Organizational Perspective, p. 95–103. Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3487664.3487678
5. Bertl, M., Shahin, M., Ross, P., Draheim, D.: Finding indicator diseases of psychiatric disorders in bigdata using clustered association rule mining. In: Proceedings SAC'2022 – the 38th ACM/SIGAPP Symposium on Applied Computing. pp. 826—-833. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3555776.3577594
6. Budtz-Jørgensen, E., Keiding, N., Grandjean, P., Weihe, P.: Confounder selection in environmental epidemiology: assessment of health effects of prenatal mercury exposure. Ann Epidemiol **17**(1), 27–35 (2007). https://doi.org/10.1016/j.annepidem.2006.05.007
7. China Data Lab: Meteorological Data – draft version (2020). https://doi.org/10.7910/DVN/TU0JDP
8. Department of the Environment: National Environment Protection (Ambient Air Quality) Measure. Federal Register of Legislative Instruments F2016C00215 (2016)

9. Draheim, D.: Future perspectives of association rule mining based on partial conditionalization. In: Proceedings of DEXA'2019 – the 30th International Conference on Database and Expert Systems Applications. p. xvi. No. 11706 in LNCS, Springer, Berlin, Heidelberg (2019)
10. Draheim, D.: Future perspectives of association rule mining based on partial conditionalizationin (DEXA'2019 keynote). In: Proceedings of DEXA'2019 - the 30th International Conference on Database and Expert Systems Applications, LNCS 11706, (2019)
11. Ewald, H., Ioannidis, J.P., Ladanie, A., Mc Cord, K., Bucher, H.C., Hemkens, L.G.: Nonrandomized studies using causal-modeling may give different answers than rcts: a meta-epidemiological study. Journal of Clinical Epidemiology **118**, 29–41 (2020). https://doi.org/https://doi.org/10.1016/j.jclinepi.2019.10.012
12. Grimes, D.A., Schulz, K.F.: Bias and causal associations in observational research. The Lancet **359**(9302), 248–252 (2002)
13. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann Publishers (2001)
14. Jager, K., Zoccali, C., MacLeod, A., Dekker, F.: Confounding: What it is and how to deal with it. Kidney International **73**(3), 256–260 (2008). https://doi.org/10.1038/sj.ki.5002650
15. Kaushik, M., Sharma, R., Peious, S.A., Draheim, D.: Impact-driven discretization of numerical factors: Case of two- and three-partitioning. In: Srirama, S.N., Lin, J.C.W., Bhatnagar, R., Agarwal, S., Reddy, P.K. (eds.) Big Data Analytics. pp. 244–260. Springer International Publishing, Cham (2021)
16. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: A systematic assessment of numerical association rule mining methods. SN Computer Science **2**(5), 1–13 (2021)
17. Kleinbaum, D.G., Sullivan, K.M., Barker, N.D.: A Pocket Guide to Epidemiology. Springer, New York (2006)
18. Lee, P.H.: Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? J Epidemiol **24**(2), 161–167 (Dec 2013)
19. Pearl, J.: Causal inference in statistics: An overview. Statistics Surveys **3**, 96 – 146 (2009). https://doi.org/10.1214/09-SS057
20. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
21. Robins, J.M., Greenland, S.: Causal inference without counterfactuals: comment. Journal of the American Statistical Association **95**(450), 431–435 (2000)
22. Rothman, K.J.: Modern Epidemiology. Lippincott Williams and Wilkins, 4 edn. (2021)
23. Schuster, N.A., Twisk, J.W., Ter Riet, G., Heymans, M.W., Rijnhart, J.J.: Noncollapsibility and its role in quantifying confounding bias in logistic regression. BMC Medical Research Methodology, doi:10.1186/s12874-021-01316-8 **21**, 1–9 (2021)
24. Sharma, R., Kaushik, M., Peious, S.A., Bazin, A., Shah, S.A., Fister, I., Yahia, S.B., Draheim, D.: A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. IEEE Access **10**, 12792–12813 (2022). https://doi.org/10.1109/ACCESS.2022.3142537
25. Sharma, R., Kaushik, M., Peious, S.A., Shahin, M., Yadav, A.S., Draheim, D.: Towards unification of statistical reasoning, olap and association rule mining: Semantics and pragmatics. In: Bhattacharya, A., Lee Mong Li, J., Agrawal, D., Reddy, P.K., Mohania, M., Mondal, A., Goyal, V., Uday Kiran, R. (eds.) Database Systems for Advanced Applications. pp. 596–603. Springer International Publishing, Cham (2022)

26. Sharma, R., Kaushik, M., Peious, S.A., Yahia, S.B., Draheim, D.: Expected vs. unexpected: Selecting right measures of interestingness. In: Song, M., Song, I.Y., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Big Data Analytics and Knowledge Discovery. pp. 38–47. Springer International Publishing, Cham (2020)
27. Snowden, J.M., Rose, S., Mortimer, K.M.: Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. American journal of epidemiology **173**(7), 731–738 (2011)
28. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. ACM SIGMOD Record **25**(2), 1–12 (Jun 1996). https://doi.org/10.1145/235968.233311
29. Suarez, D., Borràs, R., Basagaña, X.: Differences between marginal structural models and conventional models in their exposure effect estimates: a systematic review. Epidemiology **22**(4), 586–588 (2011)
30. UN General Assembly: Transforming Our World: the 2030 Agenda for Sustainable Development. Resolution A/RES/70/1. United Nations (2015)
31. Vandenbroucke, J.P., Broadbent, A., Pearce, N.: Causality and causal inference in epidemiology: the need for a pluralistic approach. International journal of epidemiology **45**(6), 1776–1786 (2016)
32. Weiss, N.S.: Clinical Epidemiology: The Study of the Outcome of Illness, Monographs in Epidemiology and Biostatistics, vol. 36. Oxford University Press (2006)

Appendix III

III

S. Arakkal Peious, R. Sharma, M. Kaushik, S. Mahtab, and D. Draheim. On observing patterns of correlations during drill-down. In P. D. Haghighi, E. Pardede, G. Dobbie, V. Yogarajan, N. A. Sanjaya, G. Kotsis, and I. Khalil, editors, *Proceedings of iiWAS'2023 – the 25th International Conference on Information Integration and Web-based Applications and Services*, volume 14416 of *Lecture Notes in Computer Science*, pages 134–143, Cham, 2023. Springer.
doi:10.1007/978-3-031-48316-5_16

On Observing Patterns of Correlations During Drill-Down

Sijo Arakkal Peious[1 [0000−0002−7858−9463]], Rahul Sharma[1 [0000−0002−9024−8768]], Minakshi Kaushik[1 [0000−0002−6658−1712]], Mahtab Shahin[1 [0000−0002−5784−6301]], and Dirk Draheim[1 [0000−0003−3376−7489]]

Information Systems Group, Tallinn University of Technology, Estonia
{sijo.arakkal,rahul.sharma,minakshi.kaushik,
mahtab.shahin,dirk.draheim}@taltech.ee

Abstract. Drill-down is a natural and extensive data analysis method that is widely used to analyse aggregate values of data at different levels of granularity. As such, drill-down has proven as an essential tool for informed decision-making in various scenarios. In this paper, we argue that drill-down can be equally utilised to analyse the behaviour of data patterns at various levels. To evaluate the usefulness of such an approach, we investigate the behaviour of Pearson correlation at different drill-down levels in the well-known meteorological data set CMDC. We test the hypothesis that the Pearson correlation between various attributes is preserved during drill-down and provide a systematic discussion of the outcome of these tests.

Keywords: Pearson correlation · drill-down.

1 Introduction

Decision-makers need to do frequent data analysis to generate proper decisions. Going through a small amount of data is acceptable. However, there is a drastic change in the volume of the data produced every day [8]. It is undoubtedly good for decision-makers to understand all the available data before making any decision [14]. Due to the inability to go through each layer, decision-makers analyse the outermost margin or drill-down until they feel comfortable making the decision. They assume that the rest of the data follows the same pattern, which might result in the wrong conclusion.

Accessing the deeper layers of an organised dataset or a file structure is known as drill-down. Drill-down allows the decision-makers to analyse the granular layers of the dataset by combining more constraints. Drill-down can provide insights into each layer of data to decision-makers other than a marginal [16, 10]. For example, a drill-down report which shows the salary distribution of each state in a country will also provide the capability to view and compare the salary distribution of each province or profession in the state. The drill-down method allows decision-makers to go deeper into each level for in-depth knowledge of each layer and compare it with each other. In the drill-down salary distribution report,

decision-makers can compare the salaries in different provinces to identify which is the high-paying profession and high-paying province for the same profession. Seeing data from a different point of view will give a different perspective about the data; drill-down allows decision-makers to analyse the same data from a different point of view and compare it with different layers of results to make a better understating of data.

The rest of the paper is structured in the following manner: Section 2 delves into the issue or challenge being addressed in this paper. Then, Section 3 explains the data preprocessing method which is used in this study. In Section 4, we answer the question we raised in Section 2. Finally, Section 5 offers a concise overview of the findings from this study before proceeding to conclude the paper.

2 Problem Statement

In this paper, we calculate the bivariate Pearson correlation in each level and compare it with the outer marginal level. In bivariate analysis, the relationship of two variables is studied simultaneously [6, 12]. Pearson correlation specifies the existence of a correlation between two variables and discovers the magnitude of the correlation between them. This method is commonly used for numerical variables [11]. A correlation shows the influence of one variable on another, but the actual causality might be in a different direction than we assume, so the correlation would not indicate causation. The correlation values vary from 1 (strong positive correlation) to -1 (strong negative correlation). The correlation value of uncorrelated variables will be 0 [2]. This correlation value shows, how a variable will behave when an increase or decrease happens to the other variable [15]. According to David(1938), the recommended sample size for calculating Pearson correlation is greater than or equal to 25 ($n \geq 25$) [3]. So, we can consider this number(25) as our minimum threshold for this study. The equation for the Person correlation of two variables is the sum of the covariance of variables divided by the sum of the square root of covariance [1].

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

In recent times, individuals, particularly politicians and the media, often make inaccurate assertions regarding the causality of misleading or inappropriate correlations. These claims have a significant impact on decision-making. As we explained in the previous paragraph, correlation is not always the answer to causality. It does not mean that the correlated variable has a causal impact on each other. A vast volume of data is created on a daily basis, necessitating decision-makers to thoroughly review the data for informed decision-making. To alleviate this workload, they may opt to review the data at a superficial level or discretize the values. Selecting pertinent information and suitable features from the dataset has historically posed a challenge for decision-makers [14]. The system's performance has a vital role in finding these features and information [7].

Choosing unsuitable features will mislead the decision-makers, resulting in incorrect conclusions or confusion regarding the variables' impact.

Utilizing a drill-down approach enhances decision-makers' understanding of the data. In this study, we integrate drill-down capabilities with Pearson correlation to gain a deeper understanding of the data. This combined approach aids decision-makers in reaching more informed conclusions. To help the decision-makers, we have created a tool(Grandreport [5][1]) with multiple data mining techniques [13] and ACIF generator function [17]. The Grandreport will produce numerous rows of report. This is due to the absence of constraints on support and confidence, enabling generalized association rule mining. Analysts utilize this approach to integrate every line of the report into their decision-making process [5]. In Grandreport, we have integrated association rule mining, Pearson correlation and regression to improve the decision-making process. In association rule mining, the target columns are discretized for numeric values to facilitate the mining process [4]. In the Grandreport, we utilize values in their original form to achieve improved results. This stands as a primary advantage of our system. The main disadvantage of this tool is that it reports all possible combinations of influencing factors and generates a lengthy report. To tackle this problem, we decided to report only the exciting and valuable factors by analysing the output with the measures provided by integrated data mining techniques before showing it to the user. From this process, we noticed that some of the variables with high correlation in the marginal level are not showing correlation during the drill-down, and most decision-makers are not considering this pattern.

In this paper, we are trying to answer,

- Does the correlation shown in the marginal for a variable follow the same pattern during drill-down, or will it behave differently?

Our hypothesis is that the correlation observed at the marginal level for a variable will maintain a consistent pattern during drill-down.

3 Experimental Study

Within this research, we examine the patterns present at the marginal layer in comparison to each subsequent drill-down layer within a real-world dataset. Subsequently, we discuss our findings regarding the patterns observed in each dataset.

3.1 CMDC Dataset

For this study, Meteorological Data in China [9] (CMDC [2]) is used. The establishment of this portal aimed to facilitate the sharing of daily meteorological data,

[1] http://grandreport.me/
[2] http://data.cma.cn/en

ultimately advancing science and technology. The CMDC dataset encompasses various meteorological attributes, including precipitation amount, sea level pressure, snow depth, temperature, visibility, wind speed, and more. Each attribute comprises one year's worth of data from 31 provinces.

3.2 Preprocessing

Pearson correlation involves a bivariate analysis and is most effective with numerical values. To enhance performance and reduce computation time, we refined the dataset by generating distinct tables for various attributes. Each column represents the mean value of a specific attribute for each day within a particular province(table 1). The comparison between the provinces will generate an overview of climate differences in each province.

Table 1: CMDC dataset separated by province.

Date	Beijing	Tianjin	Hebei	Shanxi	Neimenggu	Liaoning	Jilin	Heilongjiang
01-Jan	3.300	2.114	2.386	3.160	3.103	3.272	3.001	8.108
02-Jan	4.000	2.914	2.844	3.004	3.442	3.028	4.059	6.415
03-Jan	3.100	2.084	2.986	3.214	3.569	3.069	3.251	5.038
04-Jan	3.600	3.330	3.362	3.661	4.332	2.673	3.095	4.738
05-Jan	2.700	3.951	3.663	3.144	3.926	2.988	2.730	4.100
06-Jan	5.900	4.711	3.747	3.158	4.867	3.943	2.959	3.723
07-Jan	10.800	8.600	5.164	3.800	4.293	8.369	3.475	3.923
08-Jan	4.300	5.597	3.536	3.504	3.912	4.142	3.458	5.008
09-Jan	3.500	2.654	2.767	3.405	2.884	2.501	3.112	4.038
10-Jan	10.900	6.989	4.223	3.780	4.126	3.429	2.972	3.592

3.3 Setup

The objective of this study is to determine the correlation between variables and compare the Pearson correlation value of each drill-down layer against the marginal level. sing the above-described dataset (CMDC), we computed the Pearson correlation for each drill-down and compared it with the province's marginal layer. To ensure accuracy, we omitted correlation results with a count of less than 25 [3]. Put simply, if a month is absent or the number of values for a specific month is below 25 in any province, those months are excluded from the comparison. We developed a Python command line application to calculate correlation. To ensure accuracy and efficiency, we utilized the 'corrcoef' function from the 'NumPy' library.

As an example of the output generated by the application, see Table 3, which shows the correlations of the provinces Tianjin and Hunan for various months.

4 Evaluation

As previously mentioned, utilizing the drill-down approach enhances our understanding of the data. The outcomes confirm that drill-down helps to generate a deeper understanding of the data. Our experimentation with the CMDC dataset to verify the hypothesis, revealing diverse patterns across the dataset. The results shown here are for different meteorological attributes of Tianjin province. In this section, we address the questions posed earlier in this study.

- Whether the correlation shown in the marginal for a variable will follow the same pattern during drill-down or will it behave differently?

To answer this, we can compare the Figs. 1, 2, 3, 4 and 5. Our investigation revealed that nearly all combinations exhibit a consistent pattern with marginal level, with a few exceptions. The result shown in the Figs. 1, 2, 3, 4 and 5 can be divided into three groups.

1, The correlation patterns of drill-down have the same pattern as the marginal.
2, The correlation patterns of the majority drill-down are not the same as the marginal.
3, The correlation patterns of the drill-down and marginal behave in opposite directions.

The 'X' and 'Y' axes represent the correlation and number of elements.

Figures 1 and 2 show the same pattern for the marginal and drill-down. In this, the marginal shows a weak correlation, and the drill-down also shows the same pattern for most variables. However, a handful of variables (less than 5%) show strong correlations. Around 65% of attributes in the CMDC dataset follow this pattern.

Figures 3 and 4 show different patterns for drill-down and marginal. Here, the marginal will have a strong positive or negative correlation, while drill-down, we are getting different correlations. A strong positive correlation is generated for the marginal in Figs. 3 and 4. However, drill-down shows a weak or moderate correlation for the same dataset. Some exceptional variables(less than 10%) also show strong positive correlations. 25% of the attributes in the CMDC dataset follow this pattern.

The drill-down and marginal correlation patterns behave in opposite directions in Fig. 5. A negative correlation is generated for the marginal, while drill-down, strong or moderate positive correlations are generated for most of the variables. More than 60% of the variables behave in opposite directions. Figure 5 exemplifies the importance of drill-down techniques in data analysis. Only 10% of CMDC dataset attributes exhibit this pattern.

This result shows that it is always recommended to delve deep into the dataset to find the actual pattern before making any conclusions. In some cases, the correlation between the marginal and drill-down behaves differently (like Fig. 5). Moreover, the existence of the statistical paradoxes and other data fallacies (like confounding effects) have a significant impact on the outcome.

Table 2: Outer marginal correlations between (a) the mean temperature of 'Tianjin' and the mean temperature of various other provinces, and (b) the weighted mean sea level pressure of 'Tianjin' and the weighted mean sea level pressure of other provinces.

Province	(a) mean temperature correlation with 'Tianjin'	(b) sea level pressure correlation with 'Tianjin '
Anhui	0.9311	-0.4885
Fujian	0.8471	0.0413
Hebei	0.9897	-0.4860
Heilongjiang	0.9672	-0.4951
Jiangsu	0.9191	0.1049
Jiangxi	0.8910	-0.6206
Liaoning	0.9768	-0.0667
Neimenggu	0.9729	0.0736
Shanghai	0.8863	-0.8132
Shandong	0.9696	-0.1986
Zhejiang	0.8858	-0.7389

Table 3: Drill-down correlations between (a) the mean temperature of 'Tianjin' and 'Hunan', and (b) the weighted mean sea level pressure of 'Tianjin' and 'Hunan'.

Layer	(a) mean temperature correlation between 'Tianjin' and 'Hunan'	(b) sea level pressure correlation between 'Tianjin' and 'Hunan'
Year (Marginal)	*0.9018*	*0.0086*
January	0.2872	0.2913
February	0.6760	0.5680
March	0.4209	0.8667
April	0.5133	0.5599
May	-0.0876	0.0572
June	0.1998	0.3018
July	0.3676	0.9029
August	0.4704	-0.9963
September	-0.7657	0.8607
October	0.0351	0.7164
November	0.8734	0.8754
December	0.2183	0.9346

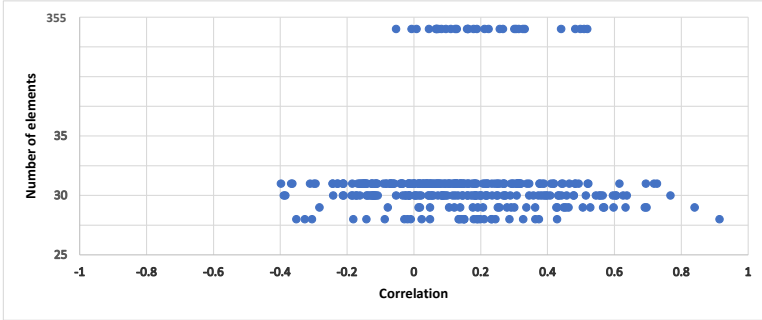


Fig. 1: Correlation plot for inverse-distance weighted maximum wind gust.

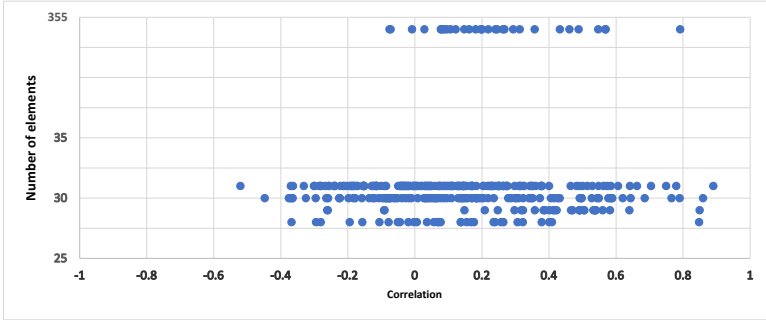


Fig. 2: Correlation plot for inverse-distance weighted maximum sustained wind speed.

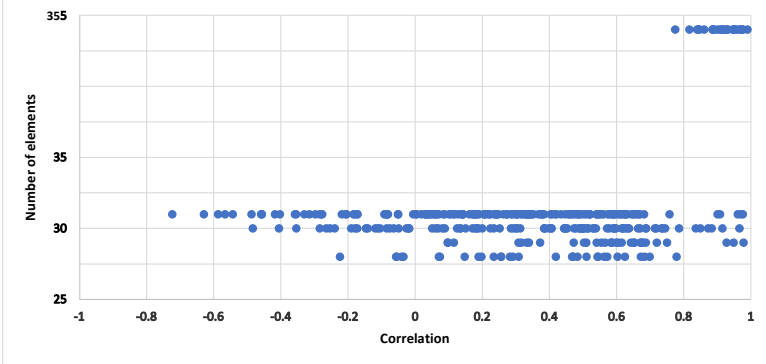


Fig. 3: Correlation plot for inverse-distance weighted mean temperature.

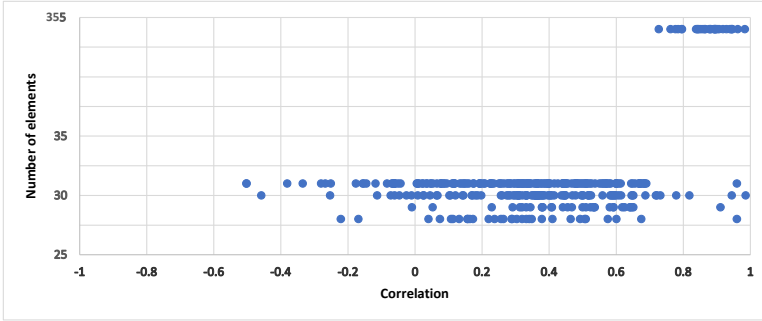


Fig. 4: Correlation plot for inverse-distance weighted mean dew point.

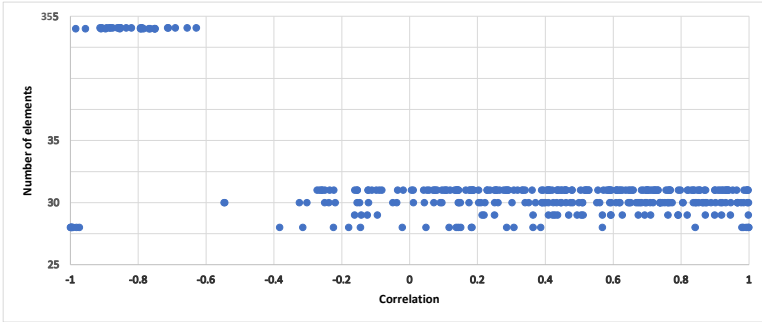


Fig. 5: Correlation plot for inverse-distance weighted mean sea level pressure.

5 Conclusion

The overall aim of this work is to identify the importance of the drill-down approach in data analysis. Given the importance of data analysis in decision-making, we wanted to understand the data more deeply and identify the patterns of each layer with the impact on the outcome. We calculated the correlation of each layer and compared it with the marginal correlation. This approach aims to identify the different patterns in drill-down and marginal. The results show three different patterns for CMDC data in this work.

The main confrontation faced in this work was to identify the proper statistical method to find the pattern to such scenarios. The experiment is carried out only on a small scale, with a limited number of meteorological attributes of CMDC data for one year. So, the presented results deliver only a general view of the importance of drill-down analysis. Further changes and more statistical analytical methods are required to enhance the results. For now, we are only using the correlation method to identify the patterns; however, as a next step, we would like to add regression to generate different perspectives and find the confounding effect of each attribute. Also, as part of future work, we would like to investigate on a larger scale(with hundreds of real-world datasets) to understand the patterns of drill-down in different types of datasets.

6 Acknowledgements

This work has been conducted in the project “ICT programme" which was supported by the European Union through the European Social Fund.

References

1. Berman, J.J.: Chapter 4 - understanding your data. In: Berman, J.J. (ed.) Data Simplification, pp. 135–187. Morgan Kaufmann, Boston (2016). https://doi.org/https://doi.org/10.1016/B978-0-12-803781-2.00004-7, https://www.sciencedirect.com/science/article/pii/B9780128037812000047
2. Berman, J.J.: 11 - indispensable tips for fast and simple big data analysis. In: Berman, J.J. (ed.) Principles and Practice of Big Data (Second Edition), pp. 231–257. Academic Press, second edition edn. (2018). https://doi.org/https://doi.org/10.1016/B978-0-12-815609-4.00011-X, https://www.sciencedirect.com/science/article/pii/B978012815609400011X
3. David, F.N.: Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples. Cambridge Univ. Press (1938)
4. Draheim, D.: Generalized Jeffrey Conditionalization: A Frequentist Semantics of Partial Conditionalization. Springer (2017)
5. Draheim, D.: Future perspectives of association rule mining based on partial conditionalizationin (DEXA’2019 keynote). In: Proceedings of DEXA’2019 - the 30th International Conference on Database and Expert Systems Applications, LNCS 11706, (2019)

6. Field, A.: Discovering statistics using IBM SPSS statistics. sage (2013)
7. Gavel, S., Raghuvanshi, A.S., Tiwari, S.: Maximum correlation based mutual information scheme for intrusion detection in the data networks. Expert Systems with Applications **189**, 116089 (2022). https://doi.org/https://doi.org/10.1016/j.eswa.2021.116089, https://www.sciencedirect.com/science/article/pii/S095741742101424X
8. Han, J., Kamber, M.: Data mining concepts and techniques. San Francisco, CA pp. 335–391 (2001)
9. Lab, C.D.: Meteorological Data (2020). https://doi.org/10.7910/DVN/TU0JDP, https://doi.org/10.7910/DVN/TU0JDP
10. Morar, N., Baber, C., McCabe, F., Starke, S.D., Skarbovsky, I., Artikis, A., Correai, I.: Drilling into dashboards: Responding to computer recommendation in fraud analysis. IEEE Transactions on Human-Machine Systems **49**(6), 633–641 (2019). https://doi.org/10.1109/THMS.2019.2925619
11. Nettleton, D.: Chapter 6 - selection of variables and factor derivation. In: Nettleton, D. (ed.) Commercial Data Mining, pp. 79–104. Morgan Kaufmann, Boston (2014). https://doi.org/https://doi.org/10.1016/B978-0-12-416602-8.00006-6, https://www.sciencedirect.com/science/article/pii/B9780124166028000066
12. Omilion-Hodges, L.: Sage Encyclopedia of Communication Research Methods, pp. 855–858. SAGE publications (01 2017). https://doi.org/10.4135/9781483381411.n293
13. Peious, S.A., Sharma, R., Kaushik, M., Shah, S.A., Yahia, S.B.: Grand reports: A tool for generalizing association rule mining to numeric target values. In: Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery. Lecture Notes in Computer Science, vol. 12393, pp. 28–37. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59065-9_3
14. Peious, S.A., Suran, S., Pattanaik, V., Draheim, D.: Enabling Sensemaking and Trust in Communities: An Organizational Perspective, p. 95–103. Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3487664.3487678
15. Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., Sabeti, P.: Detecting novel associations in large data sets. Science (New York, N.Y.) **334**, 1518–24 (12 2011). https://doi.org/10.1126/science.1205438
16. Shabaninejad, S., Khosravi, H., Indulska, M., Bakharia, A., Isaias, P.: Automated insightful drill-down recommendations for learning analytics dashboards. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. p. 41–46. LAK '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3375462.3375539, https://doi.org/10.1145/3375462.3375539
17. Sharma, R., Kaushik, M., Peious, S.A., Bazin, A., Shah, S.A., Fister, I., Yahia, S.B., Draheim, D.: A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. IEEE Access **10**, 12792–12813 (2022). https://doi.org/10.1109/ACCESS.2022.3142537

Appendix IV

IV

S. Arakkal Peious, S. Suran, V. Pattanaik, and D. Draheim. Enabling sensemaking and trust in communities: An organizational perspective. In E. Pardede, M. Indrawan-Santiago, P. D. Haghighi, M. Steinbauer, I. Khalil, and G. Kotsis, editors, *Proceedings of iiWAS'2021 – the 23rd International Conference on Information Integration and Web Intelligence*, pages 95–103. Association for Computing Machinery, 2021. doi:10.1145/3487664.3487678

Enabling Sensemaking and Trust in Communities: An Organizational Perspective

Sijo Arakkal Peious
Information Systems Group,
Tallinn University of Technology
Tallinn, Estonia
sijo.arakkal@taltech.ee

Shweta Suran
Information Systems Group,
Tallinn University of Technology
Tallinn, Estonia
shweta@taltech.ee

Vishwajeet Pattanaik
Information Systems Group,
Tallinn University of Technology
Tallinn, Estonia
vishwajeet.pattanaik@taltech.ee

Dirk Draheim
Information Systems Group,
Tallinn University of Technology
Tallinn, Estonia
dirk.draheim@taltech.ee

ABSTRACT

The large volume of information being produced in organizations today poses new challenges to the accuracy and effectiveness of any organizations' decision-making processes. These challenges, namely sensemaking and trust, can critically impact the decision-making processes, even if the organizations are relying on business intelligence (BI) strategies. Given the critical impact an organizations' BI can have on its sustainability and thus its success, in this work, we attempt to draw insights from the literature on collective intelligence and, based on these, present a novel artifact that aims to empower organizations' BI by supporting the organizations' employees in establishing trust and sense when working up with new ideas and solutions. The proposed artifact utilizes a novel reputation model, which calculates reputation based on an individual's area of expertise and reputation score, in order to assist in establishing trust among system users, and thus helps improve decision-making processes.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Software and its engineering** → *Development frameworks and environments*; *Use cases*; Abstraction, modeling and modularity; Designing software.

KEYWORDS

business intelligence, collective intelligence, crowdsourcing, sensemaking, trust, reputation model

ACM Reference Format:
Sijo Arakkal Peious, Shweta Suran, Vishwajeet Pattanaik, and Dirk Draheim. 2021. Enabling Sensemaking and Trust in Communities: An Organizational Perspective. In *The 23rd International Conference on Information Integration and Web Intelligence (iiWAS2021), November 29-December 1, 2021, Linz, Austria.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487664.3487678

iiWAS2021, November 29-December 1, 2021, Linz, Austria

ACM ISBN 978-1-4503-9556-4/21/11…$15.00
https://doi.org/10.1145/3487664.3487678

1 INTRODUCTION

Today's business organizations face endless instabilities and volatilities, which can lead to creation of massive volumes of data; being produced by organizations both internally and externally [22]. To harness the possibilities of this transformation, several organizations now aspire (but often even struggle) to convert these large volumes of existing data into a clear understandable chunks that could be utilized in their business processes. In order to achieve this businesses reply on Business Intelligence (BI); a strategy that enables organizations to examine their past actions and decisions, and thus consequently, predict the future. BI denotes a wide range of technologies, processes and applications that assist organizations in gathering, storing, evaluating, and granting access to data for refining business's processes and over-all decision-making [17, 39]. It aids organizations by continuously collecting and analyzing organizational information (including performance metrics) and assists by making the decision-making processes more efficient.

Although BI is a powerful tool and can be typically used in an organization's almost all decision-making processes (both long-term and short-term), however, business organizations today only use BI for day-to-day (i.e., short-term) decision-making [20] and, presently, BI abilities are not necessarily utilized for identifying the organizations' long-term progression, which could indeed help them in improving their methods when undertaking tactical decisions [8]. Another problem that can arise when using BI (which is also often discussed in literature) is sensemaking [36]; this is a key precondition to reach an informed decision and is based on the prior actions of humans [3]. This is to say, that given BI relies on both machine intelligence and human intelligence, when assisting organizations in decision-making; the humans involved in analysis tasks can often get confused by the lack of sense in an idea or an outcome.

Now given that by gaining a better 'sense' of the organization overall, managers (and other decision makers) could better understand their business's organizational environment and hence make healthier decisions [35]; BI applications and related strategies can

play a critical part in sensible decision making, and even added advantages beyond conventional decision-making. It is key to note here that, the decisions that are made using BI should be both sensible and explainable and should cover various potential possibilities.

In BI, data/information is used to create reports, summarize past actions, forecast actions, and to understand current and future risks. When relying on BI, the precision of predictions (made using BI strategies) depends on the quality of the information and its sources [40]; if the information and it's source are not trustable, the entire action and its outputs can become futile (even counter-productive). Managers and decision makers who use these outputs typically understand on-going scenarios, and hence create productive decisions or implement their decisions keeping the scenarios in mind [10]; however, a key factor that can influence the decision-making process in such scenarios is 'trust'; specially since humans are involved in the process. Consider this for example, if managers from distinct departments/sections of an organization are working together on creating a solution for given scenario, individuals who have encountered similar scenarios before might be able to contribute more to the solution, however, if the managers are not aware of the past experiences of their colleagues, they might end up not considering ideas of the individual that could contribute the most. This is in line with literature, where researchers have found that, when working together in groups humans tend to make better decisions when there is trust among group-members [30].

That said, in this work we attempt to tackle the above mentioned issues of sensemaking and trust, and propose a novel platform designed as discussion forum oriented towards managers, decision-makers and other employees working in organizations. To achieve this, we draw influence from another domain (one that dates back to Aristotle), that is, Collective Intelligence (CI, defined as, "groups of individuals acting collectively in ways that seem intelligent" [21]); as the domain has recently gained traction in a wide variety of domains [30]. So much so, that it is actively being used by both governing bodies and organizations today; not only to collect citizen/end-user feedback, but also in the design processes for solving critical issues and developing new products, respectively (for example, in Crowd4Roads and CAPSELLA [31], and openIDEO and Threadless [30]). In general, through its fundamental concepts of collection and collaboration, CI has allowed organizations to make better use of the (collective) intelligence of their employees and their users, and thus helps enhance their decision-making processes, when gathering information from numerous sources and creating valuable outputs using CI methods.

With this in mind, the overall aim of this study, is to discover how BI strategies could contribute to better decision-making in presence of sensemaking and trust. The study mainly focuses on the organizations' employee's perspective and tries to identify factors that generate trust between employees and attempts to understand how this trust helps in sensible decision-making processes. In particular, we would like to answer to answer the following research questions:

Q1: *Can we solve the issues of trust and sensemaking in BI using the concepts from CI?*

Q2: *How can we design a reputation model for such a BI system while solving well-known challenges related to reputation in CI platforms?*

The remaining paper is organized as follows, in Section 2, background and related work of BI, CI, trust and reputation systems are described. Then, Section delves into the novel reputation model of trust and sensemaking, and Section discusses the proposed artifact (i.e., the CI-Forum). In Section 5 we describe the evaluation process for the developed forum and reputation model; and finally, Section provides a brief discussion on the findings of this work before concluding the paper.

2 BACKGROUND AND RELATED WORK

Business organizations' performance relies on real-time and effective organizational information. BI systems analyze this information and identify shortcomings and problems within an organization, they provide businesses with insights and suggestions in real-time and support decision-makers in coming up with better conclusions; which subsequently helps organizations sustain and improve productivity [2, 34]. By implementing innovative ideas and new technologies in the their processes, businesses can achieve competitive advantage and success in rapidly changing business conditions [13, 23].

2.1 Business Intelligence

The decision-making processes change according to the information businesses are using to make decisions within their organizations [18]. We can characterize a BI system as a framework that collects, makes modification and generates business's organizational information from different resources. This reduces the time required for analyzing important business information and helps managers to make efficient decisions that can be utilized to improve business strategies. BI is the process of combining different series of actions and business information to provide a competitive advantage to business organizations by helping decision-makers [26]. It is a system and generates answers to support decision-makers to understand the economic situations of the business organizations [24]. Conventionally, BI uses methodological models and numerical functionalities for analysis, used for mining valuable business information and data from basic information to help managers and decision-makers [32]. These business information mining processes and analysis procedures enhance forecasting and help decision-makers understand the progression and problems of any business organization [27].

2.2 Collective Intelligence and Crowdsourcing

General intelligence, as understood by psychologists is the (single) statistical factor that predicts variance in performance, when an individual performs some cognitive tasks (e.g., [11]); it includes an individuals capacity for logic, understanding, learning, reasoning, planning, creativity, critical thinking, problem-solving and many other aspects. When a group of individuals (human or machine) work together and use their individual intelligence, the aggregated intelligence of the group can be understood as CI.

In Information and Communications Technologies (ICT), CI has several definitions (for example, the most prominent ones are by Levy [11] and Malone [21]); each defines CI as having, three components: "individuals (with data/information/knowledge), coordination and collaboration activities (according to a predefined set of rules), and means/platform for real-time communication (viz.,

hardware/software)"–together these "enable intelligent behavior in groups or crowds" [30]. That said, that advent of the Internet has allowed for mobilization and harnessing of CI in truly novel ways, and this has enabled creation of web-based group discussion platforms that play a key role decision making today [29]. This has opened the gates to a wide variety of emerging research topics, including for example, research where scientists and academicians are trying to understand the influence of group discussion platforms on performance improvement in the quality, efficiency, and effectiveness of decision-making when using such platforms [25]. Some researchers are also focusing on how users behave, group members carry out activities, and knowledge that is generated on group discussion platform by a user and their groups. There also have been studies which focuses on collaborative IT solutions and group discussion systems (designed as web-based platforms), and aim to explain how BI is being used in business organizational context [8].

Another application of CI, that is gaining tremendous interest in research is crowdsourcing (defined as, process where a group of people work together and carry out a task, typically involving collection of data/information or building a solution; that was conventionally done by a single individual [7]. CI (also, crowdsourcing) involves group of people working together, but its key that the individual members of group are diverse [30]. Some researchers have expressed that the main aim of crowdsourcing is to distribute the task of one person to a group of people, and by doing this the overall workload can be divided and hence the task can be carried out effortlessly [6]. Such crowdsourcing activities are divided into three categories. First, directed crowdsourcing, where, a coordinator asks a specific question (with relevant explanation) oriented towards participants, and participants earn some kind of rewards or benefits to the effort and time they contribute. The second category is self-directed, where, participants contribute due to intrinsic motivations. Here participants comes to a common platform and discuss various topics according to their volition and try to come up with decisions or actions according based on the topic at hand. The third and final category is passive, where crowdsourcing is only a side effect of output produced by some action. Here, participant are not obliged to generate the output, or might not be even aware that are participating in a crowdsourced system [37].

A first popular example of crowdsourcing that has is often discussed in literature is the Goldcorp Inc.'s initiative from the year 2000, where they used crowdsourcing to identify gold mines in the Red Lake. The participants were awarded around 0.5 million, and Goldcorp agreed to share the information about the gold mines if they were able to find 6 million ounces of gold, from an identified site. Geologists and engineers from various counties started analysing the information provided by Goldcorp and the company started to receive replies (i.e., potential sites with gold) in a short amount of time. The results produced by participants were verified by a panel decided by Goldcorp, and the end of the competition the panel members were surprised by the both the creativity of the participants and the results produced by them. Goldcorp drilled at the best 5 locations suggested by participants, and found gold from at each of the locations. A key finding of the competition was that participants were able to find gold from all of these locations, without even the locations once. The competition also illustrated how intelligent individuals are, and that by utilizing humans (collective) intelligence combined with technology, organizations could come up with novel and innovative solutions (which could not be achieved conventionally) [5, 38].

2.3 Trust

Reputation and trust are considered key factors of a civilized society [12]. In CI systems too, trust is considered a key property [14, 30]. The success rate of a CI platform can be judged by measuring the trust and openness among the users [4, 14]. An easy method to assess the trustworthiness can be to just ask the users if they trust the source of information [33]. Dworken et al. [15] explained how trust perceived by organizations using examples from the news industry. They claimed that news coverage over the years has changed dramatically, and this is because users have started to analyze both the news and its source to check the reliability of the information [15]. Trust is also a key component in decision making as well as in collaborative working environments [12]. Trust is the belief that the trusted person or the organization will accomplish a particular task according to the task givers expectation [16]. BI applications provide trustable descriptions of various business situations and deliver numerous outcomes for understanding business organizational risks; whoever, as we eluded to earlier, even with the trustable nature of BI applications, trust and sensemaking still remain a challenge to some extent.

2.4 Reputation Systems

Reputation systems are mathematical functions used to calculate a user or objects trustworthiness or value as perceived by fellow users, and is calculated based on user feedback (which can represented using up-votes, stars, like etc.). Theses user score provided by fellow users can be used as a benchmark to identify the level of user trustworthy, and the aggregate of votes and feedback are considered as the reputation score. Literature indicates that theses votes/feedback and thus reputation score can often be violated, thus providing untruthful feedback to gain reputation (supporting non-worthy users) or to decrease the reputation of other users [28]. Reputation systems also face numerous other challenges [1], for instance, Sybil attacks, where attackers (or malicious users) create multiple fake accounts to up-vote their contributions in order to gain higher reputation score, or excessive use of self-promotion, or users with high negative reputations tend to delete old accounts and create new ones (this is referred to as whitewashing). A solution to whitewashing however is that the time duration it takes for an individual to gain reputation can be studied (as was done in [19]), as true good reputation typically only grows gradually. Another challenge to reputation systems is oscillation attack, where, the attacker creates a user account and behaves fairly to achieve good reputation, and then changes their behaviour, hence misleading noble users who trusting the reputation of the attacker [9]. This these challenges in mind, in this work, we aim to develop a novel reputation model that would attempt to tackle some of the challenges described above.

To summarize, this section presented a brief background of literature of BI, CI and reputation systems; this is critical as the review of the literature allowed us to illustrate the purpose of the study, the questions, limits and advantages. It also provides theoretical

viewpoints, current views for identifying the study questions and a review of related experimental studies concerning the respective fields. The following section explains the proposed reputation model and how it is different from existing models and systems.

3 NOVEL REPUTATION MODEL

The study proposes a novel approach to the reputation system which aims to avoid the problems explained in the reputation system's literature review. The proposed approach follows a decentralized reputation system. To some extent, this model is similar to existing distributed reputation models like the one used by 'Stack Overflow'. The users give feedback through positive and negative votes (i.e., up-down votes). Whenever a user receives a vote, their reputation score is altered dynamically according to the received votes. In the proposed method, users would not getting the same score every time they receive an up/down vote; instead the amount of score that would added or reduced would depend on the reputation score of the user giving the vote. This means, if a user has a high reputation and they they give an up-vote to another user then the receiving user's reputation score would increase by a higher value, and if the user giving the up-vote does not have any reputation then the receiving user will get the minimal increase in their reputation score. In this approach, the overall score is not calculated while making the vote; but rather the votes are calculated with respect to the category tags (on the individuals profile, i.e., only the topics the user is familiar with) and points are also calculated according to these category tags.

In the proposed reputation model scores are calculated based on the category tags. Whenever a user casts their vote, the system first checks for the reputation score of that user according to the category. If the user has a reputation score, then the system divides that score (that will be added to the receivers reputation) using the total number of votes that the user has received for the particular category. If the calculated score is less than the minimum score, then the receiving user receives the minimum score else they receive the calculated score.

Take for example the following scenario, let us assume there are ten users ($A_1, A_2, A_3, \ldots, A_{10}$) and three categories ($C_0, C_1$, and C_2). At time T_0, all users will start with reputation scores of 1.0 (both, overall and for individual categories).

Now lets assume, by time T_1, A_6 has received five up-votes in category C_0 (each from $A_1, A_2, A_3, A_4 and A_5$), and hence A_6's overall and category reputation score (for C_0) will be 6.0.

At time T_2, A_6 gives an up-vote to A_7 in category C_0. So A_7's new reputation score for C_0 will be calculated by dividing A_6's total score in C_0 with total up-votes (as received by A_6 in C_0) and adding to A_7's reputation score for C_0; i.e.,

$(6/5) + 1 = 2.2$

Hence, A_7's overall and category reputation score (for C_0) will be 2.2.

Now at time T_3, A_6 gives an up-vote to A_8 for C_1; since A_6 started with 1.0 for category C_1 and has not received a single up- or down-vote in due time, a score of 1.0 would be added to A_8's category reputation score (for C_1). Hence, A_8's new C_1 category reputation score will be 2.0.

Finally at time T_4, A_9 gives an up-vote to A_6 for category C_2; since A_9 has not received a single up- or down-vote in due time category C_2, a minimum score of 1.0 will be added to A_6's C_2 category reputation score.

At this point, the final reputation scores for users ($A_1, A_2, A_3, \ldots, A_{10}$) will be as follows,

- $A_1, A_2, A_3, A_4, A_5, A_9, A_{10} = 1.0$
- $A_6 = 8.0 (C_0 = 6.0, C_1 = 2.0)$
- $A_7 = 2.2 (C_0 = 2.2)$
- $A_8 = 2.0 (C_1 = 2.0)$

When generating scores for negative votes (i.e., down votes) the exact same strategy is used, but with subtraction is used instead of addition.

To summarise, in this section, a novel reputation model has been described. The main advantage of the proposed reputation system is, that users can identify the expertise of every user by viewing an overall reputation score and separate score based on every category (the individual contributes/has contributed to). Now to validate this reputation model we have created an artifact, which we delve into in Section 4.

4 PROPOSED CI-FORUM

To study how sensemaking and trust can influence on user behaviour, and to evaluate the previously proposed reputation model here we present as discussion forum (named "CI-Forum"). The proposed artifact allows users to post questions and reply to the questions posted by other users. Users can the platform share knowledge and help other users to solve problems. Users can up-vote or down-vote other users comments and feedback, which in turn is used to calculate user reputation. Users can view posts by using filters, for example sorted based on the reputation scores of the user who posted the question/comment; and thus should be able to identify individuals experts (based on the best answers/comments). The primary notion behind the artifact is that such a CI based forum could potentially be used in line with BI strategies, and would allow organizations to use the collective intelligence of their employees when carrying out decision-making processes.

4.1 Coding and Implementation

The user interface for the artifact is designed using HTML, CSS and JavaScript. To send and receive data, AJAX POST method is used. The CI-Forum website communicates with the server and collects information in the form of JSON objects and files. To make the design process easier and to master coding, pages are used. On the server side, C# is used as the main programming language, together with a layered architecture. The application consists of four layers, i.e., a main project layer, a business logic layer, a data access layer, and a business object layer. The main project layer contains the '.aspx' files. The business logic layer provides all of the logical functionalities for the application. The layer works as a linking layer between the data access layer and the main project layer. The data access layer communicates with the business logic layer and collects data from the database. The business object layer contains objects and their values. Oracle 12C is used as the database.

4.2 System Features

The application has almost all functionality required by a question and answer (Q and A) forum. In addition to this, the application also shows overall and separate reputation scores for each category tag. This view helps the users to identify the best answer concerning the keywords and user. The main functionalities of the application are user creation, login, creating posts, viewing posts, viewing a single post with its answers, viewing reputation scores for every user and a user dashboard. The list of posts can be ordered in several ways, e.g., according to the latest posts, most viewed posts, most commented posts, or most favourites posts. The forum also has the feature to search posts by their titles and tags. The posts are listed in the form of a table, and each row consists of titles, contents, main category, and last three participant posts. Additionally,total number of comments to the post, the total number of viewers, date/time when the post was create are also visible to the users. Users can click on each participants name and view their basic information (including the name of the participant, when they joined the platform, overall reputation scores, reputation scores per category and achieved badges). These attributes were chosen so as to provide users with an overall idea of who their co-members are, thereby assisting in establishing a sort of trustworthiness among members of the community.

Users can click on each post, which then opens the post as a separate page. The post page shows users the posted question, their answers, comments, and edit-options for each post. Each post itself contains the contributor's name, the date when the post was created, its description, up and down vote options, its count and on option to mark the post as favourite. In addition to the question post, there are also options to create answers, make edits and add comments to the post. On the same page, users can see the basic information about the contributor by clicking on the contributor's name. To create a new post, users can select the 'Create New Post' option from the provided menus. Under the 'Create New Post' form, user can add the title, main category, subcategory, description and also upload relevant documents. The options to select tags is provided in the main and subcategory fields. Under the subcategory field, user can select multiple categories, as per their convenience.

To reiterate, a key advantage of designed artifact is that users can view the overall and individual reputation of all their co-members. This would helps users identify the best answers/contributions. The application also has the option to give votes to the other users based on the posts/contributions and behaviour. The code for the designed artifact and the associated database files are openly available as a repository on GitHub (https://github.com/ssijopious/CI-Forum). This is done so that the results presented in the work, can be reproduced and built upon by others.

5 EVALUATION

In this section we attempt to answer the questions we raised previously in this work. The first question, how to implement CI methods in the BI platform so as to solve business organization's decision-making problems related to trust and sensemaking in the process of decision making.

As we eluded to earlier, BI systems can help resolve issues and support in the process of business organizational sensemaking and trust, however it there is a need to create crowd-based platforms to make ensure data quality, flexibility and risk management. And maintaining data quality, requires that the source of the data are given higher priority. To make sure the integrity of the source, we can utilize the collective knowledge of humans using crowd-sourcing methods within BI systems. To entrust a source or user, would require time, and trustable users would need to contribute trustworthy information while also cooperating with other users of the system. The continuous interactions of the user would help develops trust in the platform. This accuracy of trust will have a high impact on the business organizational decision-making processes.

To solve the next question, this study proposes a new reputation model to identify the problems of the CI reputation model and support the BI system to make more trust and sensible decisions. To evaluate this artifact quantitative research method is used. A question and answer platform are created to implement this new reputation system (CI-Forum). A target group is selected for testing this platform and making the evaluations. In this evaluation, we tried to identify the target group's general understanding and habits of the reputation model. The target users are software engineers and IT specialists. Most of the participants have experience in using question and answer platform. The target group is from two different countries. To collect the evaluation, a questionnaire is created.

5.1 Experimental Procedure

To evaluate the designed artifact we conducted lab experiments with multiple users. The candidates for the experiments were identified through social media (primarily Facebook), by using snowballing. More that 50 potential candidates were identified and given presentation on how to use the platform. After the presentation, the candidates (i.e., participants or users) were provided the web address of the application (which was hosted online during the experiments). Each participant was asked to create separate user profiles, and were instructed to create multiple posts (questions, answers and comments). After this, the participants were asked to actively use the platform over the next two weeks. It is important to note here that all participants had a background in software development, hence they were asked to use the platform in the daily workflows. At the end of two weeks, more than 75 questions with multiple answers had been posted on the platform.

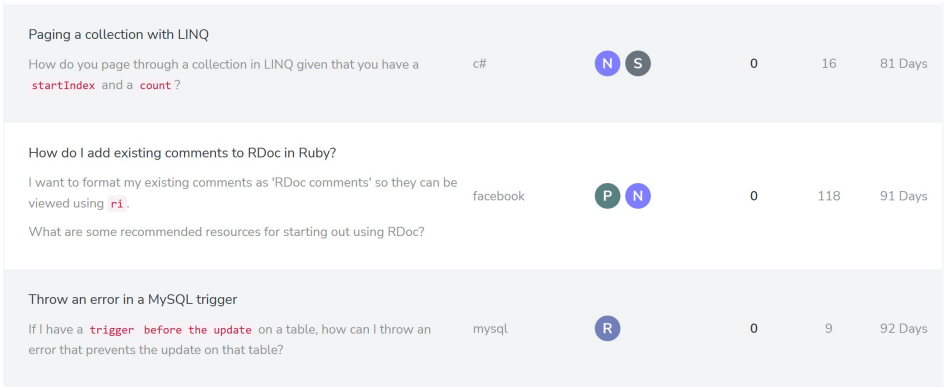
After this, all participants were forwarded survey questionnaires, and were given two days to fill in the same. In total, 68 questionnaires were collected at the end of the experiment. Only 45 valid opinions we found, and hence the remaining were 23 questionnaire responses were rejected.

5.2 Reputation Model

To assess the reputation model, the participants we asked questions related to identification of trustable users. This included three questions (given below), and participants were asked to score the questions through Likert scale ranging from (1) indicating 'Completely Disagree' to (5) indicating 'Completely Agree'. The results of the participants feedback is illustrated in Table and Figure .

- Did the CI-forum help the participant to identify the trustworthy user?

Figure 1: A screenshot of list of posts as viewed by end-users on the proposed CI-Forum



- Did the CI-forum help to analyze user expertise?
- Did the CI-Forum provide more overview of the users?

Table 1: User's assessment of the Reputation Model

Sub-factors		Level of Agreements					Mean	
		1	*2*	*3*	*4*	*5*		
Trust	*N*	1	5	11	18	10	3.7	Agree
	%	2.2	11.1	24.4	40	22.2		
User	*N*	1	4	11	16	14	3.9	Agree
Expertise	%	2.2	8.9	24.4	35.6	31.4		
Overview	*N*	2	4	19	19	2	3.3	Agree
of Users	%	4.4	8.9	42.2	42.2	2.2		
Total		*4*	*13*	*41*	*53*	*26*	*3.6*	*Agree*

The participants feedback illustrates that the proposed reputation model helped users in identify trustful users. By showing a separate reputation for each category, users were able to identify the area of expertise of their co-members. The platform also helped users gain a better overview of their co-members overall. As indicated in Table and Figure for every question, most of the participants voted for 'Agree' and the average score was more than 3, so we conclude that the reputation model successfully assists users in making sensible decisions through the use of reputation score. The overall score of 3.6 indicates that all participants agreed with the new reputation model approach and were ready to accept the reputation scores. If a user had a high reputation score, then their fellow users considered them as a trustworthy users and accepted their answers. These results also answer the second research question raised in this work. We can create a reputation model to solve the trust problem in BI by showing separate reputation score for each category, as this method benefits users by helping them identify the experts and helps users select the best inputs according to this information. This further aids BI to maintain data quality thereby assisting in sensible decision-making. We argue that this approach compels users to contribute consistently and mimics reputation as it exists in the real-world.

5.3 Usability of CI-Forum

To assess usability, the questionnaire (presented to the participants) contained four questions all revolving around the systems user interface and features. Answers to these provide us an overview of user interactions and the usability and ease-of-use of the designed CI-Forum. These questions again were supposed to be answered using a Likert scale ranging from (1) indicating 'Completely Disagree' to (5) indicating 'Completely Agree'.

- CI-Forum is easy to use or not?
- Are you willing to continue using the CI-Forum?
- Is CI-Forum having a clearer and easier operating interface?
- CI-Forum will be recommended to family and friends?

Table 2: User's feedback regarding the usability of the proposed CI-Forum

Sub-factors		Level of Agreements					Mean	
		1	*2*	*3*	*4*	*5*		
Easy to use	*N*	2	7	17	16	3	3.2	Agree
Easy to use	%	4.4	15.6	37.8	35.6	6.7		
Will continue	*N*	0	7	12	20	6	3.6	Agree
to use	%	0.0	15.6	26.7	44.4	13.3		
Easier	*N*	0	7	15	18	6	3.6	Agree
operating	%	0.0	15.6	33.3	40.0	13.3		
interface								
Recommended	*N*	1	6	17	12	9	3.5	Agree
to family and	%	2.2	13.3	37.8	26.7	20.0		
friends								
Total		*3*	*27*	*61*	*66*	*24*	*3.6*	*Agree*

As Table and Figure indicate, the users found the system's interface easy-to-use and the forum in general usable. The users' interaction with CI-Forum were meaningful as they did not face any issues while using the application. Most of the users stated that they would to continue as well as recommended to their friends and family. The mean value of every question was more than 3.

Figure 2: A screenshot of a single posts as viewed by end-users on the proposed CI-Forum

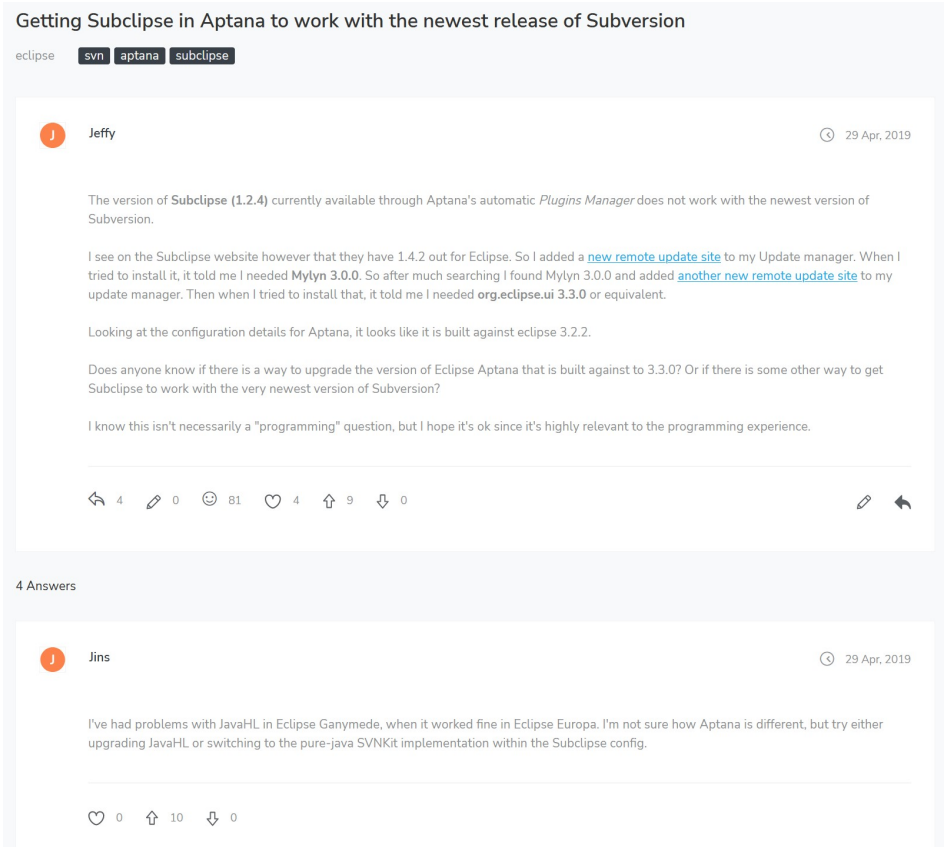
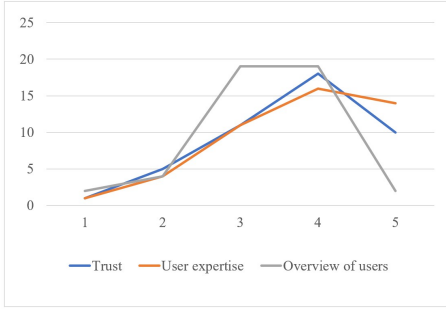
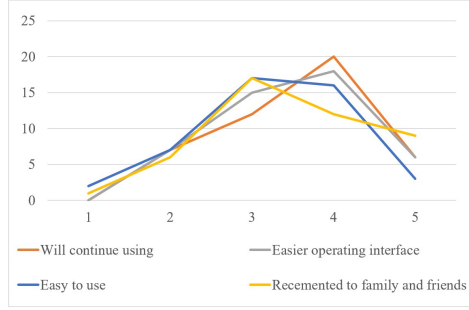


Figure 3: Users' evaluation of the proposed Reputation Model (left) and usability of CI-Forum (right)





The average of the mean value was 3.5, which means that all users were satisfied with their interactions with the CI-Forum. Most users agreed that CI-Forum is useful for their purposes.

During the development phase of the CI-forum, additional feedback was gathered from industry experts, especially those working in the field of software development and testing. These feedback were used to enhance the systems functionalities and usability. Most feedback gathered during this process was positive, and although the experiments with participants was carried out at a smaller sample size, almost all participants simulated actual real-world end-users the CI-forum is oriented towards–as mostly confirmed by the obtained results. The results of the experiments and its following quantitative analysis will be utilized in future to improve the CI-Forum further. The results of the above experiments are only limited by the number and homogeneity of the participant sample, and further user tests are required to develop more conclusive outcomes.

6 CONCLUSION

The overall aim of this work was two main challenges that are encountered when using BI strategies today, these are, sensemaking and trust. Given the critical nature of BI strategies in solving business organizational issues and in supporting organizational decision-making processes; we set out to solve the issues of sensemaking and trust by drawing influences from research in CI. We proposed a novel crowdsourcing approach to reputation models, built around a novel discussion-forum, with focus on organizational employees' perspective and helps establish trust among employees when using BI systems and strategies. By showing users separate reputation scores for each area of expertise, users were able to identify the experts among their fellow users. The idea, behind the approach was that if trustable users works together, the information and results generated by them would be by those organizations is more trustable and sensible to the organizations, specially when compared with non-expert/trustable employees.

The main challenges encountered in this work was that current technologies are still not well adapted to such scenarios. The evaluation of both the reputation model and the CI-forum were only carried out at a small scale, with limited number of participants. Hence the accumulated results only present a superficial view of the usability and usefulness of the proposed contributions. Further changes and fine-tuning is required to enhance the developed artifact. For now, the artifact allows users to identify users with expertise in specific tasks, however, for the next iteration of the forum we would like to develop it so that it can accommodate multi-organization scenarios. Also, as part of future work, we would like to investigate (on a larger scale) and understand the long-term effects of use of reputation scores within organizations and their BI systems.

ACKNOWLEDGMENTS

This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

REFERENCES

[1] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (2013), 76–81. https://doi.org/10.1109/MIC.2013.20

[2] B. Azvine, Z. Cui, D.D. Nauck, and B. Majeed. 2006. Real Time Business Intelligence for the Adaptive Enterprise. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*. 29–29. https://doi.org/10.1109/CEC-EEE.2006.73

[3] Richard J. Boland. 2008. *Decision Making and Sensemaking*. Springer Berlin Heidelberg, Berlin, Heidelberg, 55–63. https://doi.org/10.1007/978-3-540-48713-5_3

[4] Efthimios Bothos, Dimitris Apostolou, and Gregoris Mentzas. 2012. Collective intelligence with web-based information aggregation markets: The role of market facilitation in idea management. *Expert Systems with Applications* 39, 1 (2012), 1333–1345. https://doi.org/10.1016/j.eswa.2011.08.014

[5] Daren C. Brabham. 2008. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence* 14, 1 (2008), 75–90. https://doi.org/10.1177/1354856507084420

[6] Daren C. Brabham. 2013. *Crowdsourcing*. MIT Press. https://mitpress.mit.edu/books/crowdsourcing

[7] Thierry Buecheler, Jan Henrik Sieg, Rudolf Marcel Füchslin, and Rolf Pfeifer. 2010. Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. In *The 12th International Conference on the Synthesis and Simulation of Living Systems, Odense, Denmark, 19–23 August 2010*. MIT Press, 679–686. https://doi.org/10.21256/zhaw-4094

[8] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36, 4 (2012), 1165–1188. http://www.jstor.org/stable/41703503

[9] Xiaowen Chu, Xiaowei Chen, Kaiyong Zhao, and Jiangchuan Liu. 2010. Reputation and trust management in heterogeneous peer-to-peer networks. *Telecommunication Systems* 44, 3 (01 Aug 2010), 191–203. https://doi.org/10.1007/s11235-009-9259-5

[10] Thomas D. Clark, Mary C. Jones, and Curtis P. Armstrong. 2007. The Dynamic Structure of Management Support Systems: Theory Development, Research Focus, and Direction. *MIS Quarterly* 31, 3 (2007), 579–615. http://www.jstor.org/stable/25148808

[11] Ian J. Deary. 2012. Intelligence. *Annual Review of Psychology* 63, 1 (2012), 453–482. https://doi.org/10.1146/annurev-psych-120710-100353 arXiv:https://doi.org/10.1146/annurev-psych-120710-100353 PMID: 21943169.

[12] Pierpaolo Dondio and Luca Longo. 2011. Trust-Based Techniques for Collective Intelligence in Social Search Systems. In *Next Generation Data Technologies for Collective Computational Intelligence*, Nik Bessis and Fatos Xhafa (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 113–135. https://doi.org/10.1007/978-3-642-20344-2_5

[13] R. Duane Ireland and Justin W. Webb. 2007. Strategic entrepreneurship: Creating competitive advantage through streams of innovation. *Business Horizons* 50, 1 (2007), 49–59. https://doi.org/10.1016/j.bushor.2006.06.002

[14] Colette Dumas. 2010. Hosting Conversations for Effective Action. *Journal of Knowledge Globalization* 3, 1 (2010), 99 – 116. https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=51902286&site=eds-live

[15] Mark Dworkin, Lois Foreman-Wernet, and Brenda Dervin. 1999. Sense-Making and television news: An inquiry into audience interpretations. *The Electronic Journal of Communication* 9, 2 (1999). http://www.cios.org/EJCPUBLIC/009/2/009217.html

[16] Diego Gambetta et al. 2000. Can we trust trust. *Trust: Making and breaking cooperative relations* 13 (2000), 213–237.

[17] John Hancock and Roger Toren. 2006. *Practical Business Intelligence with Sql Server 2005* (first ed.). Addison-Wesley Professional.

[18] Borut Hočevar and Jurij Jaklič. 2010. Assessing benefits of business intelligence systems–a case study. *Management: journal of contemporary management issues* 15, 1 (2010), 87–119. https://hrcak.srce.hr/53609

[19] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. 2009. A Survey of Attack and Defense Techniques for Reputation Systems. *ACM Comput. Surv.* 42, 1, Article 1 (Dec. 2009), 31 pages. https://doi.org/10.1145/1592451.1592452

[20] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. 2011. Big data, analytics and the path from insights to value. *MIT sloan management review* 52, 2 (2011), 21–32. https://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

[21] Thomas W. Malone and Michael S. Bernstein (Eds.). 2015. *Handbook of Collective Intelligence*. The MIT Press, Cambridge, MA. 230 pages.

[22] Morteza Namvar, Jacob L. Cybulski, and Luckmika Perera. 2016. Using business intelligence to support the process of organizational sensemaking. *Communications of the Association for Information Systems* 38 (3 2016), 330–352. Issue 1. https://doi.org/10.17705/1cais.03820

[23] Ojelanki K. Ngwenyama and Noel Bryson. 1999. Making the information systems outsourcing decision: A transaction cost approach to analyzing outsourcing decision problems. *European Journal of Operational Research* 115, 2 (1999), 351–367. https://doi.org/10.1016/S0377-2217(97)00171-9

[24] Muhmmad I. Nofal and Zawiyah M. Yusof. 2013. Integration of Business Intelligence and Enterprise Resource Planning within Organizations. *Procedia Technology* 11 (2013), 658–665. https://doi.org/10.1016/j.protcy.2013.12.242 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013.

[25] O Pilli. 2014. LMS Vs. SNS: Can social networking sites act as a learning management systems. *American International Journal of Contemporary Research* 4, 5 (2014), 90–97. http://www.aijcrnet.com/journals/Vol_4_No_5_May_2014/9.pdf

[26] V. Pirttimaki. 2007. Conceptual analysis of business intelligence. *SA Journal of Information Management* 9, 2 (2007). https://doi.org/10.4102/sajim.v9i2.24

[27] Mahesh S Raisinghani. 2003. *Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks: Opportunities, Limitations and Risks.* Idea Group Pub. https://books.google.ee/books?id=xKszZbc7RhYC

[28] Paul Resnick and Richard Zeckhauser. 2002. Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system. Advances in Applied Microeconomics, Vol. 11. Emerald Group Publishing Limited, 127–157. https://doi.org/10.1016/S0278-0984(02)11030-3

[29] J.P. Shim, Merrill Warkentin, James F. Courtney, Daniel J. Power, Ramesh Sharda, and Christer Carlsson. 2002. Past, present, and future of decision support technology. *Decision Support Systems* 33, 2 (2002), 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7 Decision Support System: Directions for the Nest Decade.

[30] Shweta Suran, Vishwajeet Pattanaik, and Dirk Draheim. 2020. Frameworks for Collective Intelligence. *Comput. Surveys* 53, 1 (may 2020), 1–36. https://doi.org/10.1145/3368986

[31] Shweta Suran, Vishwajeet Pattanaik, Sadok Ben Yahia, and Dirk Draheim. 2019. Exploratory Analysis of Collective Intelligence Projects Developed Within the EU-Horizon 2020 Framework. In *Computational Collective Intelligence*, Ngoc Thanh Nguyen, Richard Chbeir, Ernesto Exposito, Philippe Aniorté, and Bogdan Trawiński (Eds.). Springer International Publishing, Cham, 285–296.

[32] Carlo Vercellis. 2011. *Business intelligence: data mining and optimization for decision making.* Wiley Online Library. https://bit.ly/3FrU9fZ

[33] C. Nadine Wathen and Jacquelyn Burkell. 2002. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 134–144. https://doi.org/10.1002/asi.10016

[34] Hugh J. Watson and Barbara H. Wixom. 2007. The Current State of Business Intelligence. *Computer* 40, 9 (2007), 96–99. https://doi.org/10.1109/MC.2007.331

[35] K.E. Weick. 2012. *Making Sense of the Organization, Volume 2: The Impermanent Organization.* Wiley. https://bit.ly/3oEHIY4

[36] K.E. Weick and K.E.W. Weick. 1995. *Sensemaking in Organizations.* SAGE Publications. https://bit.ly/3BkaKzT

[37] Michael Weiss. 2016. Crowdsourcing Literature Reviews in New Domains. *Technology Innovation Management Review* 6 (02/2016 2016), 5–14. https://doi.org/10.22215/timreview/963

[38] Sean Wise, Robert A. Paton, and Thomas Gegenhuber. 2012. Value co-creation through collective intelligence in the public sector. *VINE* 42, 2 (01 Jan 2012), 251–276. https://doi.org/10.1108/03055721211227273

[39] Barbara Wixom and Hugh Watson. 2010. The BI-Based Organization. *International Journal of Business Intelligence Research (IJBIR)* 1, 1 (2010), 13–28. https://doi.org/10.4018/jbir.2010071702

[40] Öykü Işık, Mary C. Jones, and Anna Sidorova. 2013. Business intelligence success: The roles of BI capabilities and decision environments. *Information & Management* 50, 1 (2013), 13–23. https://doi.org/10.1016/j.im.2012.12.001

Appendix V

V

S. Arakkal Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia. Grand reports: A tool for generalizing association rule mining to numeric target values. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery*, volume 12393 of *Lecture Notes in Computer Science*, pages 28–37, Cham, 2020. Springer. doi:10.1007/978-3-030-59065-9_3

Grand Reports: A Tool for Generalizing Association Rule Mining to Numeric Target Values

Sijo A. Peious[1][0000−0002−7858−9463], Rahul Sharma[1][0000−0002−9024−8768], Minakshi Kaushik[1][0000−0002−6658−1712], Syed Attique Shah[2][0000−0003−2949−7391], and Sadok Ben Yahia[3][0000−0001−8939−8948]

[1] Information Systems Group, Tallinn University of Technology, Estonia
{sijo.arakkal,rahul.sharma,minakshi.kaushik}@taltech.ee
[2] Faculty of Information and Communication Technology
BUITEMS, Quetta, Pakistan
attique.shah@buitms.edu.pk
[3] Software Science Department, Tallinn University of Technology, Estonia
sadok.ben@taltech.ee,

Abstract. Since its introduction in the 1990s, association rule mining(ARM) has been proven as one of the essential concepts in data mining; both in practice as well as in research. Discretization is the only means to deal with numeric target column in today's association rule mining tools. However, domain experts and decision-makers are used to argue in terms of mean values when it comes to numeric target values. In this paper, we provide a tool that reports mean values of a chosen numeric target column concerning all possible combinations of influencing factors – so-called grand reports. We give an in-depth explanation of the functionalities of the proposed tool. Furthermore, we compare the capabilities of the tool with one of the leading association rule mining tools, i.e., RapidMiner. Moreover, the study delves into the motivation of grand reports and offers some useful insight into their theoretical foundation.

Keywords: Grand report · association rule mining · relational algebra

1 Introduction

The continuous development of information technology created a massive amount of data [8]. Data mining can be defined, in general, as the process of finding rules and patterns from large data sets. Association rule mining (ARM) is one of the leading data mining techniques. ARM helps to find relationships between attributes and to create rules according to these relationships. In ARM, impractical rules are created along with important rules. The ARM algorithms compare association rules with ancestor association rules to eliminate redundant and impractical rules [12]. The target column and its influencing factors are the heart

of an association rule. Nevertheless, in RapidMiner, it is often called as a conclusion and premises. Support and confidence are the two essential measures of interestingness. Support and confidence calculate the strength of the association between itemsets.

The implementation of support and confidence with a minimum threshold could eliminate the association which are below than this minimum threshold. Some useful association rules might miss due to this method. The decision-makers need to observe all the valuable association and its measures in order to make productive decisions. The listing of all possible combinations and its measures called as a grand report [4]. A grand report contains the list of associations and their measures on every attribute in a data set. The ARM tools discretize the target numeric value columns for easy handling. The discretized values association rules give an idea of the association and its measures, however, decision-makers need to have a better picture. The mean value method could help decision-makers to overcome this situation.

This paper attempts to implement the grand report and the mean value method in ARM with the help of a new tool[4]. This tool generated a grand report for the data set and calculated the mean value for numeric target columns concerning the influencing factors. The tool also calculates the support, lift and conditional probability for the target columns. Grand report and mean value calculations are generated with the help of relational algebra functions. A comprehensive description of the grand report and its calculations is given in section 2. Section 3 illustrates the details about ARM and discretization of the target numeric values. Whereas, development and comparison of the tool are described in section 4.

2 Grand Reports

A *grand report* is the complete print-out of a generalized association rule. In the grand report, work with most minimal minimum supports (i.e., support threshold larger than zero but smaller than $1/N$, with N being the number of rows) and the minimal minimum confidences (i.e. zero). The grand report could produce all possible combinations of the influencing factors against the target columns. A grand report is the complete unfolding of the pivot table. The grand report will generate many rows as a report. This is because it does not have any constraints on support and confidence. It is generalized ARM through which analyst incorporate every line of the report in decision making [4]. The grand report generates 2n combinations of influencing factors by using the sum of the binomial coefficients, where n is the number of columns. Usually, a grand report is massive in size, so users might feel inconvenience to read the entire report.
Let T be a table with columns C where $C = X_1 : T_1, \ldots, X_n : T_n$,
$X_1 \ldots X_n$ are the column names and $T_1 \ldots T_n$ are the column types.
To generate the report for table T,

$$\forall 1 \leq \psi \leq n \tag{1}$$

[4] http://grandreport.me/

$$\forall D = \{X_1^{'} : D_1, \ldots, X_{\psi-1}^{'} : D_{\psi-1}\} \subseteq C \qquad (D_i = d_1, \ldots, d_{ni}) \tag{2}$$

$$\forall d_1^{'} \in D_1, \ldots, d_{\psi-1}^{'} \in D_{\psi-1} \tag{3}$$

Here, D is the subset of C and the influencing factor.

$$\forall Y : \Re \in C \; or \; Y = X_{ij} : B, X_i : d_i \in C \tag{4}$$

Y is the target column, $\Re$ is the real-valued numbers. One can use a select query with "where" condition in relation algebra, to generate the grand report. The select query returns the average value of the target column. In "where" conditions, it is necessary to specify the influencing factors:

$$\text{SELECT AVG}(Y) \text{ FROM } T \text{ WHERE } X_1' = d_1, \ldots, X_{\psi-1}' = d_{\psi-1} \tag{5}$$

In SQL we can use "group by" instead of "where" conditions:

$$\text{SELECT AVG}(Y) \text{ FROM } T \text{ GROUP BY } X_1', \ldots, X_{\psi-1}' \tag{6}$$

3 Association Rule Mining (ARM)

Association rule mining(ARM) is the process of finding the association of frequent itemsets in a large data set and generate rules according to the associations [2]. The ARM first introduced by a research team from Vancouver, British Columbia in 1993 [1]. There are numerous algorithms used in ARM; each of it has its advantages and disadvantages [10]. The 'Apriori' algorithm is the most popular and commonly used algorithm in the ARM [2]. Various types of constraints that can be applied to identify interesting association rules from a data set [7]. Sometimes, These constraints will generate different rules according to their property, and these rules might be conflicting [13]. The most popular constraints in association mining are support and confidence with a minimum threshold [9]. Support is the percentage of transactions which contain a particular itemset. For an itemset X, supp(X) is the percentages of transaction which contain X. Confidence defines how often itemset Y occurs during the transaction with itemset X.

$$\text{supp}(X) = \{t \in T \mid t \text{ satisfies } X\}/|T| \tag{7}$$

$$\text{conf}(X \Rightarrow Y\ ) = \text{supp}(X \cup Y)/\text{supp}(X) \tag{8}$$

$$T = Transactions,\ X \subseteq T,\ Y \subseteq T,\ usually : X \cap Y = \theta, |Y| = 1$$

A sample binary representation of data is shown in Table 2. In the table, every row is corresponds to a transaction (T_1, T_2, T_3, T_4), and each column corresponds to a data item. If an item is present in a transaction, then its value is 1 else it is 0 in the table.

$$\begin{array}{l} T_1 : \{Milk, Bread, Diaper, Beer\} \\ T_2 : \{Diaper, Beer\} \\ T_3 : \{Milk, Bread, Diaper, Beer\} \\ T_4 : \{Milk, Bread, Beer\} \end{array} \tag{9}$$

Table 1. Binary data set.

TID	**Milk**	**Bread**	**Diaper**	**Beer**
1	1	1	0	0
2	0	0	1	1
2	1	1	1	1
2	1	1	0	1

In a database, let I be a set of M binary attributes $\{i_1, i_2, i_3, \ldots, i_m\}$ called database items. T be a set of n Transactions $\{t_1, t_2, t_3, \ldots, t_n\}$, each transaction t has a unique ID and is a subset of the Items in I, i.e. t $\subseteq I$.An Association Rule may be represented as an implication of the form

$$X \Rightarrow Y \tag{10}$$

where $X, Y \subseteq I$ (Item set) and $X \cap Y = \phi$. The left-hand side of the implication known as the antecedent and right hand side of the implication is known as consequent:

$$X \Rightarrow Y := X = \{x_1, \ldots, x_n\} \subseteq I \Rightarrow Y = \{y_1, \ldots, y_n\} \subseteq I \tag{11}$$

$$\{Bread, Butter\} \Rightarrow \{Milk, Butter\} \tag{12}$$

In this association rule example $\{Bread, Butter\}$ is antecedent and $\{Milk\}$ is consequent. Generally, an association rule may be represented as a production rule in the expert system, if statement in programming and implication in logic.

In ARM, the target cluster method is used to generate association rules for numeric target values [11]. Target clustering and discretization of the target column are equivalent. In association rule mining numeric target columns are generally discretized for easy mining [3]. Garćia et al. (2013) [6] and Fayyad & Irani (1993) [5] described well about the discretization of the target column. Once the discretization applied on a target column, then it will be easy to identify those columns as binary values. For example, the column age contains the value from 0 to 140, and column age discretized into different groups. Age 70, 140 group is considered as older people in this example [6]. Most of the interesting measures of ARM are only adaptive with binary target columns [7]. Sometimes misinterpretation of association rules or loss of information occurred by discretization of the target column. Determining the median of the target column, calculation of mean value and identifying the variance of target attributes are the different possible way to find the association rules on numeric target column. Determining the mean value of a numeric target column is much easier than discretization. This tool is using the generalized selection of relational algebra to find the mean value of a numeric target column. For example, 'SELECT' syntax with 'AVG' function.

4 The Proposed Tool

In this study, a web application created to generate a grand report and verify data set. As explained in the previous section, the grand report is complete printout of a generalized association rule. The tool computes all possible combinations of influencing factors against the target column to generate generalized association rules. A data set of four columns has used to test the application. This data set creates seven combinations of influencing factors and calculates the aggregate value or conditional probability of the target column for each combination. This tool is capable of accessing data from the Oracle database and Excel file and is a combination of RapidMiner association mining and relational algebra functions.

4.1 Development and Functionalities

ASP.NET, an open-source framework, is used to develop this tool. The Ajax request method used to establish the communication between the server-side and client-side. For a smooth data transfer, JSON serialization and deserialization functions are used. As mentioned earlier, this tool is capable of accessing data from the Oracle database and Excel files. Furthermore, the Oracle Data provider and OLE DB methods used to access the Oracle database and Excel file.

Table 2. Technologies used for the development of the tool.

C#	Language
ASP.NET	Framework
Ajax	To send and retrieve data from a server asynchronously
Oracle Data Provider (ODP)	Data access to the Oracle database
OLE DB	Data access to the Excel file

Determining the support, lift and conditional probability or aggregate values are the main functionality of this tool. A few steps need to be carried out to find these results. At first, the user needs to select the input source. It can be either Oracle or Excel source. The user needs to provide host address, port number, sid, table name, username and password to connect the Oracle database. Whereas, the user needs to upload the file for Excel. If the uploaded Excel file contains more than one sheet, then the user needs to select the sheet name as well. After these steps, the tool will load each column head with a radio button and checkbox. The radio button is to set the column as the target column and checkbox is to set the column as an influencing factor. If a column is selected as a target column, then it can not be selected for influencing factor and vice versa. After selecting the target column and influencing factors, press the report button to generate the report. While generating the report, the tool will identify the target column type.

The aggregate function is used for the numeric value target column. In aggregate, the average value of the target column calculated against the influencing

factors (select AVG(target column) from table group by influencing factors). If the target column type is categorical, the tool will calculate the conditional probability of the target column (select (conditional probability of target column under influencing factor) from table group by target column and influencing factors). For both functions, support and lift also calculated. The order of columns for numeric target column report is support, lift, the average value of the target column and the influencing factors. The column order for categorical target column report is support, lift, conditional probability, target column and the influencing factors. A sample pseudo-code for finding the combinations of influencing factors and retrieving data from the Oracle database is given in Listing 1.

Listing 1 Pseudo-code for finding the combinations of influencing factors and retrieving data from the Oracle database.

```
FUNCTION column_combination(
  influencingColumns:STRING[], numberofColumns:INTEGER,
  startPosition:INTEGER, columns:STRING[]
)
IF numberofColumns != 0 THEN
FOR i FROM startPosition TO lengthOf(influencingColumns)-1
columns[lengthOf(columns)-numberofColumns] := influencingColumns[i]
call: column_combination(
influencingColumns, numberofColumns-1, i + 1, columns
    )
ENDFOR
ENDFUNCTION
FUNCTION generate_report (
  tableName:STRING, targetColumn:STRING,
  influencingColumns:STRING[], numberofColumns:INTEGER
)
columnCombination := call: column_combination(
                           influencingColumns, numberofColumns,
                           0, columns
                           )
orclQuery := "
  Select Count(*)/ (Select count(*) from  tableName ) AS SUPPORT,
  (Select Avg(targetColumn)from tableName) / Avg(targetColumn) AS LIFT,
  Avg(targetColumn) AS AVG_targetColumn, columnCombination
  from tableName group by columnCombination order by columnCombination";
ENDFUNCTION
```

There are three different colours used in the report. Red colour indicates influencing factors; green used for target/principle measures. Principle measures are the average value of the target column or conditional probability of the target column, and target means of the target column. Either conditional probability or the average value will be present in the report. Blue is used to further measures

like support and lift. Support and lift showed in the first, in order to maintain the uniqueness. For numeric target columns, there is only one column in green because the average column is the representation of the target columns aggregate value. In categorical target column, there are two columns in green. The first column is the conditional probability of the target column, and the second one is the target column value.

In categorical target column, it shows the fibrillation mechanism. It means the tool will compute the conditional probability for all instances of the target column. For example, consider a target column called 'education' and its values are 'A, B, C, D, E'. The column name 'education' is the factor, and its values are the instances of the factor. In report generation, the tool will calculate the conditional probability for each instance.

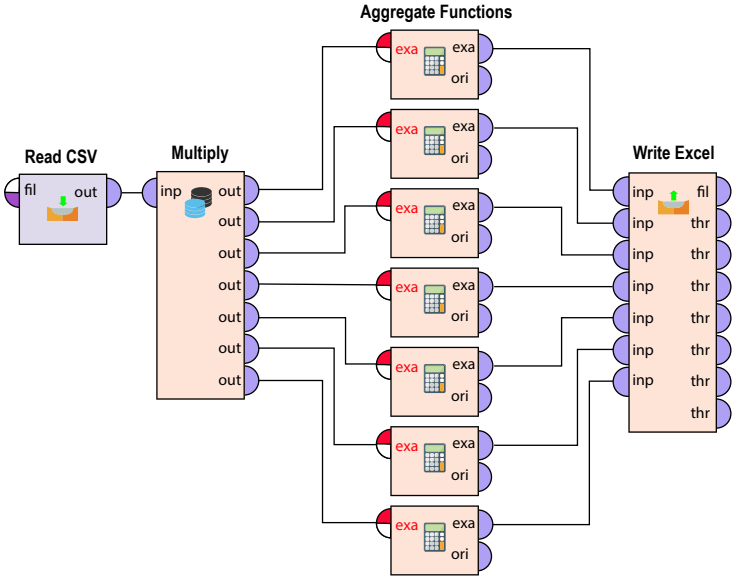
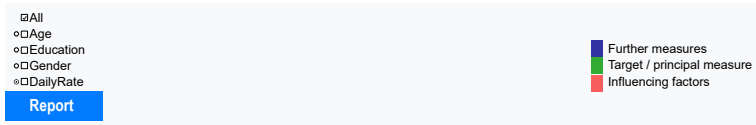


Fig. 1. A sample project for generating all possible combinations of influencing factors against target column in RapidMiner.

4.2 Comparison and Advantages

This tool is a combination of RapidMiner association rule mining and relational algebra. In RapidMiner average value of numeric target column against all pos-



#	SUPPORT	LIFT	AVG_DailyRate	Age	Education	Gender
1	0.222	0.97	825.44	20-30		
2	0.116	0.98	822.42		A	
3	0.400	0.99	808.27			Female
4	0.056	0.96	837.75	20-30	A	
5	0.082	0.93	836.38	20-30		Female
6	0.018	0.86	928.96	20-30	A	Female
7	0.041	0.96	836.97		A	Female

Fig. 2. First record of all combinations in grand report.

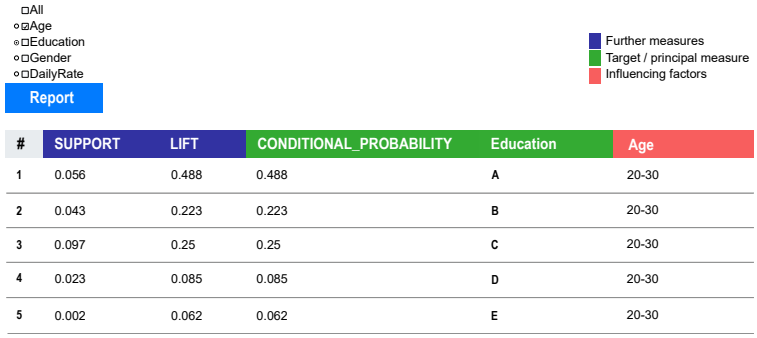
sible combinations of influencing factors, the user needs to create different functions for different combinations (Fig.1.). Similarly, for different data set user needs to modify the columns and its combinations. That means users need to create a different project for different data sets. In this tool, the user needs to select the target column and select all from the list to generate all combinations of influencing factors. The tool will automatically identify the combination and generate the report (Fig.2.).

Education = A	Education = B	Education = C	Education = D	Education = E
true	false	false	false	false
false	true	false	false	false
true	false	true	false	false
false	false	true	true	false
false	true	false	false	true

Fig. 3. Binominal column conversion in RapidMiner.

In RapidMiner categorical columns are converted into binominal columns (Fig.3.). The binominal columns are treated as separate columns, and a separate report generated for each column. RapidMiner shows influencing factors and its

values in the column called 'conclusion'. Meanwhile, target columns and values in the column called 'premises' (Fig.4.). In conclusion and premises column, combinations are shown in the style of "factor=value, factor=value". It is hard for users to identify each factor and its instance. The grand report tool creates separated columns for each factor to avoid this. In this tool, all the measures are located on the left side of the table. It is beneficial for users to verify the output. In the next stage, the column will have the options for filtering the output.



#	SUPPORT	LIFT	CONDITIONAL_PROBABILITY	Education	Age
1	0.056	0.488	0.488	A	20-30
2	0.043	0.223	0.223	B	20-30
3	0.097	0.25	0.25	C	20-30
4	0.023	0.085	0.085	D	20-30
5	0.002	0.062	0.062	E	20-30

Fig. 4. Premises and conclusion in RapidMiner.

Premises	Conclusion
Education = D	Gender, Age = 30-40
Gender	Age = 30-40, Education = D

Fig. 5. Grand report of a categorical target column.

5 Conclusion

The outcome of the present study shows that the study could able to generate the grand report for a data set. The study tried to calculate the mean value for the numeric target columns. The grand report, which generated in this tool, is providing more association rules and giving a better understating of associations between the attributes. The discretization and mean value calculation creates different kinds of the association on numeric target columns. One of the attractions of the studied tool is that the decision-makers can quickly identify different

measures and influencing factors with the help of different colours. It observed that the grand report generates numerous number of records. It is deemed that a filtering option must be needed in place to make the grand report more user friendly. The tool can be accessed publicly (http://grandreport.me/).

6 Acknowledgements

This work has been conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. pp. 207–216 (1993)
2. Cox, M.T., Funk, P., Begum, S.: Case-based Reasoning Research and Development: 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings, vol. 11156. Springer (2018)
3. Draheim, D.: Generalized Jeffrey Conditionalization: A Frequentist Semantics of Partial Conditionalization. Springer (2017)
4. Draheim, D.: Future perspectives of association rule mining based on partial conditionalizationin (DEXA'2019 keynote). In: Proceedings of DEXA'2019 - the 30th International Conference on Database and Expert Systems Applications, LNCS 11706, (2019)
5. Fayyad, U., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI). pp. 1022–1027 (1993)
6. Garcia, S., Luengo, J., Sáez, J.A., Lopez, V., Herrera, F.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering **25**(4), 734–750 (2012)
7. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR) **38**(3), 9–es (2006)
8. Han, J., Kamber, M.: Data mining concepts and techniques. San Francisco, CA pp. 335–391 (2001)
9. Hornik, K., Grün, B., Hahsler, M.: arules-a computational environment for mining association rules and frequent item sets. Journal of Statistical Software **14**(15), 1–25 (2005)
10. Kumbhare, T.A., Chobe, S.V.: An overview of association rule mining algorithms. International Journal of Computer Science and Information Technologies **5**(1), 927–930 (2014)
11. Moreland, K., Truemper, K.: Discretization of target attributes for subgroup discovery. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition. pp. 44–52. Springer (2009)
12. Srikant, R., Agrawal, R.: Mining generalized association rules. Future generation computer systems **13**(2-3), 161–180 (1997)
13. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems **29**(4), 293–313 (2004)

Curriculum Vitae

1. Personal data

Name	Sijo Arakkal Peious
Date and place of birth	24 July 1988 Kerala, India
Nationality	India

2. Contact information

Address	Tallinn University of Technology, School of Information Technologies, Department of Software Science, Akadeemia tee 15A, 12618 Tallinn, Estonia
Phone	+372 58794695
E-mail	sijo.arakkal@taltech.ee

3. Education

2019-...	Tallinn University of Technology, School of Information Technologies, Estonia, PhD studies
2017-2019	Tallinn University of Technology, School of IT Estonia, MSc in E-Governance Technologies and Services
2010–2012	Annamalai University, Faculty of Computer and Information Science, Tamil Nadu, India, Master of Computer Application, MSc
2006–2009	University of Calicut, Faculty of Science, Bachelor of Computer Science, BSc

4. Language competence

Malayalam	native
English	fluent

5. Professional employment

Since 2019	Tallinn University of Technology, School of Information Technologies, Department of Software Science, Early Stage Researcher
2016–2017	Code IT Solutions, Qatar, Software Engineer
2014–2016	GISI E-creations Pvt. Ltd., Software Engineer
2013–2014	Iser Global Solutions Pvt Ltd., ASP .NET Developer
2010–2013	Smartway Solutions Pvt Ltd., ASP .NET Developer
2009–2010	Xtrem InfoTech, Junior Software Developer

6. Defended theses

- 2019,“ Enabling Sensemaking and Trust in Communities; An Organizational Perspective”, MSc, supervisor Prof. Dirk Draheim, co-supervisor Shweta Suran, Tallinn University of Technology, School of IT, Department of Software Science.

7. Field of research

- 4.6. Computer Science

- 4.7. Information and Communications Technologies

8. Scientific work

1. S. Arakkal Peious, M. Kaushik, M. Shahin, R. Sharma, and D. Draheim. On choosing the columnar in-memory database Hyrise as high-performant implementation platform for the GrandReport tool. In *Proceedings of NGISE'2025 – the 1st International Conference in Next Generation Information Systems Engineering*, pages 1–7. IEEE, 2025. to appear, preprint available at IEEE TechRxiv, doi: 10.36227/techrxiv.174015861.15410981/v1

2. S. Arakkal Peious, M. Kaushik, S. A. Shah, R. Sharma, S. Suran, and D. Draheim. On measuring confounding bias in mixed multidimensional data. In J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, editors, *Proceedings of ICTIS'2024 – the 8th International Conference on ICT for Intelligent Systems, Volume 6*, volume 1112 of *Lecture Notes in Network and Systems*, pages 329–342, Singapore, 2024. Springer Nature Singapore. doi:10.1007/978-981-97-6684-0_27

3. S. Arakkal Peious, R. Sharma, M. Kaushik, S. Mahtab, and D. Draheim. On observing patterns of correlations during drill-down. In P. D. Haghighi, E. Pardede, G. Dobbie, V. Yogarajan, N. A. Sanjaya, G. Kotsis, and I. Khalil, editors, *Proceedings of iiWAS'2023 – the 25th International Conference on Information Integration and Web-based Applications and Services*, volume 14416 of *Lecture Notes in Computer Science*, pages 134–143, Cham, 2023. Springer. doi:10.1007/978-3-031-48316-5_16

4. R. Sharma, M. Kaushik, S. Arakkal Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, and D. Draheim. A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. *IEEE Access*, 10:12792–12813, 2022. doi:10.1109/ACCESS.2022.3142537

5. R. Sharma, M. Kaushik, S. Arakkal Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *Proceedings of ADBIS 2022 – the 26th International Conference on New Trends in Database and Information Systems*, volume 1652 of *Communications in Computer and Informations Science*, pages 67–76, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-15743-1_7

6. R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. S. Yadav, and D. Draheim. Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, and R. Uday Kiran, editors, *Proceedings of DASFAA 2022 – the 27th International Conference on Database Systems for Advanced Applications*, volume 13245 of *Lecture Notes in Computer Science*, pages 596–603, Cham, 2022. Springer International Publishing. 10.1007/978-3-031-00123-9_48

7. R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim. Why not to trust big data: Discussing statistical paradoxes.

In U. K. Rage, V. Goyal, and P. K. Reddy, editors, *Proceedings of DASFAA 2022 International Workshops – the 27th International Conference on Database Systems for Advanced Applications*, volume 13248 of *Lecture Notes in Computer Science*, pages 50–63, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-11217-1_4

8. R. Sharma, H. Garayev, M. Kaushik, S. Arakkal Peious, P. Tiwari, and D. Draheim. Detecting Simpson's paradox: A machine learning perspective. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA 2022 – the 33rd International Conference on Database and Expert Systems Applications*, volume 13426 of *Lecture Notes in Computer Science*, pages 323–335, Cham, 2022. Springer International Publishing. doi:10.1007/978-3-031-12423-5_25

9. R. Sharma, M. Kaushik, S. Arakkal Peious, M. Shahin, A. Vidyarthi, and D. Draheim. Existence of the Yule-Simpson effect: An experiment with continuous data. In *Proceedings of Confluence 2022 – the 12th International Conference on Cloud Computing, Data Science and Engineering*, pages 351–355. IEEE, 2022. doi:10.1109/Confluence52989.2022.9734211

10. M. Kaushik, R. Sharma, M. Shahin, S. Arakkal Peious, and D. Draheim. An analysis of human perception of partitions of numerical factor domains. In M. Indrawan-Santiago, E. Pardede, I. L. Salvadori, M. Steinbauer, I. Khalil, and G. Kotsis, editors, *Proceedings of iiWAS 2022 – the 24th International Conference on Information Integration and Web Intelligence*, volume 13635 of *Lecture Notes in Computer Science*, pages 137–144, Cham, 2022. Springer Nature Switzerland. doi:10.1007/978-3-031-21047-1_13

11. S. Arakkal Peious, S. Suran, V. Pattanaik, and D. Draheim. Enabling sensemaking and trust in communities: An organizational perspective. In E. Pardede, M. Indrawan-Santiago, P. D. Haghighi, M. Steinbauer, I. Khalil, and G. Kotsis, editors, *Proceedings of iiWAS'2021 – the 23rd International Conference on Information Integration and Web Intelligence*, pages 95–103. Association for Computing Machinery, 2021. doi:10.1145/3487664.3487678

12. M. Kaushik, R. Sharma, S. Arakkal Peious, S. Mahtab, S. B. Yahia, and D. Draheim. A systematic assessment of numerical association rule mining methods. *SN Computer Science*, *2*(5):1–13, 2021. doi:10.1007/s42979-021-00725-2

13. M. Kaushik, R. Sharma, S. Arakkal Peious, and D. Draheim. Impact-driven discretization of numerical factors: Case of two- and three-partitioning. In S. N. Srirama, J. C.-W. Lin, R. Bhatnagar, S. Agarwal, and P. K. Reddy, editors, *Proceedings of BDA'21 – the 9th International Conference on Big Data Analytics*, pages 244–260, Cham, 2021. Springer International Publishing. doi:10.1007/978-3-030-93620-4_18

14. M. Shahin, S. Arakkal Peious, R. Sharma, M. Kaushik, S. Ben Yahia, S. A. Shah, and D. Draheim. Big data analytics in association rule mining: A systematic literature review. In *Proceedings of BDET'23 – the 3rd International Conference on Big Data Engineering and Technology*, pages 40–49. Association for Computing Machinery, 2021. doi:10.1145/3474944.347495

15. M. Shahin, S. Saeidi, S. A. Shah, M. Kaushik, R. Sharma, S. Arakkal Peious, and D. Draheim. Cluster-based association rule mining for an intersection accident dataset. In *Proceedings ICE Cube 2021 – the 1st International Conference on Computing, Electronic and Electrical Engineering*, pages 1–6. IEEE, 2021. doi:10.1109/ICECube53880.2021.9628206

16. S. Arakkal Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia. Grand reports: A tool for generalizing association rule mining to numeric target values. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery*, volume 12393 of *Lecture Notes in Computer Science*, pages 28–37, Cham, 2020. Springer. doi:10.1007/978-3-030-59065-9_3

17. M. Kaushik, R. Sharma, S. Arakkal Peious, M. Shahin, S. B. Yahia, and D. Draheim. On the potential of numerical association rule mining. In T. Dang, J. Küng, M. Takizawa, and T. Chung, editors, *Proceedings of FDSE'20 – the 7th International Conference on Future Data and Security Engineering*, volume 1306 of *Communications in Computer and Information Science*, pages 3–20. Springer, 2020. doi:10.1007/978-981-33-4370-2_1

18. R. Sharma, M. Kaushik, S. Arakkal Peious, S. B. Yahia, and D. Draheim. Expected vs. unexpected: Selecting right measures of interestingness. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK 2020 – the 22nd International Conference on Big Data Analytics and Knowledge Discovery*, volume 13428 of *Lecture Notes in Computer Science*, pages 38–47, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-59065-9_4

Elulookirjeldus

1. Isikuandmed

Nimi	Sijo Arakkal Peious
Sünniaeg ja -koht	24.07.1988, Kerala, India
Kodakondsus	Indian

2. Kontaktandmed

Aadress	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Tarkvarateaduse Instituut, Akadeemia tee 15A, Tallinn 12618 Eesti
Telefon	+372 58794695
E-post	sijo.arakkal@taltech.ee

3. Haridus

2019-...	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, doktoriõpe
2017-2019	Tallinna Tehnikaülikool, Eesti, MSc e-valitsemise tehnoloogiate ja teenuste alal
2010–2012	Annamalai ülikool, Arvuti- ja infoteaduste teaduskond, Tamil Nadu, India, Arvutirakenduse magister, MSc
2006–2009	Calicuti ülikool, loodusteaduskond, Arvutiteaduse bakalaureusekraad, BSc

4. Keelteoskus

Malayalam	emakeel
inglise keel	kõrgtase

5. Teenistuskäik

2019-...	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Tarkvarateaduse instituut, nooremteadur
2016–2017	Code IT Solutions, Qatar, Tarkvarainsener
2014–2016	GISI E-creations Pvt. Ltd., Tarkvarainsener
2013–2014	Iser Global Solutions Pvt Ltd., ASP .NET Arendaja
2010–2013	Smartway Solutions Pvt Ltd., ASP .NET Arendaja
2009–2010	Xtrem InfoTech, Noorem Tarkvaraarendaja

6. Kaitstud lõputööd

- 2019,"Võimaldades Sensemaking Ja Usalduse Kogukondades; Organisatsiooni Vaatenurgast", MSc, juhendaja Prof. Dirk Draheim, kaasjuhendaja Shweta Suran, Tallinna Tehnikaülikooli Infotehnoloogia teaduskond, Tarkvarateaduse instituut.

7. Teadustöö põhisuunad

- 4.6. Arvutiteadus
- 4.7. Info- ja sidetehnoloogiad

8. Teadustegevus

Teadusartiklite, konverentsiteeside ja konverentsiettekannete loetelu on toodud ingliskeelse elulookirjelduse juures.

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-290-8 (PDF)