TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Verkehrswissenschaften „Friedrich List“ Institut für Automobiltechnik Dresden – IAD

Lehrstuhl Kraftfahrzeugtechnik

Fusion von Unfallszenarien für die Repräsentativitätsüberprüfung eines Testszenarienkataloges zur Absicherung automatisierter Fahrfunktionen

Linda Dziuba-Kaiser

Geboren am: 8. August 1992 in Braunschweig
Matrikelnummer: 3909241
Immatrikulationsjahr: 2012

Diplomarbeit

zur Erlangung des akademischen Grades

Diplomingenieur (Dipl.-Ing.)

Betreuer

Dipl.-Ing. Maximilian Bäumler, MBA

Betreuender Hochschullehrer

Prof. Dr.-Ing Günther Prokop

Eingereicht am: 9. Dezember 2019

Selbstständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich reiche sie erstmals als Prüfungsleistung ein. Mir ist bekannt, dass ein Betrugsversuch mit der Note „nicht ausreichend“ (5,0) geahndet wird und im Wiederholungsfall zum Ausschluss von der Erbringung weiterer Prüfungsleistungen führen kann.

Name:	Dziuba-Kaiser
Vorname:	Linda
Matrikelnummer:	3909241

Dresden, den 9. Dezember 2019, ...

Linda Dziuba-Kaiser

Kurzreferat

Gegenstand dieser Arbeit ist die Bewertung und Durchführung der Fusion von zwei Datensätzen, die auf Basis der Statistik der Straßenverkehrsunfälle des statistischen Bundesamtes konstruiert werden. Für die Fusionierung wird die Methode der statistischen Datenfusion angewendet. Die zu fusionierenden Datensätze werden auf die Ausgangslage der Datenfusion und Unfalldatenbanken angepasst. Anhand der Zusammenhangsstärke und Verteilung werden die passenden Variablen, die für die Datenfusion verwendet werden können, identifiziert und ausgewählt. Für die Datenfusion werden verschiedene nichtparametrische Verfahren unter der bedingten Unabhängigkeitsannahme (Distanz-Hot-Deck, Random-Hot-Deck) und unter der Beibehaltung der Unsicherheit (Imprecise Imputation) durchgeführt. Zusätzlich werden Qualitätsstufen mit einbezogen, um die Auswirkung von veränderten Variablen auszuwerten. Dabei zeigt sich, dass die Datenfusion unter der bedingten Unabhängigkeit allgemein eine unsichere Methode ist, die jedoch unter Umständen für bivariate Analysen vielversprechende Ergebnisse erzielen kann.

Abstract

The object of this thesis is the valuation and implementation of a merger of two datasets, which are constructed on basis of the statistics elevated from the german Statistisches Bundesamt. For the merger, the method for statistical data fusion is used. The datasets which are to be merged are being fitted to the initial situation of the data fusion and the accident databases. Matching variables that can be used for the data fusion are identified and choosen by the strength of their connection as well as by their distribution. Different nonparametric procedures are being carried out under consideration of presumption of the conditional independence (Distance-Hot-Deck and Random-Hot-Deck) while retaining the uncertainty (Imprecise Imputation). Furthermore the quality levels are being included to evaluate the effect of altered variables. Thereby it becomes clear that the data fusion under consideration of the conditional independence is generally an unreliable method, but that it can achieve promising results for bivariate analyses given the circumstances.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

Formelzeichen	[Einheit]	Beschreibung
A, B, C	[-]	Datensätze
BC	[-]	Bhattacharyya-Koeffizient
α	[-]	Intercept
β	[-]	Regressionskoeffizient
C	[-]	korrigierte Kontingenzkoeffizient (Sakoda's adjusted Pearson's C)
d	[-]	Distanz
$E()$	[-]	Erwartungswert
E	[-]	Vorhersagefehler
$f()$	[-]	Wahrscheinlichkeits- /Dichtefunktion
$\hat{F}()$	[-]	Kumulierte Verteilungsfunktion
g	[-]	Strafterm
H	[-]	Entropie
H_D	[-]	Hellinger-Distanz
H_{D_m}	[-]	Anzahl an Kategorien einer Teilmenge von X Variablen
I	[-]	Indikatorfunktion
I, J, K	[-]	Anzahl Kategorien
i, j, k	[-]	Kategorie ($i = 1, ..., I;\ j = 1, ..., J;\ k = 1, .., K$)
k, l, m, o	[-]	Zellenwert
K	[-]	Anzahl Parameter
L	[-]	Likelihood-Funktion
M	[-]	Teilmenge an Variablen

n	[-]	Anzahl an Einheiten
$\bar{n}$	[-]	Spärlichkeit
P	[-]	Anzahl gemeinsame Variablen
Pr	[-]	Wahrscheinlichkeit
p	[-]	relative Häufigkeit
R	[-]	Spannweite
s	[-]	Ähnlichkeitskoeffizient
U	[-]	Unsicherheitskoeffizient
u	[-]	Unsicherheit
V	[-]	Cramers V
w	[-]	Zuordnungsvariable
X	[-]	gemeinsame Variable
x, y, z	[-]	Beobachtungen im Datensatz
Y, Z	[-]	spezifische Variablen
χ^2	[-]	Chi-Quadrat
δ	[-]	Kronecker-Delta
$\mathbb{R}$	[-]	reelle Zahl
θ	[-]	(mehrere) Verteilungsparameter
Θ	[-]	Parameterraum
$\mathbf{\Sigma_{XX}}$	[-]	Kovarianzmatrix
$\mathbf{\Sigma_{XX}^{-1}}$	[-]	Inverse Kovarianzmatrix

Indizes

Index	Beschreibung
A,B	Datensätze
a,b	Einheit ($a = 1, ..., n_A;\ b = 1, ..., n_B$)
D	Dice

E	Euclidean
G	Gower
i,j,k	Kategorien
J	Jaccard
K	Anzahl Parameter
M	Manhattan
Ml	Mahalanobis
RT	Rogers&Tanimoto
SM	Simple-Matching
1,2	Laufvariable

Abkürzungsverzeichnis

Abkürzung	Beschreibung
CIA	Conditional Independence Assumption
DESTATIS	Statistisches Bundesamt
EU	Europäische Union
EUSka	Elektronische Unfalltypensteckkarte
FDZ	Forschungsdatenzentrum
GIDAS	German In-Depth Accident Study
i.i.d	unabhängig und identisch verteilt
JS	Jensen-Shannon-Divergenzmaß
KL	Kullback-Leibler-Divergenzmaß
kNN	k nearest neighbour
MAR	Missing not at random
MCAR	Missing completely at random
ML	Maximum-Likelihood
MNAR	Missing not at random
NDS	Naturalistic Driving Studies
PRE	proportional reduction of error
SePIA	Szenarien basierte Plattform zur Inspektion automatisierter Fahrfunktionen

1 Einleitung

Fahrerassistenzsysteme sind eine wichtige Entwicklung in den letzten Jahren und bereits in vielen Fahrzeugen eingebaut. Des Weiteren spielt für das Verkehrssystem der Zukunft die Weiterentwicklung zu hoch- und vollautomatisierten Fahrfunktionen eine immer wichtigere Rolle. Eines der bedeutendsten Ziele dabei ist es, den Straßenverkehr allgemein sicherer zu machen und Unfälle zu vermeiden. Eine Voraussetzung dafür ist, dass die automatisierten Fahrfunktionen fehlerfrei funktionieren. Das Fahren unter Zuhilfenahme von automatisierten Fahrfunktionen muss mindestens so sicher sein wie das Fahren eines rein menschlichen Fahrers (Prokop et al., 2017, vgl. S.1). Um diese geforderte Zuverlässigkeit beziehungsweise Sicherheit zu gewährleisten, müssen während und nach der Entwicklung die automatisierten Fahrfunktionen umfangreich getestet und bewertet werden. Während Fahrzeuge auf ihren technischen Zustand hin und ihre Tauglichkeit geprüft werden (zum Beispiel bei der Hauptuntersuchung), gibt es solche Test- und Prüfkonzepte für automatisierte Fahrfunktionen bislang nicht.

Motivation

Für die Homologation und Effektivitätsbewertung von automatisierten Fahrfunktionen ist eine Bewertungsmethodik mit umfangreichen Qualitätsanforderungen erforderlich. Aus diesem Grund ist beispielsweise das Ziel des Projektes SePIA (Szenarien basierte Plattform zur Inspektion automatisierter Fahrfunktionen), eine Plattform mit realistisch simulierten Szenarien zu entwickeln, die zum Testen und Bewerten von automatisierten Fahrfunktionen dienen. Dabei müssen Szenarien generiert werden, die typische Straßenverkehrssituationen umfangreich abdecken und insbesondere die für Unfälle kritischen Szenarien enthalten. Diese Szenarien können anschließend in Testszenarienkatalogen zusammengefasst werden, mit denen Fahrfunktionen abgesichert werden.

Eine Möglichkeit für das Genieren von Szenarien ist die Verwendung von Realfahr- und Unfalldaten. Diese werden in Datenbanken wie beispielsweise GIDAS[1], EUSka[2] und NDS[3]-Daten erfasst. Jedoch umfassen diese Datengrundlagen teilweise verschiedene Schwerpunkte und sind hinsichtlich der erfassten Merkmale mitunter unterschiedlich.
Für die Erstellung der Szenarien des Testszenarienkatalogs besteht eine zeit- und kostengünstige Möglichkeit darin, unterschiedliche Unfalldatenbanken zu fusionieren, um den Informationsgehalt einer Datenbank zu erhöhen.

[1] German In-Depth Accident Study
[2] Elektronische Unfalltypensteckkarte
[3] Naturalistic Driving Studies

Die dadurch gewonnene umfangreiche Variablenauswahl steigert die Repräsentativität der Testszenarien, da anschließend die Möglichkeit besteht, Variablen zu untersuchen, die nie in einer Erhebung zusammen beobachtet wurden.

Problemstellung

Die Generierung von Szenarien mittels fusionierten Daten stellt nur dann eine Möglichkeit dar, wenn mit den resultierenden Datensätzen festgelegte Qualitätsziele erreicht werden und die Daten zu keinen falschen Schlüssen führen. Die Datenfusion wurde hauptsächlich im Bereich der Sozialwissenschaften und Marktforschung angewendet und erforscht (Rässler, 2002). Für den Bereich der Unfallforschung gibt es bisher jedoch nur die Studie des Trace Projektes (Grömping et al., 2007, Kap.4), welche sich mit der Datenfusion von Unfalldatenbanken beschäftigt.

Aus diesem Grund ist das Ziel dieser Arbeit, mit Hilfe simulierter Datensätze auf Basis der Straßenverkehrsunfallstatistik Datenfusionen durchzuführen und die fusionierten Datensätze in Bezug auf die Qualität auszuwerten.

Aufbau der Arbeit

Aufgrund der zuvor genannten Problemstellung gliedert sich der Aufbau der Arbeit folgendermaßen. In **Kapitel 2** wird zunächst eine allgemeine Einführung in die Theorie der Datenfusion gegeben. Dabei wird ein Überblick über die Problemstellung, verschiedene Analysemethoden und Vorbereitungsschritte vorgestellt. Das Kapitel schließt mit verschiedenen Bewertungskriterien zur Beurteilung der Qualität von fusionierten Datensätzen ab. Anschließend werden in **Kapitel 3** zwei simulierte Datensätze auf Grundlage der Straßenverkehrsunfallstatistik konstruiert und sowohl an die allgemeine Ausgangssituation der Datenfusion, als auch exemplarisch an die Rahmenstrukturen der beiden Datenbanken GIDAS und EUSKa, die auch in dem SePIA-Projekt genutzt werden, angepasst. In **Kapitel 4** werden zunächst die Variablen in den simulierten Datensätzen durch Vorbereitungsschritte an die Datenfusion angepasst und ausgewählt. Darauf aufbauend wird die Auswahl der Methoden für die Datenfusion mit unterschiedlichen Qualitätsstufen begründet und durchgeführt. Nachfolgend werden in **Kapitel 5** die Ergebnisse dargestellt und diskutiert. Die Arbeit schließt in **Kapitel 6** mit einer Zusammenfassung und einem Ausblick ab.

2 Grundlagen

Datenfusionen bieten eine vielversprechende Möglichkeit, Datenquellen miteinander zu verbinden. Dieses Kapitel zeigt daher die Grundlagen für die Durchführung von Datenfusionen. Im Rahmen dieser Arbeit erfolgt die Datenfusion anhand der Straßenverkehrsunfallstatistik, die an die Rahmenstruktur der in SePIA verwendeten GIDAS und EUSKa Datenbanken angepasst wird. Aus diesem Grund bietet dieses Kapitel zunächst einen Überblick über die beiden Datenbanken sowie die Straßenverkehrsunfallstatistik. Um ein allgemeines Verständnis der Datenfusion zu erhalten, wird die Theorie zunächst erläutert und anschließend werden verschiedene Analysemethoden und Vorbereitungsschritte vorgestellt, die mit den Datensätzen vor der eigentlichen Datenfusion durchgeführt werden müssen. Den Abschluss dieses Kapitels bildet ein Überblick über die Bewertungskriterien der Datenfusion.

2.1 Unfalldatenbanken

Für die Datengrundlagen werden unter anderem Datenbanken wie GIDAS und EUSKa verwendet, aus denen Fahr- und Unfalldaten extrahiert werden können, um anschließend daraus Szenarien zu generieren. Jede der Datenbanken weist eine unterschiedliche Anzahl an Variablen und Eigenschaften auf. Nachfolgend wird die Straßenverkehrsunfallstatistik näher erläutert.

2.1.1 GIDAS

GIDAS (German In-Depth Accident Study) gehört zu den In-Depth-Untersuchungen („Unfallerhebungen am Unfallort“) (Johannsen, 2013, vgl. S.32f). Bei GIDAS handelt es sich um eine Stichprobe mit einer umfangreichen und detaillierten Dokumentation von Unfalldaten. Pro Unfall werden durchschnittlich 3500 verschiedene Informationen, wie zum Beispiel Einlauf- und Kollisionsgeschwindigkeiten, sowie allgemeine Angaben zur Unfallumgebung aufgenommen (Brunner und Georgi, 2003). In Deutschland werden in zwei verschiedenen Erhebungsgebieten (Dresden und Hannover) seit 1999 jeweils ca. 1000 Unfälle jährlich mit Personenschaden aufgezeichnet. Das Erhebungsgebiet in Dresden setzt sich dabei aus der Region Dresden sowie Teilen der Landkreise Meißen, Riesa-Großenhain, Weißeritzkreis, Sächsische Schweiz, Bautzen und Kamenz zusammen (Hautzinger et al., 2006). Die Unfallerhebung- und rekonstruktion erfolgt in Dresden durch die Verkehrsunfallforschung an der Technischen Universität Dresden.

2.1.2 EUSka

Die EUSka (Elektronische Unfalltypensteckkarte) ist ein Erfassungs- und Analysesystem zur Auswertung von Verkehrsunfalldaten.

EUSKa wird unter anderem von der Polizei verwendet und stellt die Verkehrsunfalldatenbank der sächsischen Polizei dar. Pro Unfall werden ca. 110 Variablen von der Polizei erfasst. Dabei handelt es sich beispielsweise um Informationen zu Unfallort, Unfallfolgen oder Unfalltyp. (Bönninger, 2018)

2.1.3 Straßenverkehrsunfallstatistik

Die Datenbank des Statistischen Bundesamtes (DESTATIS) enthält statistische Informationen zu verschiedenen Themengebieten wie beispielsweise Wirtschaft, Gesellschaft, Umwelt und ähnlichem. Ein Themengebiet sind des Weiteren die Verkehrsunfälle, die sich in Schienen- und Straßenverkehrsunfälle aufgliedern. Die Straßenverkehrsunfallstatistik ist eine Sekundärstatistik und wird daher nicht direkt durch DESTATIS erhoben, sondern sie umfasst alle Straßenverkehrsunfälle, die durch die Polizei im gesamten deutschen Bundesgebiet aufgenommen werden. Das bedeutet, dass Unfälle bei denen die Polizei nicht hinzugezogen wird, nicht in der Straßenverkehrsunfallstatistik aufgeführt sind. Außerdem beinhaltet die Datenbank keine Unfälle, bei denen ausschließlich Fußgänger involviert sind. (Statistisches Bundesamt (DESTATIS), 2017)

Die Unfallszenarien, die in der Statistik beschrieben werden, lassen sich in zwei Bereiche einteilen: Angaben direkt zum Unfall und Angaben zu Unfallbeteiligten. Die Angaben zum Unfall umfassen dabei Informationen hinsichtlich des Unfallhergangs, der -ursache sowie zur -umgebung. Tabelle 2.1 stellt einen Ausschnitt über verschiedenen Merkmale dar, die sich den Bereichen zuordnen lassen.

Tabelle 2.1: Ausschnitt der Merkmale zum Unfall in der Straßenverkehrsunfallstatistik

Bereich	Merkmal	Beschreibung
Unfallumgebung	Straßenklasse	Art der Straße, z.B. Bundesstraße
	Lichtverhältnis	Lichtbeschaffenheit, z.B. Tageslicht
Unfallhergang	Unfalltyp	Der Unfalltyp beschreibt die Konfliktsituation
	Unfallart	Beschreibt den Unfall, z.B. Kollision

Zu den Unfallbeteiligten werden alle Fahrzeuge, Fußgänger sowie andere Personen gezählt, die selbst oder deren Fahrzeug Schaden erlitten oder hervorgerufen haben (Statistisches Bundesamt (DESTATIS), 2019a, vgl. S.12). Dabei beziehen sich die Merkmale in der Statistik zum Beispiel auf die Eigenschaften der Beteiligten wie Geschlecht oder Fahrzeugklasse, aber auch auf die Ursachen, die durch die Unfallbeteiligten zu einem Unfall geführt haben.

Der Zugriff auf die anonymisierten Mikrodaten der Straßenverkehrsunfallstatistik des Jahres 2013 (FDZ der Statistischen Ämter des Bundes und der Länder, 2013) finden an einem Gastwissenschaftsarbeitsplatz im Forschungsdatenzentrum (FDZ) in Dresden statt. Aus Datenschutzgründen ist es jedoch nur möglich, die vollständigen Daten am Arbeitsplatz innerhalb des FDZs zu nutzen. Anschließend dürfen lediglich vom FDZ kontrollierte aggregierte Daten außerhalb des Forschungsdatenzentrum und damit in dieser Arbeit verwendet werden.

2.2 Datenfusion

Die Datenfusion gehört zu dem Bereich der Datenintegration und beinhaltet das Ziel, Informationen aus unterschiedlichen Datenquellen miteinander zu verknüpfen. Grundsätzlich wird dabei zwischen „Record Linkage“ und „Datenfusion“ unterschieden.

Beim Record Linkage werden Datensätze miteinander verknüpft, die aus unterschiedlichen Quellen stammen, aber jeweils dieselben Einheiten betrachten. Statistische Einheiten werden auch als Merkmalsträger bezeichnet und sind jeweils die Einzelobjekte, die untersucht und zu denen bestimmte Informationen (Merkmale) gesammelt werden (Fahrmeir et al., 2016, S.13). Allgemein stammen die Merkmalsträger dabei aus unterschiedlichen Bereichen, die der Tabelle 2.2 zu entnehmen sind (Stocker und Steinke, 2016, vgl. S. 21).

Bereich	Beispiel
Personen	Studenten
Gegenstände	Fahrzeuge
geographische Orte	Länder
Vorgänge	Unfälle
Ereignisse	Geburten

Tabelle 2.2: Zuordnung von Merkmalsträgern

Beim Record Linkage werden beispielsweise Daten aus zwei Befragungen miteinander verknüpft, die jeweils an den gleichen Studenten durchgeführt wurden. Anhand eindeutiger gemeinsamer Identifikationsschlüssel (wie Steuernummer, Name, Adresse, Personenkennziffer) können die Datensätze zusammengeführt werden. Schwierigkeiten bei diesem Verfahren treten auf, wenn Identifikationsschlüssel fehlerbehaftet sind (durch Tipp- oder Erfassungsfehler) oder unterschiedliche Schreibweisen aufweisen (zum Beispiel ae anstatt ä). Der Vorgang basiert daher beim Record Linkage auf einer Ähnlichkeitssuche in den Identifizierungsschlüsseln. (Cielebak und Rässler, 2014, S.368)

Bei der Datenfusion hingegen werden zwei oder mehrere Datensätze miteinander verknüpft, in denen unterschiedliche Einheiten betrachtet werden. Im Vergleich zum Record Linkage werden also nicht Informationen über identische Einheiten zusammengeführt, sondern Beobachtungen möglichst ähnlicher Einheiten, die sich aber auf die gleiche Grundgesamtheit beziehen (Kiesl und Rässler, 2005, vgl. S. 18f). Zusätzlich zu der Bezeichnung Datenfusion finden sich in der Literatur weitere Begrifflichkeiten, wie „statistical matching“, „synthetical matching“ (D’Orazio et al., 2006, S.1f), „data merging“, „data matching“, „mass imputation“, „microsimulation modeling“ oder „file concatenation“ (Rässler, 2002, S.2). Im Folgenden wird der Begriff Datenfusion verwendet.

Variablentypen und Skalierungsarten

Die Einheiten, die in den jeweiligen Datensätzen betrachtet werden, enthalten Informationen, dargestellt in Form von Merkmalen (Variablen). Diese können unterschiedliche Werte, auch als Merkmalsausprägungen bezeichnet, annehmen (Fahrmeir et al., 2016, vgl. S. 13). Die konkreten Werte der Merkmalsausprägungen, die im Datensatz pro Einheit enthalten sind, werden Daten oder Beobachtungen genannt. Merkmalsausprägungen können hinsichtlich ihrer Eigenschaften unterschiedlich klassifiziert werden. Bei der Durchführung von Datenanalysen und auch hinsichtlich der Datenfusion gibt es viele verschiedene Methoden, die teilweise nur bei gewissen Variablentypen möglich oder sinnvoll sind. Das Gewicht einer Person (quantitative Variable) wird beispielsweise anders ausgewertet als das Geschlecht (qualitative Variable). Aus diesem Grund ist es wichtig, einen Überblick über die verschiedenen Variablentypen zu bekommen. Abbildung 2.1 zeigt einen Überblick über die verschiedenen Klassifizierungen der Merkmalsausprägungen.

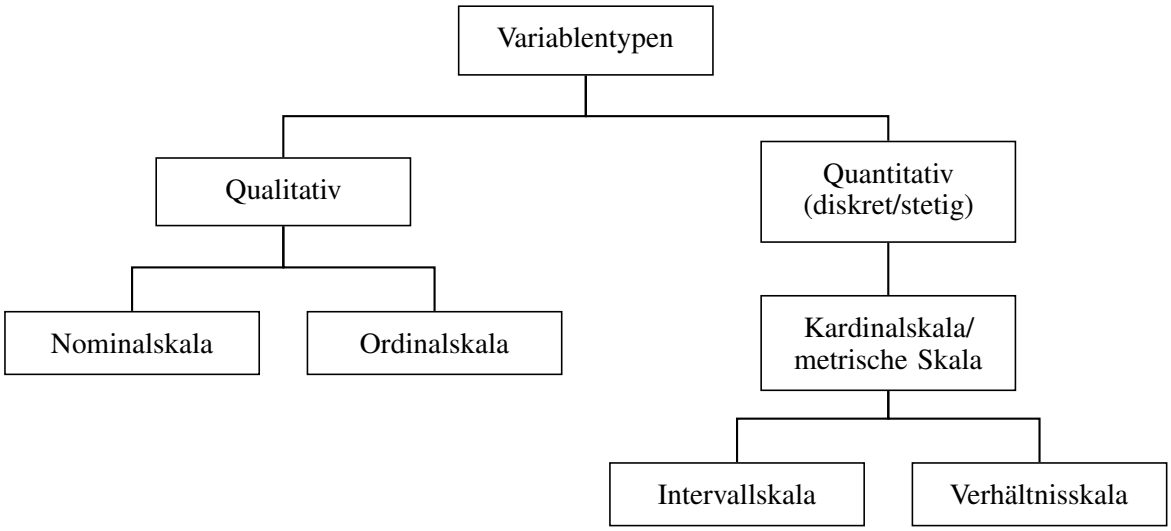


Abbildung 2.1: Merkmalstypen und Skalenniveaus von Variablen (Eigene Darstellung)

Merkmalsausprägungen von quantitativen Variablen können durch Zahlen, teilweise mit einer bestimmten Maßeinheit, dargestellt werden und lassen im Gegensatz zu qualitativen Variablen bestimmte mathematische Rechenoperationen zu (Fahrmeir et al., 2016, vgl. S.17). Quantitative Variablen lassen sich weiter in diskrete und stetige Variablen unterteilen. Diskrete Variablen besitzen eine endliche oder abzählbar unendliche Anzahl von Merkmalsausprägungen (Stocker und Steinke, 2016, vgl. S.23). Diskrete Variablen treten häufig bei Angaben von Anzahlen auf, beispielsweise Anzahl von Personen im Haushalt oder Anzahl der Unfallbeteiligten bei einem Unfall. Im Gegensatz dazu können bei stetigen Variablen die Merkmalsausprägungen theoretisch unendlich viele verschieden Werte, häufig auf ein Intervall bezogen, annehmen (Stocker und Steinke, 2016, vgl. S.23). Die Körpergröße ist beispielsweise eine typische stetige Variable, da sie innerhalb eines bestimmten Intervalls jeden beliebigen Wert annehmen kann.

Qualitative Variablen, auch kategoriale Variablen genannt, enthalten eine endliche Anzahl von Merkmalsausprägungen und lassen sich durch Namen oder Kategorien ausdrücken (Stocker und Steinke, 2016, vgl. S.23). Die Merkmalsausprägungen unterscheiden sich demnach begrifflich voneinander und nicht durch Zahlen.

Mitunter werden qualitative Variablen jedoch metrisch kodiert, beispielsweise wird aus technischen Gründen das Geschlecht „männlich“ mit dem Zahlenwert 0, das Geschlecht „weiblich“ mit dem Zahlenwert 1 ausgedrückt. Ihre Bedeutung bleibt dabei weiterhin qualitativ, quantitative Rechenoperationen lassen sich nicht sinnvoll interpretieren (Stocker und Steinke, 2016, vgl. S.23).

Variablen werden des Weiteren hinsichtlich ihres Skalenniveaus eingeteilt, das Rückschlüsse darauf gibt, welche Berechnungsmöglichkeiten zulässig und wie die Variablen zu interpretierten sind. Die Merkmalsausprägungen der nominalskalierten Variablen haben keine Wertigkeit und lassen sich lediglich miteinander vergleichen. Dabei kann ausschließlich eine Aussage darüber getroffen werden, ob sie entweder gleich oder ungleich sind. (Stocker und Steinke, 2016, vgl. S.24)
Beispiele für nominalskalierte Variablen sind Geschlecht, Orte, Namen oder Unfallarten. Die Merkmalsausprägungen von ordinalskalierten Variablen enthalten im Vergleich dazu eine natürliche Rangfolge (Stocker und Steinke, 2016, vgl. S.24). Dadurch ist es möglich, eine Aussage über die Relationen zu treffen, inwiefern die Merkmalsausprägung einer Variable beispielsweise besser/schlechter oder größer/kleiner als eine andere Merkmalsausprägung der gleichen Variable ist. Allerdings ist es nicht möglich, eine Aussage über den Abstand zu treffen. Bei der Betrachtung von Schulnoten in Deutschland ist die Note „1“ besser ist als die Note „2“, allerdings ist es nicht möglich, den Abstand zu interpretieren oder Differenzen zu bilden. Weitere Beispiele für ordinalskalierte Variablen sind Präferenzrankings, Empfinden oder Zufriedenheit.

Berechnungen hinsichtlich ihres Abstandes sind mit kardinalskalierten Variablen möglich. Kardinalskalierte Variablen werden als metrisch bezeichnet und gehören dementsprechend zu quantitativen Variablen. Sie lassen sich in intervall- und verhältnisskaliert unterteilen. Der Intervallskala gehören Variablen an, deren Merkmalsausprägungen Zahlen sind und deren Abstände zueinander berechnet werden können. Allerdings weisen sie keinen Nullpunkt auf, wodurch keine Verhältnisse abgebildet werden können. Ein Beispiel für eine Variable, die in dieses Skalenniveau eingeteilt werden kann, ist die Temperatur in Grad Celsius. Es können Differenzen zwischen Temperaturen gebildet werden, aber eine Temperatur von 20°C ist nicht doppelt so warm wie 10°C. Anders verhält es sich bei verhältnisskalilierten Merkmalsausprägungen, diese haben einen Nullpunkt und Verhältnisse können berechnet werden. Beispiele sind Größenangaben und Zeitdauern. (Fahrmeir et al., 2016, S.16)

Ausgangssituation Datenfusion

In dieser Arbeit liegt der Schwerpunkt auf der Fusion zweier Datensätze. Die Ausgangssituation kann daher durch zwei Datensätze A und B mit der zugehörigen Beobachtungsanzahl n_A und n_B beschrieben werden. Die Beobachtungen unterliegen der zentralen Annahme, dass sie unabhängig und identisch (i.i.d) verteilt und durch die wahre gemeinsame Verteilung generiert sind (D'Orazio et al., 2006, vgl. S.3). Daher ist für die praktische Umsetzung der Datenfusion eine wichtige Voraussetzung, dass die Datensätze, die meistens Stichproben sind, aus der selben Grundgesamtheit stammen (Grömping et al., 2007, vgl. S.127). Abbildung 2.2 beschreibt die allgemeine Datensituation bei der Datenfusion.

Datensatz	Y	$X_1 \ldots X_P$	Z
A	y_1	$x_{11} \ldots x_{1P}$	
	y_a	$x_{a1} \ldots x_{aP}$	
	...	...	
	y_{n_A}	$x_{n_A1} \ldots x_{n_AP}$	
B		$x_{11} \ldots x_{1P}$	z_1
		$x_{b1} \ldots x_{bP}$	z_b
		...	...
		$x_{n_B1} \ldots x_{n_BP}$	z_{n_B}

Abbildung 2.2: Schematische Darstellung der Ausgangssituation der Datenfusion (Eigene Darstellung nach D'Orazio et al., 2006, S.5)

Der zusammengefasste Datensatz aus Abbildung 2.2 wird auch als $A \cup B$ bezeichnet und enthält

- gemeinsame Variablen $\mathbf{X} = (X_1, ..., X_P)$, eine Menge von P Variablen, die in beiden Datensätzen vorkommen,
- spezifische Variablen Y und Z, die jeweils nur in Datenquelle A (Y) und in Datenquelle B (Z) vorkommen und
- fehlende Beobachtungen (grau hinterlegt).

Es ist in diesem Zusammenhang anzumerken, dass fett gedruckte Buchstaben für Vektoren stehen. Die gemeinsamen Variablen sind demnach zu einem Vektor zusammengefasst. Zudem kann ein Vektor aufgestellt werden, der die Werte der einzelnen Variablen pro Einheit zusammenfasst (Handl und Kuhlenkasper, 2017, vgl. S.15f). Demnach ergibt sich bei Datensatz A (Y, X) für jede Einheit a mit $a = 1, ..., n_A$ der Vektor

$$(y_a, \mathbf{x}_a) = (y_a, x_{a1}, ..., x_{aP})$$

und für die Werte in Datensatz B (Z, X) für jede Einheit b mit $b = 1, ..., n_B$

$$(z_b, \mathbf{x}_b) = (z_b, x_{b1}, ..., x_{bP}).$$

Allgemein könnten auch mehrere spezifische Variablen in Datensätzen vorkommen. Im Rahmen dieser Arbeit wird die Grundvoraussetzung mit jeweils einer spezifischen Variable angenommen. Diese Ausgangssituation entspricht der Situation in Abbildung 2.2.

Das allgemeine Ziel der Datenfusion ist es, die Beziehung zwischen allen Variablen **X**, Y und Z beziehungsweise zwischen den spezifischen Variablen Y und Z zu untersuchen (Meinfelder, 2013, vgl. S.84). Jedoch ist Abbildung 2.2 zu entnehmen, dass keine gemeinsamen Beobachtungen von Y und Z vorliegen. Demzufolge werden keine gleichzeitigen Informationen über die spezifischen Variablen eingeholt. Dennoch ist es möglich, anhand der gemeinsamen Variable **X** die Datensätze miteinander zu verbinden, um anschließend die gemeinsame (unbeobachtete) Verteilung $f_{\mathbf{X}YZ}(x,y,z)$ zu schätzen und die Beziehung zwischen Variablen zu analysieren, die nicht zusammen untersucht wurden. Bei diskreten Variablen wird die Verteilung durch die Wahrscheinlichkeitsfunktion $f_{\mathbf{X}YZ}(\mathbf{x},y,z) = Pr_{\mathbf{X}YZ}(\mathbf{X} = x, Y = y, Z = z)$ beschrieben, die aufgrund der endlichen Anzahl an Merkmalsausprägungen jedem Wert (x,y,z) eine konkrete Wahrscheinlichkeit zuordnet. Bei stetigen Variablen ist dies aufgrund der unendlichen Anzahl an Merkmalsausprägungen nicht möglich und es wird stattdessen von einer Dichtefunktion $f_{\mathbf{X}YZ}(x,y,z)$ gesprochen, die angibt wie das Intervall verteilt ist. Die Dichtewerte selbst sind keine Wahrscheinlichkeiten, sondern das Integral der Dichtefunktion über einem Intervall gibt die Wahrscheinlichkeiten an. (Stocker und Steinke, 2016, vgl. S. 247-251)

Ergebnis der Datenfusion

Bei der Datenfusion kann in Bezug auf das Ergebnis zwischen zwei Ansätzen unterschieden werden:

- Makroansatz und
- Mikroansatz.

Bei einem Makroansatz wird kein neuer Datensatz erstellt, sondern es werden die gemeinsame Verteilung von (Y, Z) beziehungsweise (X, Y, Z) oder beispielsweise auch Parameter, die die Beziehung zwischen den spezifischen Variablen beschreiben, geschätzt. Bei der Betrachtung von kategorialen Daten kann das zum Beispiel die Kontingenztabelle für die spezifischen Variablen sein und bei der Betrachtung von quantitativen Daten die Schätzung der Korrelation zwischen den spezifischen Variablen. (D'Orazio et al., 2006, vgl. S.2f)

Im Gegensatz dazu steht der Mikroansatz, bei dem ein vollständiger Datensatz erstellt wird. Er beinhaltet alle Variablen aus den Datensätzen A und B, die von Interesse sind. Der neu erzeugte Datensatz wird oft als „synthetisch" bezeichnet, da die spezifischen Variablen zu keinem Zeitpunkt zusammen beobachtet werden. (D'Orazio et al., 2006, vgl. S.2)
Ein wichtiger Aspekt, der bei der weiteren Analyse des synthetischen Datensatzes nicht vergessen werden darf, ist die Tatsache, dass die Ergebnisse nicht aus einer direkten Beobachtung stammen, sondern durch Informationen aus zwei oder mehreren Datensätzen zusammengetragen sind.

Bei der Erstellung eines neuen synthetischen Datensatzes (Mikroansatz) wird unterschieden, ob bei der Fusion symmetrisch oder asymmetrisch vorgegangen werden soll. Der asymmetrische Ansatz ist der Darstellung a) in Abbildung 2.3 zu entnehmen und sagt aus, dass die Fusion nur in eine Richtung stattfindet. Bei diesem Ansatz werden die spezifischen Variablen aus einem Datensatz nur um die des anderen Datensatzes ergänzt.

Bei der symmetrischen Datenfusion ($A \cup B$), vergleiche Darstellung b) in Abbildung 2.3, wird im Gegensatz dazu die spezifische Variable $\tilde{Z}$[4] im Datensatz A ergänzt und ebenso die spezifische Variable $\tilde{Y}$ in B. (Kiesl und Rässler, 2005, vgl. S.19)

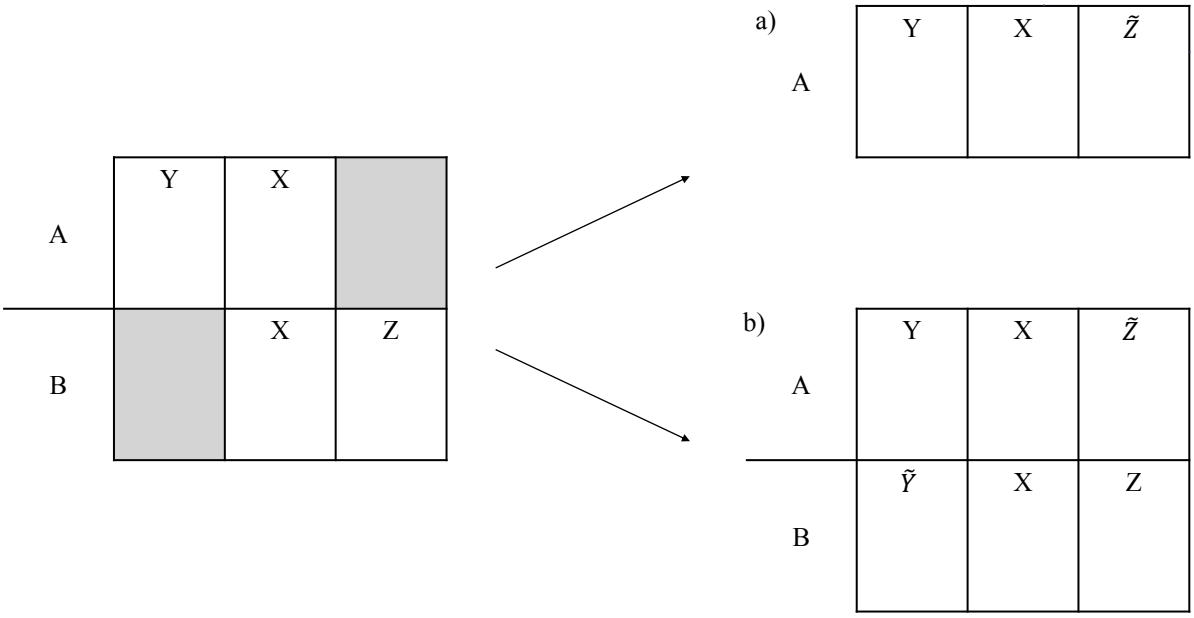


Abbildung 2.3: Asymmetrische a) und symmetrische b) Variante des Mikroansatzes: a) Ergänzung des Datensatzes A durch $\tilde{Z}$ und b) Datensatz A wird durch $\tilde{Z}$ ergänzt und Datensatz B durch $\tilde{Y}$ (Eigene Darstellung)

Der Nachteil des asymmetrischen Verfahrens ist, dass dieser ineffizienter ist, da nicht alle Beobachtungen am Ende im synthetischen Datensatz verwendet werden. Im Gegensatz dazu sorgt bei der symmetrischen Variante eine große Differenz zwischen den Datensatzgrößen für ein größeres *Matching Noise*[5], welches die Randverteilung der spezifischen Variable im kleineren Datensatz beeinflusst.

Aufgrund der fehlenden spezifischen Variablen in den Datensätzen weist die Datenfusion ein Problem fehlender Daten auf, auch genannt *Missing-Data Problem* (Cielebak und Rässler, 2014, vgl. S.371). Bei der Datenfusion gibt es jedoch einen signifikanten Unterschied zu gewöhnlichen *Missing-Data Problemen*, aufgrund der Tatsache, dass die spezifischen Variablen Y und Z nie gemeinsam beobachtet werden. In Abbildung 2.4 ist im Gegensatz dazu ein gewöhnliches Ausfallmuster zu erkennen, bei dem alle möglichen Paare an Variablen zu einem bestimmten Teil gemeinsam beobachtet werden (fehlende Werte grau hinterlegt). Somit lassen sich die Zusammenhänge direkt ableiten beziehungsweise schätzen (Kiesl und Rässler, 2005, vgl. S.19).

[4] Die Tilde kennzeichnet jeweils die fusionierten und damit künstlichen Werte.

[5] Matching Noise ist definiert als die Diskrepanz zwischen der gemeinsamen Verteilung des synthetischen Datensatzes und der zugrunde liegenden Verteilung (D'Orazio et al., 2006, vgl. S.10)

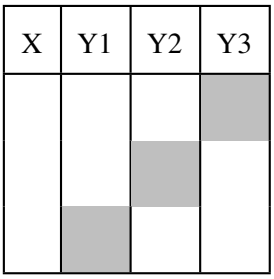


Abbildung 2.4: Herkömmliche Ausfallmuster bei fehlenden Daten (fehlende Werte sind grau hinterlegt) (Eigene Darstellung nach Kiesl und Rässler, 2005, vgl. S.20)

Diese Schätzung ist bei der Datenfusion nicht möglich, da keine Informationen über die Beziehung zwischen den spezifischen Variablen vorliegt. Dies bedeutet wiederum, dass nicht nur ein *Missing-Data Problem*, sondern auch ein sogenanntes Identifikationsproblem vorliegt. (Kiesl und Rässler, 2005, vgl. S.19f)
Um die Datensätze erfolgreich zu fusionieren, sind beide Probleme zu lösen.

Missing Data

Es können verschieden Ursachen auftreten, die zu einem *Missing-Data Problem* führen. Da der Mechanismus, der zu einem Datenausfall führt, Einfluss darauf nimmt, wie mit einem Ausfall umgegangen werden kann oder muss, ist es wichtig, diesen zu betrachten (Cielebak und Rässler, 2014, vgl. S.373). Rubin (1976) definiert dafür drei Ausfallmechanismen: "*Missing completely at random*„ (MCAR), „*Missing at random*" (MAR)und „ *Missing not at random*" (MNAR). Im Anhang A.2 ist eine genaue Beschreibung mit Beispiel der Mechanismen aufgelistet.

Werden die Daten im unvollständigen Datensatz $A \cup B$ mit $n_A + n_B$ Beobachtungen betrachtet, wobei A und B Zufallsstichproben aus derselben Grundgesamtheit sind, so kann der Datenausfall als MCAR angesehen werden (Meinfelder, 2013, S.86). Bei MCAR wird davon ausgegangen, dass die Ursache für die fehlenden Werte unabhängig von den beobachteten und nicht beobachteten Variablen **X**, Y, Z ist. Aufgrund dieser Unabhängigkeit kann der Ausfallmechanismus ignoriert und die beobachtete Stichprobenverteilung (Vergleich Abbildung 2.2) aus den Randverteilungen $f_{\mathbf{X}Y}(\mathbf{x}_a, y_a; \boldsymbol{\theta})$ und $f_{\mathbf{X}Z}(\mathbf{x}_b, z_b; \boldsymbol{\theta})$ mit $n_A + n_B$ Beobachtungen durch die Likelihood-Funktion (2.1) angegeben werden (Rässler, 2002, vgl. S. 78).

$$L(\boldsymbol{\theta}) = \prod_{a=1}^{n_A} f_{\mathbf{X}Y}(\mathbf{x}_a, y_a; \boldsymbol{\theta}) \cdot \prod_{b=1}^{n_B} f_{\mathbf{X}Z}(\mathbf{x}_b, z_b; \boldsymbol{\theta}) \tag{2.1}$$

Kann die beobachtete Strichprobenverteilung einer Verteilungsfamilie[6] zugeordnet werden, können anhand der Stichprobe die Parameter der Verteilungsfunktion geschätzt werden.

[6]Eine Verteilungsfamilie umfasst eine bestimmte Menge an ähnlichen Verteilungen, die sich hinsichtlich einer bestimmten Anzahl von Verteilungsparametern θ unterscheiden. Dementsprechend gehört zu jeder einzelnen Verteilung ein bestimmter Parameter θ beziehungsweise Parametervektor $\boldsymbol{\theta}$. Zusammengefasst gibt es pro Verteilungsfamilie einen Parameterraum Θ, der alle zulässigen Parametervektoren $\boldsymbol{\theta}$ zusammenfasst. (Stocker und Steinke, 2016, S.344f)

Dafür lässt sich beispielsweise die Likelihood-Funktion verwenden, die die Wahrscheinlichkeit (Likelihood) für das Eintreten der Parameter bei den beobachteten Daten der Stichprobe liefert (Kohn, 2005, vgl. S.346f).

Bei einem herkömmlichen Datenausfall, wie für Abbildung 2.4 beschrieben, können durch das Aufstellen der Likelihood-Funktion für die beobachteten Daten alle bivariaten Zusammenhänge zwischen den Variablen geschätzt werden (Cielebak und Rässler, 2014, vgl. S.372f). Bei der Datenfusion ist dies nicht möglich, da wie aus Gleichung 2.1 ersichtlich, anhand der Likelihood-Funktion für die beobachteten Daten keine Beziehung zwischen den spezifischen Variablen Y und Z abgeleitet werden kann (Rässler, 2002, vgl. S. 78). Grund dafür ist das Identifikationsproblem.

Identifikationsproblem

Da es aufgrund des Identifikationsproblems nicht möglich ist den direkten Zusammenhang beziehungsweise die gemeinsame Verteilung von (X, Y, Z) zu schätzen, müssen bestimmte Annahmen oder Informationen über die Zusammenhänge der spezifischen Variablen Y und Z getroffen beziehungsweise hinzugefügt werden. In der Literatur werden dafür verschiedene Ansätze vorgestellt, von denen in den folgenden Unterkapiteln drei genauer beschrieben werden.

Um einen eindeutigen Wert für die Parameter (Punktschätzer) schätzen zu können, kann die Annahme der bedingten Unabhängigkeit verwendet werden (Sims, 1972). Außerdem kann das Hinzuziehen von zusätzlichen Informationen, in Form von Datensätzen oder externen Informationen, eine weitere Möglichkeit sein, die Parameter zu schätzen (Kadane, 1978). Für den dritten Ansatz werden keine Annahmen getroffen und die Unsicherheit aufgrund des Identifikationsproblems beibehalten (Kadane, 1978). Im Vergleich zu den anderen beiden Ansätzen kann bei diesem Ansatz für die Parameter kein Punktschätzer geschätzt werden, sondern ausschließlich ein Intervallschätzer. Dieser liefert basierend auf den beobachteten Daten ein Intervall (Ober- und Untergrenze) für die unbeobachteten Parameter. (D'Orazio et al., 2006, vgl. S.9)

2.2.1 Die bedingte Unabhängigkeitsannahme (CIA)

Ein in der Literatur weit verbreiteter Ansatz ist die bedingte Unabhängigkeitsannahme, im englischen *Conditional Independence Assumption* (CIA). Als erster wies Sims auf die Annahme der bedingten Unabhängigkeit hin, die besagt, dass die spezifischen Variablen Y und Z gegeben der gemeinsamen Variablen **X** unabhängig voneinander sein müssen (Sims, 1972). Trifft diese Annahme zu, reichen die Informationen aus A und B aus, um die gemeinsame Verteilung von (**X**, Y, Z) abzuleiten. Allgemein kann die gemeinsame Verteilung anhand der Gleichung 2.2 geschätzt werden.

$$f_{\mathbf{X}YZ}(\mathbf{x}, y, z) = f_{Y|\mathbf{X}}(y|\mathbf{x}) \cdot f_{Z|\mathbf{X}}(z|\mathbf{x}) \cdot f_{\mathbf{X}}(\mathbf{x}) \tag{2.2}$$

$f_{Y|\mathbf{X}}(y|\mathbf{x})$ und $f_{Z|\mathbf{X}}(z|\mathbf{x})$ sind die bedingten Wahrscheinlichkeitsfunktionen von Y beziehungsweise Z, gegeben X, geschätzt aus den Datensätzen A/B und $f_{\mathbf{X}}(\mathbf{x})$ ist die Rand-Wahrscheinlichkeitsfunktion von X, geschätzt aus $A \cup B$. (D'Orazio et al., 2006, vgl. S.13)

Speziell für kategoriale Variablen kann die Berechnung der gemeinsamen Verteilung bei Zutreffen der CIA ebenfalls durch die Gleichung 2.3 beschrieben werden (D'Orazio, 2013, vgl. S.21).

$$Pr_{XYZ}(X=i,Y=j,Z=k) = Pr_{Y|X}(Y=j|X=i) \cdot Pr_{Z|X}(Z=k|X=i) \cdot Pr_X(X=i) \tag{2.3}$$

Pr steht für die Wahrscheinlichkeit und i, j und k sind die Kategorien der jeweiligen Variable mit $i = 1,...,I$; $j = 1,...,J$ und $k = 1,...,K$. I, J und K beschreiben demnach die Anzahl der Kategorien, der jeweiligen Variable. Aus Gleichung 2.3 kann wiederum Gleichung 2.4 abgeleitet werden, die die Wahrscheinlichkeiten für die Zellen der Kontingenztabelle von $Y \times Z$ unter der CIA durch das Summieren über die X Kategorien schätzt (D'Orazio, 2013, vgl. S.21). Somit kann die Verteilung zwischen den Variablen geschätzt werden, die nie zusammen beobachtet wurden.

$$Pr_{YZ}(Y=j,Z=k) = \sum_{i=1}^{I} Pr_{Y|X}(Y=j|X=i) \cdot Pr_{Z|X}(Z=k|X=i) \cdot Pr_X(X=i) \tag{2.4}$$

Bei realen Datensätzen kann die Annahme anhand von $A \cup B$ nicht getestet werden (D'Orazio et al., 2006, S.13). Daher muss die Verwendung der CIA anhand der Datengrundlage und von Erfahrungswerten abgeschätzt werden. Dabei ist jedoch zu beachten, wie Rodgers anhand eines Beispiels zeigt, dass bei fälschlicher Annahme der CIA, die Datenfusion zu fehlerhaften Schätzern führt (Rodgers, 1984). Trotzdem ist die Annahme der bedingten Unabhängigkeit ein Ansatz, der sowohl in früheren Durchführungen als auch bei aktuelleren Datenfusionen häufig zur Anwendung kommt, siehe zum Beispiel (Leulescu und Agafiţei, 2013) und (de Waal, 2015).

Für die Datenfusion unter der CIA gibt es verschiedene Methoden, die angewendet werden können. Dabei lassen sich die Methoden, wie bereits erwähnt, in einen Mikro- und Makroansatz unterteilen (Vergleich Kapitel 4) und darüber hinaus in parametrische und nichtparametrische Methoden. Es existieren auch gemischte Methoden, die eine Kombination beider Ansätze sind und die Vorteile der beiden Methoden vereinen.(D'Orazio et al., 2006, Kap. 2)
Lässt sich die gemeinsame Verteilung von (**X**, Y, Z) des Gesamtdatensatzes $A \cup B$ einer parametrischen Verteilungsfamilie zuordnen, kann eine parametrische Methode angewendet werden. Trifft dies jedoch nicht zu, kann die Verteilung $f_{\mathbf{X}YZ}(\mathbf{x},y,z)$ anhand von nichtparametrischen Methoden geschätzt werden.

In Tabelle 2.3 ist ein Überblick über die verschieden Methoden dargestellt, die von D'Orazio et al. (2006, vgl. Kap. 2) unter der CIA näher erläutert werden.
Im Allgemeinen wird bei der Datenfusion der Mikroansatz häufiger verwendet, da bei ihm ein realer Datensatz erzeugt wird und dadurch die späteren Analysen der Daten leichter sind. Auch wenn für nachfolgende Modellbetrachtungen echte Eingangsdaten benötigt werden, ist der Mikroansatz vorzuziehen.(D'Orazio et al., 2006, vgl. S.3)
In dieser Arbeit liegt der Fokus auf dem Mikroansatz, weshalb die Methoden des Makroansatzes nicht näher erläutert werden und auf D'Orazio et al. (2006, Kap. 2.1 und 2.3) verwiesen wird. Für die Beschreibung der Methoden mit dem Mikroansatz wird als Primärquelle D'Orazio et al. (2006, Kap. 2.2 und 2.4) verwendet.

Tabelle 2.3: Methoden der Datenfusion unter der Annahme der bedingten Unabhängigkeit

	Makroansatz	Mikroansatz
parametrisch	Maximum-Likelihood Schätzung	Maximum-Likelihood Schätzung und Conditional Mean Matching oder Random Draws
nichtparametrisch	Kern-Dichte-Schätzung kNN-Schätzer empirische Verteilungsfunktion	Random-Hot-Deck Rank-Hot-Deck Distanz-Hot-Deck

Mikroansatz mit parametrischer Methode

Wie bereits erwähnt, lässt sich bei parametrischen Methoden die gemeinsame Verteilung $f_{\mathbf{X}YZ}(\mathbf{x}, y, z)$ einer bekannten Verteilungsfamilie zuordnen. Jede dieser Verteilungen $f_{\mathbf{X}YZ}(\mathbf{x}, y, z; \boldsymbol{\theta})$ ist durch einen endlich-dimensionalen Parametervektor $\boldsymbol{\theta} = (\theta_1, ..., \theta_K) \in \Theta \subseteq \mathbb{R}^K$ charakterisiert (D'Orazio et al., 2006, vgl. S.14). Der Parametervektor wird mit in die Wahrscheinlichkeitsfunktion eingetragen, um die Abhängigkeit auszudrücken.

Grundsätzlich wird beim parametrischen Mikroansatz ein vollständiger Datensatz erzeugt. Zuerst werden dafür die Parameter $\boldsymbol{\theta}$ der gemeinsamen Verteilung $f_{\mathbf{X}YZ}(\mathbf{x}, y, z; \boldsymbol{\theta})$ geschätzt (Makroansatz) und anschließend anhand der geschätzten Parameter $\hat{\boldsymbol{\theta}}$ und der bereits vorhandenen Beobachtungen die fehlenden Werte abgeleitet (D'Orazio et al., 2006, vgl. S.25f). Daraus ergibt sich, dass die fusionierten Werte im künstlichen Datensatz nicht wirklich beobachtet, sondern aufgrund der Parameterschätzung künstlich erzeugt werden.

Die Schätzung der Parameter entspricht dem parametrischen Makroansatz, der deshalb kurz näher erläutert wird. In Anlehnung an Gleichung 2.2, kann nach Gleichung 2.5 die gemeinsame Verteilung $f_{\mathbf{X}YZ}(\mathbf{x}, y, z; \boldsymbol{\theta})$ anhand der Parametervektoren $\boldsymbol{\theta}_{Y|\mathbf{X}}$, $\boldsymbol{\theta}_{Z|\mathbf{X}}$ und $\boldsymbol{\theta}_{\mathbf{X}}$ identifiziert werden (D'Orazio et al., 2006, vgl. S.14).

$$f_{\mathbf{X}YZ}(\mathbf{x}, y, z; \boldsymbol{\theta}) = f_{Y|\mathbf{X}}(y|\mathbf{x}; \boldsymbol{\theta}_{Y|\mathbf{X}}) f_{Z|\mathbf{X}}(z|\mathbf{x}; \boldsymbol{\theta}_{Z|\mathbf{X}}) f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}_{\mathbf{X}}) \tag{2.5}$$

Diese unbekannten Parameter lassen sich anhand der Daten aus den vorhanden Datensätzen A und B beziehungsweise $A \cup B$ schätzten. Eine Schätzmethode, die in Bezug auf die Datenfusion häufig verwendet wird, ist die Maximum-Likelihood (ML) Schätzung (Leulescu und Agafiţei, 2013, vgl. S.16f). Der Vorteil besteht darin, dass sie leicht anzuwenden ist und recht gute Eigenschaften aufweist (Kohn, 2005, vgl. S.346). Die genaue Berechnung der Parameter unterscheidet sich in ihrer Abhängigkeit von der zugeordneten Verteilungsfamilie. In D'Orazio et al. (2006, vgl. S.15-25) wird die Vorgehensweise anhand der Maximum-Likelihood-Schätzung für die Normalverteilung und Multinomialverteilung ausführlich beschrieben. Des Weiteren verweisen D'Orazio et al. auf weitere Schätzmethoden, die für Punktschätzungen verwendet werden können (D'Orazio et al., 2006, vgl. S.19).

Nachfolgend kann anhand der geschätzten Parametern $\hat{\boldsymbol{\theta}}_{Y|\mathbf{X}}$, $\hat{\boldsymbol{\theta}}_{Z|\mathbf{X}}$ und $\hat{\boldsymbol{\theta}}_{\mathbf{X}}$ die gesamte Verteilung von ($\mathbf{X}$, Y, Z) geschätzt oder im Mikroansatz für die Erstellung des künstlichen Datensatzes verwendet werden.

Der parametrische Mikroansatz gliedert sich in zwei mögliche Ansätze: *Conditional mean matching* und *Draw based on conditional predictive distributions* (D'Orazio et al., 2006, vgl. S.26). Beispielhaft wird nachfolgend das *Conditional mean matching* kurz vorgestellt. Für Erklärungen wird im weiteren Verlauf der Arbeit der asymmetrische Fall betrachtet, bei dem Datensatz A durch Z ergänzt wird. Analog funktioniert die Fusion von Y in B.

Beim *Conditional mean matching* wird jede fehlende Beobachtung $\tilde{z}_a$ mit $a = 1, ..., n_A$ im Datensatz A nach Gleichung 2.6 durch den Erwartungswert der fehlenden Variable gegeben der beobachteten Werte x ersetzt (D'Orazio et al., 2006, vgl. S.26).

$$\tilde{z}_a = E(Z|\mathbf{X} = \mathbf{x}_a) \tag{2.6}$$

Für eine multivariate Normalverteilung ist das Verfahren gleichzusetzen mit einer Regressionsimputation (D'Orazio et al., 2006, vgl. S.26). Die fehlenden Werte $\tilde{z}_a$ mit $a = 1, ..., n_A$ im Datensatz ergeben sich nach Gleichung 2.7 durch die geschätzte Regressionsfunktion.

$$\tilde{z}_a = \hat{\alpha}_Z + \hat{\beta}_{ZX}\, x_a \tag{2.7}$$

Dabei sind $\hat{\alpha}_Z$ und $\hat{\beta}_{ZX}$ Regressionsparameter, die mittels der Maximum-Likelihood-Schätzung aus dem Datensatz B geschätzt werden und x_a jeweils der beobachtete Wert der gemeinsamen Variable X aus Datensatz A. Für eine Vertiefung des Mikroansatzes mit parametrischer Methode sei auf D'Orazio et al. (2006, Kap. 2.2) verwiesen.

Mikroansatz mit nichtparametrischer Methode

Es gibt viele verschiedene Verfahren, die nicht den parametrischen Methoden zugeordnet werden und das Mikro-Ziel befolgen, einen vollständigen Datensatz zu produzieren. Eine weitverbreitete Möglichkeit in Bezug auf die Datenfusion ist die Verwendung von Hot-Deck Verfahren. Bei diesen Verfahren werden fehlende Beobachtungen aus dem einen Datensatz durch reale Beobachtungen aus dem anderen Datensatz ersetzt (D'Orazio et al., 2006, vgl. S.35). In Bezug auf die Datenfusion gibt es drei verschiedene Hot-Deck Verfahren (D'Orazio et al., 2006, vgl. S.37):

- *Distanz-Hot-Deck*
- *Random-Hot-Deck*
- *Rank-Hot-Deck.*

Hot-Deck Verfahren werden auch bei herkömmlichen Datenimputationen eingesetzt und ersetzten in diesem Zusammenhang Beobachtungen durch reale Beobachtungen aus dem gleichen Datensatz.

Da bei der Datenfusion unter Verwendung von Hot-Deck Verfahren die fehlenden Beobachtungen aus einem anderen Datensatz fusioniert werden, muss zuvor den Datensätzen meistens eine bestimmte Rolle zugewiesen werden. Ein Datensatz wird der Empfänger und erhält für die nicht beobachtete Variable reale Beobachtungen aus dem anderen Datensatz, dem Spender. Allerdings gibt es eine Ausnahme bei der symmetrischen Methode. In diesem Fall nimmt jeder der Datensätze sowohl die Rolle des Empfängers als auch die Rolle des Spenders ein. Die Zuordnung der Datensätze wird meistens anhand der Datengröße entschieden. Dabei ist es üblich, den größeren Datensatz als Spender zu verwenden, da je größer der Spender, desto genauer kann die Verteilung der fusionierten Variable gegeben **X** geschätzt werden (D'Orazio et al., 2006, vgl. S. 36). Nimmt dagegen der kleinere Datensatz die Rolle des Spenders ein, werden einige Beobachtungen in dem Spenderdatensatz mehr als einmal in den Empfängerdatensatz imputiert, was in einer künstlichen Veränderung der Variabilität der Verteilung im fusionierten Datensatz resultiert (D'Orazio et al., 2006, vgl. S.35). Daraus wird ersichtlich, dass bei einer symmetrischen Variante die beiden Datensätze eine ähnliche Größe aufweisen sollten. Im weiteren Verlauf dieser Arbeit nimmt Datensatz A mit der spezifischen Variable Y die Rolle des Empfängers ein und Datensatz B mit der spezifischen Variable Z die Rolle des Spenders.

Distanz-Hot-Deck Verfahren

Beim *Distanz-Hot-Deck* Verfahren werden die Beobachtungen aus dem Spendersatz mit der Empfängerbeobachtung verbunden, der sie am ähnlichsten sind. Die Ähnlichkeit wird über ein Distanzmaß berechnet, das jeweils die Distanzen zwischen den gemeinsamen Variablen **X** bestimmt (D'Orazio et al., 2006, vgl. S.41). Je ähnlicher sich die Beobachtungen sind, desto kleiner sind die Werte des Distanzmaßes. Für die Einheit a im Empfängerdatensatz A wird die Beobachtung z aus dem Spenderdatensatz B fusioniert, bei der nach Gleichung 2.8 die Einheit b ($1 \leq b \leq n_B$) so ausgewählt wird, dass die Distanz d_{ab} zwischen der a-ten Beobachtung der gemeinsamen Variable x_a im Datensatz A und der Beobachtung der gemeinsamen Variable im Datensatz B x_b minimal wird.

$$d_{ab} = \min_{1 \leq b \leq n_B} |x_a^A - x_b^B| \tag{2.8}$$

Weisen mehrere Beobachtungen im Spenderdatensatz für die a-te Beobachtung die kleinste Distanz auf, wird zufällige eine Beobachtung ausgewählt. In der Literatur sind unterschiedliche Distanzmaße zu finden. Welches Distanzmaß für die Datenfusion verwendet wird, muss abhängig von den vorhandenen Daten und Variablentypen (Vergleich Abbildung 2.1) der Variablen im Datensatz entschieden werden.

In Bezug auf die Datenfusion werden für quantitative Variablen unter anderem die folgenden drei Distanzmaße verwendet (D'Orazio et al., 2006, vgl. S.215f):

- Manhattan-Distanz

$$d_{ab}^M = \sum_{p=1}^{P} |x_{ap} - x_{bp}| \tag{2.9}$$

- Euklidische Distanz

$$d_{ab}^{E} = \sqrt{\sum_{p=1}^{P} (x_{ap} - x_{bp})^2} \tag{2.10}$$

- Mahalanobis-Distanz

$$d_{ab}^{Ml} = (\mathbf{x}_a - \mathbf{x}_b)^T \sum_{\mathbf{XX}}^{-1} (\mathbf{x}_a - \mathbf{x}_b). \tag{2.11}$$

Die Gleichungen 2.9 - 2.11 sind hier für P gemeinsame X Variablen aufgelistet ($\mathbf{X} = (X_1, ..., X_P)$) mit den Beobachtungen x_{ap} aus dem Datensatz A und x_{bp} aus dem Datensatz B ($p = 1, ..., P$). Bei der Mahalanobis-Distanz (Gleichung 2.11) ist $\sum_{\mathbf{XX}}^{-1}$ die inverse Matrix der Kovarianzmatrix von $\mathbf{X}$.

Die Erklärung der Euklidischen und Manhattan-Distanz werden anhand eines Beispiels näher erläutert. In Abbildung 2.5 sind zwei Beispieldatensätze A und B dargestellt. Beide bestehen jeweils aus einer Einheit mit zwei gemeinsamen Variablen (X_1 =Alter von Kind 1 und X_2 = Alter von Kind 2).

	X_1=Alter Kind 1	X_2=Alter Kind 2
Datensatz A	$x_{a1} = 4$	$x_{a2} = 4$
Datensatz B	$x_{b1} = 8$	$x_{b2} = 7$

Abbildung 2.5: Beispieldatensatz A und B mit quantitativen Variablen

Von Interesse ist die Distanz zwischen der Beobachtung A und der Beobachtung B. Aufgrund von nur zwei Variablen, lassen sich die beiden Beobachtungen in ein Diagramm übertragen (Vergleich Abbildung 2.6).

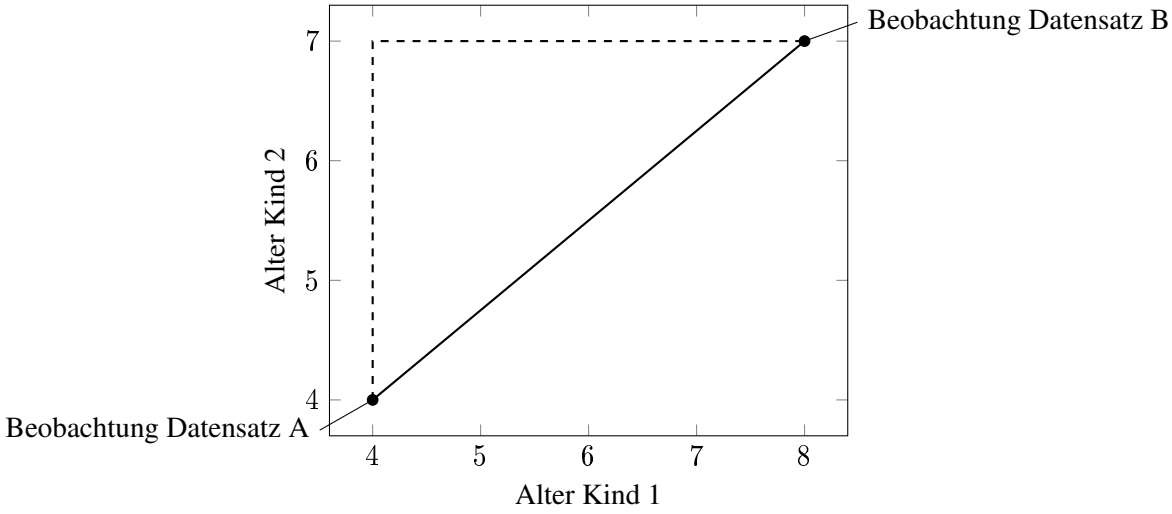


Abbildung 2.6: Euklidische Distanz (durchgängige Linie) und Manhattan-Distanz (gestrichelt) im zweidimensionalen Raum (Eigene Darstellung nach Handl und Kuhlenkasper (2017, vgl. S.94))

Punkt A gibt die Beobachtung aus dem Datensatz A wieder, bei dem sowohl Kind 1 als auch Kind 2 vier Jahre alt sind und Punkt B die Beobachtung aus dem Datensatz B, bei dem das erste Kind acht und das zweite Kind sieben Jahre alt ist. Die durchgängige Linie gibt den direkten und damit kürzesten Abstand zwischen den zwei Punkten wieder. Berechnet werden kann nach Gleichung 2.12 der Abstand d_{ab}^{E} mit Hilfe des Satzes des Pythagoras, wobei die Katheten (Vergleich Abbildung 2.6 gestrichelte Linien) die Abstände zwischen den Werten der gleichen Variable, aber aus den unterschiedlichen Datensätzen, sind (Handl und Kuhlenkasper, 2017, vgl. S. 94).

$$d_{ab}^{E} = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2} \tag{2.12}$$

$$= \sqrt{(4-8)^2 + (4-7)^2} = 5 \tag{2.13}$$

Die kürzeste Distanz zwischen den Beobachtungen ist demnach 5 Jahre. Wird die Gleichung 2.12 für höherdimensionale Räume, also für P Variablen umgestellt, ergibt sich Gleichung 2.10 (Handl und Kuhlenkasper, 2017, vgl. S.95). Dementsprechend entspricht die grünen Linie in Abbildung 2.6 der Euklidischen Distanz.

Da der direkte Abstand nicht immer der sinnvollste ist, gibt es zusätzlich die Manhattan-Distanz. Werden zwei beliebige Punkte in eine Stadt mit Schachbrettmuster (beispielsweise wie in Manhattan) übertragen, ist der direkte Weg nicht immer möglich. Es wird daher der Weg um die Häuser herum ausgewählt, der die Manhattan-Distanz beschreibt (Vergleiche beispielsweise Abbildung 2.6 gestrichelte Linien). (Handl und Kuhlenkasper, 2017, vgl. S.102)
Beide Linie ergeben zusammen ein Dreieck, bei dem die euklidische Distanz die Länge der Hypotenuse ist und die Manhattan-Distanz d_{ab}^{M} nach Gleichung 2.14 die Summe der Längen der beiden Katheten, die wiederum der Abstand zwischen den Beobachtungen der jeweiligen Variable sind.

$$d_{ab}^{M} = |x_{a1} - x_{b1}| + |x_{a2} - x_{b2}| = |4-8| + |4-7| = 7 \tag{2.14}$$

Die Gleichung 2.14 kann ebenfalls für mehr als zwei Variablen (P Variablen) umgestellt werden und ergibt anschließend die Gleichung 2.9.

Die Mahalanobis-Distanz kann nicht in das Diagramm mit eingezeichnet werden, denn sie berücksichtigt zusätzlich die Korrelation (Zusammenhang) zwischen den Daten, da die Kovarianzmatrix in die Gleichung mit einbezogen wird. Allerdings ist die Mahalanobis-Distanz mit der Euklidischen-Distanz verwandt. Entspricht die Kovarianzmatrix der Einheitsmatrix, die Daten besitzen demnach eine Varianz von 1 und weisen die gleiche Standardabweichungen auf, liegen unkorrelierte Variablen vor. Die Mahalanobis-Distanz entspricht dann der Euklidischen-Distanz. (Joenssen, 2015, vgl. S. 81)

Bei der Betrachtung von qualitativen Daten wird meistens nicht direkt ein Distanzmaß berechnet, sondern ein Ähnlichkeitsmaß, das anschließend in eine Distanz umgerechnet wird. Ähnlichkeitsmaße werden im Gegensatz zum Distanzmaß größer, je ähnlicher sich die Beobachtungen sind.

Aus diesem Grund kann ein Ähnlichkeitsmaß, dass bei qualitativen Variablen zwischen 0 und 1 normiert ist, anhand der Gleichung 2.15 in ein Distanzmaß umgerechnet werden, indem der Ähnlichkeitskoeffizient s_{ab} von 1 abgezogen wird. (Kaufmann und Pape, 1996, vgl. S. 442)

$$d_{ab} = 1 - s_{ab} \tag{2.15}$$

Aufgrund der Eigenschaften von qualitativen Merkmalen ist es nicht direkt möglich, die Distanz zu berechnen, sondern es wird überprüft, ob die Merkmalsausprägungen übereinstimmen und sich somit die Beobachtungen ähnlich oder unähnlich sind. Des Weiteren wird bei qualitativen Variablen zwischen binären und Variablen mit mehr als zwei Merkmalsausprägungen, und bei den binären darüber hinaus noch zwischen asymmetrischen und symmetrischen Variablen unterschieden (Handl und Kuhlenkasper, 2017, vgl. S. 103).

Bei asymmetrischen Variablen sind zwei Beobachtungen nicht unbedingt ähnlich, auch wenn beide Beobachtungen die Eigenschaften, die von Interesse sind, nicht besitzen. Bedeutet beispielsweise die Merkmalsausprägung 1 Krankheit vorhanden und 0 Krankheit nicht vorhanden, so bedeutet der Vergleich von zwei Beobachtungen mit Kategorie 0 nicht zwangsläufig, dass sich die Beobachtungen ähnlich sind. Beziehungsweise für die weitere Analyse ist meistens nur das Vorhandensein der Krankheit von Interesse. Aus diesem Grund wird bei der Berechnung die Übereinstimmung der Kategorie 0 nicht mit betrachtet. Bei symmetrischen Variablen sind dagegen die Beobachtungen ähnlich, auch wenn sie beide Eigenschaften nicht besitzen. (Handl und Kuhlenkasper, 2017, vgl. S. 103)

Für die Berechnung der Ähnlichkeitskoeffizienten kann die Tabelle 2.4 zu Hilfe genommen werden.

a \ b	1	0	
1	k	l	k+l
0	m	o	m+o
	k+m	l+o	k+l+m+o

Tabelle 2.4: Hilfstabelle zur Berechnung von Ähnlichkeitskoeffizienten zwischen zwei Beobachtung mit binären Variablen

In die Tabelle wird jeweils die Anzahl der Merkmalsausprägungen eingetragen, die übereinstimmen (Zelle k oder o) und nicht übereinstimmen (Zelle l oder m). Das Vorgehen wird beispielhaft an den Datensätzen aus Abbildung 2.7 gezeigt. Die Datensätze bestehen jeweils aus einer beobachteten Einheit mit vier binären Variablen.

	Geschlecht (0=männlich, 1=weiblich)	Fahrzeug fahrbereit (0=Ja, 1=Nein)	Ortslage (0=Innerorts, 1=Außerorts)	Lichtverhältnis (0=Tag, 1=Nacht)
Datensatz A	0	1	0	0
Datensatz B	0	1	1	0

Abbildung 2.7: Beispieldatensatz A und B mit binären Variablen

Beim Vergleich der Beobachtungen tritt bei der Variable *Fahrzeug fahrbereit* jeweils die Kategorie 1 auf, daher wird in die Hilfstabelle 2.5 eine 1 in die k-Zelle eingetragen. Dass der Datensatz A eine 1 aufweist und der Datensatz B dagegen eine 0 kommt bei keinem Vergleich vor, daher wird in die l-Zelle der Wert 0 eingetragen. Allerdings tritt der Gegensatz auf und Datensatz A weist bei der Variable *Ortslage* eine 0 und Datensatz B im Vergleich dazu eine 1 auf. Dementsprechend wird in die m-Zelle eine 1 eingetragen. In die o-Zelle wird eine 2 eingetragen, da bei zwei Variablen jeweils die 0 beobachtet wird.

a \ b	1	0	
1	k=1	l=0	1
0	m=1	o=2	3
	2	2	4

Tabelle 2.5: Hilfstabelle zur Berechnung von Ähnlichkeitskoeffizienten zwischen zwei Beobachtung mit binären Variablen

Anschließend können mit Hilfe der Hilfstabelle die Ähnlichkeitskoeffizienten berechnet werden. Die Koeffizienten werden in symmetrische und asymmetrischen Variablen unterteilt. Bei asymmetrischen Variablen wird der Wert der o-Zelle nicht mit einberechnet. Zusätzlich erfolgt eine Unterscheidung bei der Wichtung von Übereinstimmung und Nichtübereinstimmung. Die Übereinstimmung setzt sich für symmetrische Variablen aus den Werten der Zellen k und o zusammen und für asymmetrische Variablen aus der Zelle o. Die Nichtübereinstimmung ist in beiden Fällen aus den Werten der Zellen l und m zusammengesetzt. (Handl und Kuhlenkasper, 2017, S.105f)

Zu den symmetrischen Ähnlichkeitskoeffizienten gehört der Simple-Matching-Koeffizient, der sich nach Gleichung 2.16 aus dem Quotienten der Anzahl der Übereinstimmung zur Gesamtanzahl der Variablen ergibt (Handl und Kuhlenkasper, 2017, vgl. S. 105).

$$s_{ab}^{SM} = \frac{k+o}{k+l+m+o} \tag{2.16}$$

Für die Hilfstabelle 2.5 mit den Werten zum Beispieldatensatz aus Abbildung 2.7 wird der Simple-Matching-Koeffizient anhand der Gleichung 2.16 wie folgt berechnet:

$$s_{ab}^{SM} = \frac{1+2}{1+0+1+2} = 0,75.$$

Daraus ergibt sich die Distanz nach Gleichung 2.15:

$$d_{ab} = 1 - 0,75 = 0,25.$$

Der Rogers & Tanimoto-Koeffizient wichtet im Vergleich zum Simple-Matching-Koeffizient die Nichtübereinstimmung höher, indem, wie aus Gleichung 2.17 ersichtlich, die Anzahl der Nichtübereinstimmungen im Divisor doppelt gezählt wird (Handl und Kuhlenkasper, 2017, vgl. S.105).

$$s_{ab}^{RT} = \frac{k+o}{k+o+2(l+m)} \tag{2.17}$$

In Analogie zu den symmetrischen Variablen sind die folgenden Ähnlichekeitskoeffizienten für asymmetrische Variablen ausgelegt. Identisch zum Simple-Matching-Koeffizient wichtet der Jaccard-Koeffizient die Übereinstimmung und Nichtübereinstimmung gleich. Ohne Betrachtung der Werte aus der o-Zelle, ergibt sich nach Gleichung 2.18 der Koeffizient aus dem Quotienten der Übereinstimmungen zur Gesamtanzahl der Nichtübereinstimmungen. (Handl und Kuhlenkasper, 2017, vgl. S.106)

$$s_{ab}^{J} = \frac{k}{k+l+m} \tag{2.18}$$

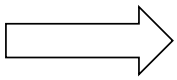
Der Dice-Koeffizient wichtet nach Gleichung 2.19 die Übereinstimmungen im Vergleich zu den Nichtübereinstimmungen doppelt so hoch (Kaufmann und Pape, 1996, vgl. S.445).

$$s_{ab}^{D} = \frac{2k}{2k+l+m} \tag{2.19}$$

Es gibt noch weitere Koeffizienten, auf die nicht eingegangen wird. Sie sind nach dem gleichen Prinzip wie die beschriebenen Koeffizienten aufgebaut, allerdings werden andere Wichtungen mit einbezogen (Beispielsweise in Kaufmann und Pape, 1996, vgl. S. 444f).

Für den Vergleich von Beobachtungen mit qualitativen Variablen mit mehr als zwei Merkmalsausprägungen ist es nicht möglich, die Ähnlichkeitskoeffizienten auf die eben genannte Weise zu berechnen. Es ist allerdings möglich, die Variablen in binäre Variablen zu codieren und anschließend die Koeffizienten zu berechnen. Bei der Umwandlung in sogenannte Dummy-Variablen werden die einzelnen Kategorien in eine Variable umgewandelt, bei der 1 dafür steht, dass die Kategorie auftritt und 0 für das nicht auftreten. In Abbildung 2.8 ist der Schritt der Umwandlung beispielhaft für die Variable „Lichtverhältnis“ dargestellt. Aus einer Variable mit 3 Merkmalsausprägungen (Kategorien) entstehen drei neue binäre Variablen. Auf diese Weise wird zum Beispiel in Leulescu und Agafiţei (2013, S.37f) für kategoriale Variablen der Dice-Koeffizient, mit Umwandlung in eine Distanz, als Maß für die Hot-Deck Methode verwendet.

Lichtverhältnis
1: Tageslicht
2: Dämmerung
3: Dunkelheit



L1	L2	L3
1	0	0
0	1	0
0	0	1

Abbildung 2.8: Beispiel Dummycodierung

Eine weitere Möglichkeit ist die Umwandlung der Variablen in numerische (quantitative) Variablen und die Verwendung der Distanzmaße für quantitative Variablen (Vergleich Gleichung 2.9 - 2.11) (D'Orazio et al., 2006, vgl. S. 216). Allerdings ist diese Betrachtung nicht bei allen Variablen sinnvoll und wird daher seltener verwendet als die Umwandlung in binäre Variablen, da bei diesem Ansatz die Variablen trotzdem als qualitative Variablen behandelt werden.

Die letzte Betrachtung von Distanzmaßen berücksichtigt gemischte Variablentypen. Zum einen ist eine Umcodierung möglich, damit alle Variablen einem Variablentypen entsprechen. Zum anderen kann nach Gleichung 2.20 die Gower-Distanz d_{ab}^{G} berechnet werden, die die unterschiedlichen Variablentypen mit einschließt (Handl und Kuhlenkasper, 2017, vgl. S.107f). Die Distanz kann jedoch auch für nur einen Variablentyp verwendet werden.

$$d_{ab}^{G} = \frac{\sum_{p=1}^{P} \delta_{ab}^{(p)} d_{ab}^{(p)}}{\sum_{p=1}^{P} \delta_{ab}^{(k)}} \tag{2.20}$$

In Abhängigkeit vom Typ der Variable p wird die Distanz $d_{ab}^{(p)}$ zwischen den beiden Beobachtungen a und b mit den Merkmalsausprägungen x_{ap} und x_{bp} berechnet.

- Für nominalskalierte Variablen beträgt nach Gleichung 2.21 der Abstand 0, wenn die Merkmalsausprägungen übereinstimmen und ansonsten 1:

$$d_{ab}^{(p)} = \begin{cases} 1, & \text{wenn } x_{ap} \neq x_{bp} \\ 0, & \text{wenn } x_{ap} = x_{bp} \end{cases} \tag{2.21}$$

- für quantitative und ordinalskalierte Variablen wird der Abstand nach Gleichung 2.22 aus dem Betrag der Merkmalsausprägungen geteilt durch die Spannweite der Merkmalsausprägungen beider Datensätze berechnet:

$$d_{ab}^{(p)} = \frac{|x_{ap} - x_{bp}|}{R_p}, \quad \text{mit } R_p = max(x_{ap}, x_{bp}) - min(x_{ap}, x_{bp}) \tag{2.22}$$

Mit δ_{ab} wird die Symmetrie binärer Variablen berücksichtigt. Handelt es sich um asymmetrische Variablen ist $\delta_{ab} = 0$, falls beide Beobachtungen den Wert 0 annehmen und sonst 1. (Handl und Kuhlenkasper, 2017, vgl. S. 107f)

Tabelle 2.9 beinhaltet eine Kombination aus den zuvor beschriebenen Beispieldatensätzen mit unterschiedlichen Variablentypen.

	Alter Kind	Geschlecht (0=männlich, 1=weiblich)	Fahrzeug fahrbereit (0=Ja, 1=Nein)	Lichtverhältnis (1=Tageslicht, 2=Dämmerung, 3=Dunkelheit)
Datensatz A	4	0	1	0
Datensatz B	8	0	1	1

Abbildung 2.9: Beispieldatensatz A und B mit quantitativen Variablen

Anhand dieser beiden Beispieldatensätze mit $P = 4$ Variablen (binäre Variablen sind symmetrisch) ergibt sich für die Gower-Distanz nach Gleichung 2.20 mit folgender Berechnung eine Distanz von 0,5.

$$d_{ab}^{G} = \frac{\left(\delta_{ab}^{(1)} \times \frac{|x_{ap} - x_{b1}|}{\max\limits_{a} x_{a1} - \min\limits_{b} x_{bp}}\right) + \left(\delta_{ab}^{(2)} \times d_{ab}^{(2)}\right) + \left(\delta_{ab}^{(3)} \times d_{ab}^{(3)}\right) + \left(\delta_{ab}^{(4)} \times d_{ab}^{(4)}\right)}{\delta_{ab}^{(1)} + \delta_{ab}^{(2)} + \delta_{ab}^{(3)} + \delta_{ab}^{(4)}}$$

$$= \frac{(1 \times \frac{|4-8|}{8-4}) + (1 \times 0) + (1 \times 0) + (1 \times 1)}{1+1+1+1} = 0,5$$

Insgesamt ist festzustellen, dass ein passendes Distanzmaß dem Variablentyp sowie der Definition der Eigenschaften der Variablen entsprechend ausgewählt werden muss.

Nach den Erläuterungen zur Berechnung des Distanzmaßes erfordert das *Distanz-Hot-Deck* Verfahren eine Auseinandersetzung mit den zwei Methoden dieses Verfahrens (D'Orazio et al., 2006, S.41f):

- unbeschränkte Methode und
- beschränkte Methode.

Fällt die Wahl auf eine Beobachtung, die der Beobachtung im Empfängerdatensatz am ähnlichsten ist, und kann diese Beobachtung anschließend erneut als Spender ausgewählt werden, so wird von einem unbeschränkten *Distanz-Hot-Deck* Verfahren gesprochen. Mit dieser Methode können immer die Beobachtungen fusioniert werden, die sich am ähnlichsten sind und somit die kleinsten Distanzen zueinander aufweisen. (D'Orazio et al., 2006, S.41f)

Das Gegenteil tritt bei der beschränkten Methode ein. Jede Beobachtung im Spenderdatensatz kann nur einmal ausgewählt werden.

Eine Voraussetzung für diese Methode ist folglich, dass der Spenderdatensatz mindestens genau so groß sein muss wie der Empfängerdatensatz ($n_B \geq n_A$) (D'Orazio et al., 2006, vgl. S.42). Kann jede Beobachtung nur einmal als Spender verwendet werden und die Auswahl der Beobachtungen findet der Reihenfolge nach statt, so kann es zu folgender Situation kommen: Wird eine Beobachtung mit dem kleinsten Abstand ausgewählt, kann es dazu führen, dass am Ende nur noch Spenderbeobachtungen mit einem sehr großen Abstand zur Verfügung stehen. Die Qualität der Fusionsergebnisse werden dadurch negativ beeinflusst. Daher ist es sinnvoll, die Gesamtdistanz mit einzubeziehen und die Beobachtungen so zuzuordnen, dass die Gesamtdistanz zwischen allen Empfängern und ihren Spendern minimal wird. Dieses Problem wird als lineares Zuordnungsproblem bezeichnet und gehört zu den Optimierungsproblemen. (D'Orazio et al., 2006, vgl. S.42)

Für die Zuordnung der Beobachtungen wird eine Zuordnungsvariable eingeführt, die nach Gleichung 2.23 die Werte 0 oder 1 annehmen kann. Mit $w_{ab} = 1$ wird signalisiert, dass das Beobachtungspaar (a, b) für die Fusion ausgewählt ist. Wenn dies nicht der Fall ist, nimmt w_{ab} den Wert 0 an.

$$w_{ab} = \begin{cases} 1 \text{ wenn das Beobachtungspaar } (a,b) \text{ ausgewählt ist} \\ 0 \text{ sonst} \end{cases} \tag{2.23}$$

Ziel ist, die Zuordnungsvariable w_{ab} so zu bestimmen, dass die Gesamtdistanz entsprechend Gleichung 2.24 minimal wird.

$$\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (d_{ab} w_{ab}) \rightarrow min \tag{2.24}$$

Allerdings müssen für die Minimierung der Gesamtdistanz Nebenbedingungen aufgestellt werden. Bei gleich großer Beobachtungsanzahl im Spender- und Empfängerdatensatz ($n_B = n_A$) sind die Nebenbedingungen so definiert, dass jede Beobachtung a aus dem Empfängerdatensatz verwendet (Gleichung 2.25) und jede Spenderbeobachtung b einer Empfängerbeobachtung a zugeordnet werden muss (Gleichung 2.26). (D'Orazio et al., 2006, vgl. S.42)

$$\sum_{b=1}^{n_B} w_{ab} = 1, \qquad a = 1, ..., n_A \tag{2.25}$$

$$\sum_{a=1}^{n_A} w_{ab} = 1, \qquad b = 1, ..., n_B \tag{2.26}$$

$$w_{ab} \in \{0; 1\}$$

Ist der Spenderdatensatz größer als der Empfängerdatensatz, müssen die Nebenbedingungen hinsichtlich der Beobachtungen im Spenderdatensatz anders definiert werden. Nicht jeder Spenderbeobachtung b wird eine Empfängerbeobachtung zugeordnet.

Dementsprechend gibt es auch Beobachtungen im Spenderdatensatz, die für die Fusion nicht verwendet werden (Gleichung 2.28). Aber dennoch gilt, dass jede Empfängerbeobachtung einmal zugeordnet werden muss (Gleichung2.27). (D'Orazio et al., 2006, vgl. S.42)

$$\sum_{b=1}^{n_B} w_{ab} = 1, a = 1, ..., n_A \tag{2.27}$$

$$\sum_{a=1}^{n_A} w_{ab} \leq 1, b = 1, ..., n_B \tag{2.28}$$

$$w_{ab} \in \{0; 1\} \tag{2.29}$$

Allgemein bedeutet dies, dass in jeder Zeile nur eine Beobachtung und in jeder Spalte abhängig von der Spendergröße eine oder keine Beobachtung ausgewählt werden kann. Die Distanzen zwischen den Beobachtungen im Empfänger- und Spenderdatensatz müssen hierfür in einer Matrix dargestellt sein (Vergleich Abbildung 2.10).

$a \backslash b$	1	...	n_B
1	d_{11}	...	$d_1 n_B$
...	...	...	...
n_A	$d_{n_A 1}$	...	$d_{n_A n_B}$

Abbildung 2.10: Schematische Darstellung der Distanzen zwischen den Beobachtungen im Empfängerdatensatz A und Spenderdatensatz B

Aus Abbildung 2.10 kann abgeleitet werden, wie groß die Matrix bei Datensätzen mit vielen Beobachtungen wird. Demzufolge benötigt die Lösung der optimalen Zuordnung einen großen Rechenaufwand (D'Orazio et al., 2006, S.42). Bereits bei einer Größe von jeweils 10 Beobachtungen in beiden Datensätzen existieren über 3,6 Millionen Kombinationsmöglichkeiten, die auf ihre Gesamtdistanz überprüft werden müssten. Für die Lösung des Zuordnungsproblems können klassische Lösungsalgorithmen für Transportprobleme, wie beispielsweise die Spaltenminimum-Methode, die Vogelsche-Approximationsmethode oder die MODI-Methode verwendet werden (Joenssen, 2015, S.113f). Da diese jedoch nicht speziell auf das lineare Zuordnungsproblem angepasst sind, werden sie in dieser Arbeit nicht näher erläutert. Eine weitere Lösung ist die Ungarische Methode (Kuhn, 1955). Dieser Algorithmus ist speziell auf das beschriebene lineare Zuordnungsproblem ausgelegt und somit deutlich schneller. Die genaue Vorgehensweise der Ungarischen Methode ist an einem Beispiel im Anhang A.3 beschrieben.

Ersichtlich wird, dass ein Nachteil des beschränkten *Distanz-Hot-Deck* Verfahrens gegenüber dem unbeschränkten Ansatz die hohe erforderliche Rechenleistung ist. Zusätzlich ist, durch die Zuordnung der Beobachtungen anhand der minimalen Gesamtdistanz, die durchschnittliche Distanz zwischen den Spender- und Empfängerbeobachtungen größer als bei der unbeschränkten Fusion. Dies führt dazu, dass nicht immer die ähnlichste Beobachtung fusioniert wird.

Andererseits ist es möglich, dass die unbeschränkte Methode zu einer Überbenutzung vereinzelter Spenderbeobachtungen führt. Demzufolge bleibt bei der beschränkten Methode eine bessere Nachbildung der Randverteilungen der spezifischen Variable Z (fusionierte Variable) erhalten. (Rodgers, 1984, S.93) (D'Orazio et al., 2006, vgl. S.43) Insgesamt hängt die Fusion beim *Distanz-Hot-Deck* Verfahren von den gemeinsamen Variablen **X** ab, weshalb ein hoher Zusammenhang zwischen den gemeinsamen Variablen und der spezifischen Variable Z zu besseren Ergebnissen führt (D'Orazio et al., 2006, vgl. S.1677f).

Random-Hot-Deck Verfahren

Anders als beim *Distanz-Hot-Deck* Verfahren werden beim *Random-Hot-Deck* Verfahren die fehlenden Beobachtungen im Empfängerdatensatz mit einer zufälligen Beobachtung aus dem Spenderdatensatz verbunden (D'Orazio et al., 2006, vgl. S.37). Daraus ergibt sich, dass jede Beobachtungseinheit aus dem Spenderdatensatz mit der gleichen Wahrscheinlichkeit für die Auswahl in Frage kommt.
Eine Alternative zur zufälligen Spenderauswahl besteht darin, die Einheiten des Empfängerdatensatz sowie die des Spenderdatensatzes in Abhängigkeit ihrer Merkmalsausprägungen in homogene Gruppen (Spenderklassen) einzuteilen (D'Orazio et al., 2006, vgl. S.37). Somit werden Beobachtungen in Datensatz A fusioniert, die in Bezug auf bestimmte Merkmale (Variablen) die gleiche Eigenschaften (Merkmalsausprägungen) aufweisen wie die Spenderbeobachtungen in B.

Die Spenderklassen werden durch die Kreuzklassifizierung der Kategorien einer bestimmten Anzahl an gemeinsamen kategorialen Variablen **X** gebildet (D'Orazio et al., 2006). In der Sozialwissenschaft und Marktforschung werden vor allem demografische Variablen wie Alter, Geschlecht oder Einkommen für die Erzeugung von Spenderklassen verwendet. Sind die Variablen ursprünglich nicht kategorial, können die Merkmalsausprägungen der quantitativen Variablen klassifiziert werden (Joenssen, 2015, vgl. S.69).

Da die fehlenden Beobachtungen im Empfängerdatensatz nur mit Beobachtungen des Spenderdatensatzes aus derselben Spenderklasse fusioniert werden können, setzt das *Random-Hot-Deck* Verfahren voraus, dass jede Spenderklasse im Spenderdatensatz mindestens eine Beobachtung zur Fusion enthält. Dies ist ein wichtiger Punkt, der bei der Auswahl der gemeinsamen Variablen, hinsichtlich der Variablenanzahl und -eigenschaft, beachtet werden muss.

Aufgrund der Spenderklassen und der dadurch entstehenden Beeinflussung der gemeinsamen Variablen **X**, ist, wie beim *Distanz-Hot-Deck* Verfahren, die imputierte Variable Z von der gemeinsamen Variable **X** abhängig und die bedingte Verteilung von Z gegeben **X** wird aus B in A reproduziert. Im Gegensatz dazu findet ohne die Verwendung der Spenderklasse kein Bezug zu **X** statt. Die zufällige Spenderauswahl impliziert demnach die Annahme, dass Z und **X** unabhängig voneinander sind. (Endres, 2019, vgl. S.112)

Rank-Hot-Deck Verfahren

Beim *Rank-Hot-Deck* Verfahren (D'Orazio et al., 2006, vgl. Kap. 2.4.2) wird die Fusion der spezifischen Variablen anhand einer ordinal- oder metrischskalierten gemeinsamen Variable X durchgeführt.

Beide Datensätze werden in Abhängigkeit der ausgewählten Variable X (Rangvariable) sortiert und anschließend die Beobachtungen anhand der Rangposition fusioniert. Die Zuordnung ist, abhängig vom Größenverhältnis von Datensatz A zu B, unterschiedlich.

Sind beide Datensätze gleich groß ($n_A = n_B$), wird die fehlende Beobachtung im Empfängerdatensatz mit der Beobachtung aus dem Spenderdatensatz verbunden, die sich an derselben Rangposition befindet. Bei einer unterschiedlich großen Anzahl ($n_A \leq n_B$) wird die empirische kumulative Verteilungsfunktion für die Verteilung der Rangvariable X in beiden Datensätze berechnet. (D'Orazio et al., 2006, vgl. S.39)
Die kumulierte Verteilungsfunktion $\hat{F}_x(x_a)$ ergibt sich nach Gleichung 2.30 mit der Indikatorfunktion 2.31 aus der Summe der Merkmalsausprägungen, die kleiner oder gleich des Beobachtungswertes x_a der Einheit a sind und durch die Gesamtzahl an Einheiten im Datensatz A geteilt werden.

$$\hat{F}_x(x_a) = \frac{1}{n_A} \sum_{t=1}^{n_A} I(x_t \leq x_a), \qquad a = 1, ..., n_A \tag{2.30}$$

$$I() = \begin{cases} 1, & (x_t \leq x_a) \\ 0, & sonst \end{cases} \tag{2.31}$$

Demnach kann jeder Einheit im Datensatz A eine kumulierte Häufigkeit zugewiesen werden, die einen Wert zwischen 0 und 1 annimmt (D'Orazio, 2013, vgl. S. 66). Analog wird die Verteilungsfunktion für den Datensatz B berechnet. Anschließend wird nach Gleichung 2.32 jeder Einheit $a = 1, ..., n_A$ im Datensatz A der Wert der Einheit b aus dem Datensatz B zugewiesen, der die kleinste Distanz zwischen den kumulierten Häufigkeiten aufweist.

$$d_{ab} = \min_{1 \leq b \leq n_B} |\hat{F}_x(x_a) - \hat{F}_x(x_b)| \tag{2.32}$$

Weisen für eine Einheit aus A zwei oder mehr Einträge aus dem Spenderdatensatz B denselben geringsten Abstand auf, wird einer der Einträge zufällig ausgewählt. Wie bei der *Distanz-Hot-Deck* Methode kann zwischen einem beschränkten und unbeschränkten Ansatz unterschieden werden (D'Orazio, 2013, vgl. S.67).

2.2.2 Hilfsinformationen

In den meisten Fällen ist es nicht möglich, anhand der vorliegenden Daten die CIA zu testen. Ist deshalb die Gültigkeit unsicher oder gilt die Annahme tatsächlich nicht, kann das Heranziehen von zusätzlichen Informationen bezüglich der spezifischen Variablen Y und Z eine weitere Möglichkeit sein. Die zusätzlichen Informationen können nach Singh et al. (1993) aus zwei Quellen herangezogen werden:

- dritter Datensatz C
- plausible Hilfswerte für die unschätzbaren Parameter von (Y,Z|X) und (Y,Z).

Der dritte Datensatz kann entweder alle Variablen (X,Y,Z) oder nur die spezifischen Variablen (Y,Z) enthalten (Singh et al., 1993, vgl. S. 60). In Abbildung 2.11 sind die beiden unterschiedlichen Datensatzkonstellationen des dritten Datensatzes C zusammen mit den Hauptdatensätzen schematisch dargestellt, die grau hinterlegten Felder markieren jeweils die fehlenden Beobachtungen.

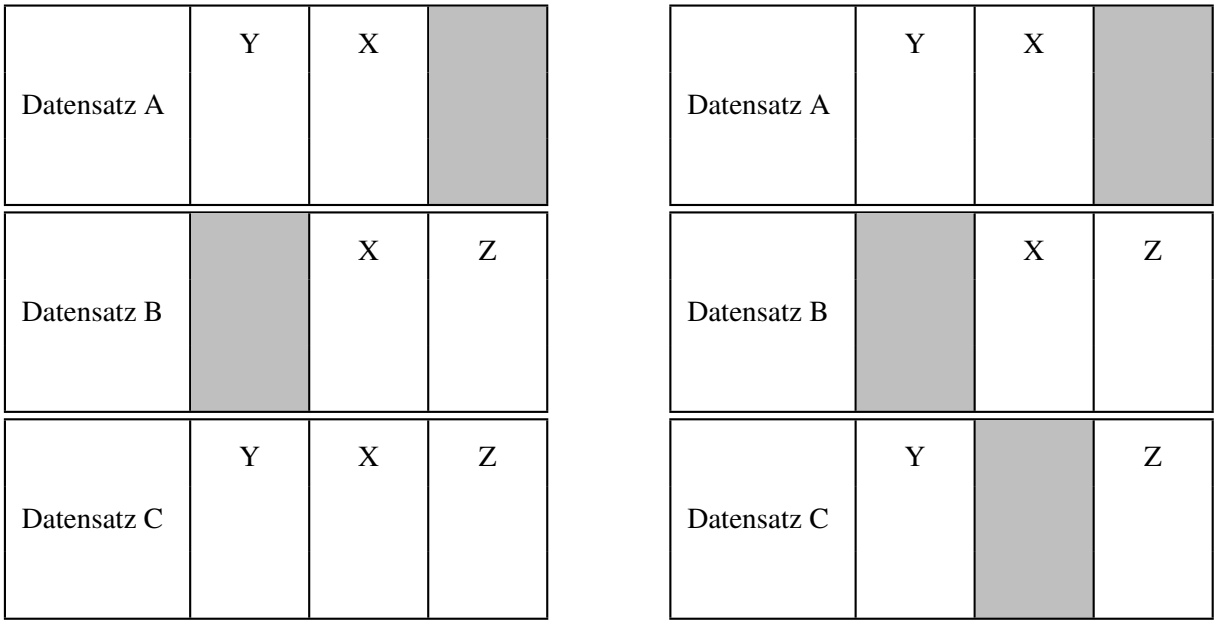


Abbildung 2.11: Schematische Darstellung der Ausgangssituation der Datenfusion beim Heranziehen eines zusätzlichen Datensatzes C, der entweder alle Variablen X, Y, Z enthält (links) oder nur die Variablen Y und Z (rechts) (Eigene Darstellung)

Für den dritten Datensatz kann entweder ein bereits bestehender Datensatz verwendet werden oder es wird gesondert eine angepasste kleine Umfrage (Ad-Hoc-Umfrage) durchgeführt (D'Orazio et al., 2006, vgl. S.67). Vor allem eine Ad-Hoc-Umfrage stellt eine gute Möglichkeit dar, da sie speziell auf die bereits vorhandenen Datensätze angepasst werden kann. Ein Nachteil ist allerdings, dass die zusätzliche Umfrage mit zeitlichem und finanziellem Aufwand verbunden ist. Die Verwendung von bestehenden Datensätzen kann einfacher sein, denn die Zusatzinformationen müssen nicht perfekt sein. Somit kann der Datensatz auch aus veralteten statistischen Untersuchungen stammen, bei denen davon ausgegangen wird, dass die Inhalte noch gültig sind. (Singh et al., 1993, vgl. S.62)

Eine Alternative zur Verwendung eines dritten Datensatzes ist die Nutzung von Proxy Variablen (D'Orazio et al., 2006, vgl. S.67). Eine Proxy Variable ist eine Variable, die eine Eigenschaft einer Variable angibt, die nicht beobachtet wurde oder nicht direkt messbar ist. Ein klassisches Beispiel stellt die Verwendung des Schulabschlusses dar über den ein mess- und vergleichbarer Beweis für die Intelligenz einer Person erbracht wird. Proxy Variablen können dementsprechend in diesem Kontext verwendet werden, wenn davon ausgegangen werden kann, dass sie ähnlich verteilt sind wie die spezifischen Variablen Y und Z (D'Orazio et al., 2006, vgl. S.67).

In Bezug auf die Methoden der Datenfusion unter dem Heranziehen von zusätzlichen Informationen kann ebenfalls zwischen einem Mikro- und Makroansatz unterschieden werden.

Die Methoden werden, wie unter der CIA, nach parametrischen und nichtparametrischen Verfahren aufgeteilt. Die Methoden, die D'Orazio et al. (2006, Kap. 3) für die Datenfusion unter Verwendung von zusätzlichen Hilfsinformationen beschreibt, sind der Tabelle 2.6 zu entnehmen. Sie ähneln grundsätzlich den Verfahren zur CIA aus Kapitel 2.2.1. Jedoch wird keine Annahme getroffen, sondern die Beziehung zwischen Y und Z aus den Zusatzinformationen mit einbezogen. Anschließend kann der Parameter θ_{YZ} geschätzt werden und die gemeinsame Verteilung von (**X**, Y, Z) lässt sich identifizieren (D'Orazio et al., 2006, vgl. S. 3.2). In Bezug auf den nichtparametrischen Makroansatz gibt es einen Unterschied. Dieser Ansatz ist ohne die CIA nicht anzuwenden. Für die Schätzung ist ein nichtparametrischer Schätzer notwendig, mit der Voraussetzung, dass vollständig beobachtete Daten vorliegen. Unter der CIA ist die Verwendung dieser Schätzer erlaubt, allerdings nicht für $A \cup B \cup C$, da Y sowie Z teilweise fehlen (Vergleich Abbildung 2.11). (D'Orazio et al., 2006, S.83)

Tabelle 2.6: Methoden der Datenfusion unter der Annahme der bedingten Unabhängigkeit (CIA)

	Makroansatz	Mikroansatz
parametrisch	Maximum-Likelihood Schätzung	Maximum-Likelihood Schätzung und Conditional Mean Matching oder Random Draws
nichtparametrisch	-	Random-Hot-Deck Rank-Hot-Deck Distanz-Hot-Deck

Für einen weiterführenden Einblick in die genannten Methoden sei auf die Literatur von D'Orazio et al. (2006, Kap. 3) verwiesen.

2.2.3 Beibehalten der Unsicherheit

Sowohl unter der CIA, als auch durch das Heranziehen von zusätzlichen Hilfsinformationen, werden Annahmen getroffen, deren Gültigkeit nicht immer gewährleistet ist. Die Verletzung dieser Annahmen kann zu fehlerbehafteten Ergebnissen führen. Daher basiert der dritte Datenfusionsansatz nicht auf zusätzlichen Annahmen, sondern darauf, die Unsicherheit, die durch das Identifikationsproblem verursacht wird, beizubehalten. Für diesen Ansatz sind vor allem parametrische Methoden bekannt. Das Beibehalten der Unsicherheit im Kontext eines nicht parametrischen Ansatzes blieb lange ein ungelöstes Problem. Erst in der aktuellen Forschung wird in Endres (2019, Kap. 3.4) ein parametrischer Mikroansatz für kategoriale Variablen beschrieben, bei dem durch die Datenfusion ein realer Datensatz erstellt wird, der die Unsicherheit beibehält.

Makroansatz mit parametrischer Methode

In Kadane (1978) wurde für die Beibehaltung der Unsicherheit als erstes ein parametrischer Ansatz für stetige Daten vorgestellt.

D'Orazio et al. (2006) erweitern den Ansatz anhand der Likelihood-Schätzung speziell für kategoriale Variablen. Im parametrischen Kontext der Datenfusion kann aufgrund der fehlenden Beobachtungen von Y und Z in $A \cup B$ für den unbekannten Parameter θ_{YZ} keine eindeutige Aussage und kein exakter Wert geschätzt werden. Jedoch sind die Parameter θ_{XY} aus dem Datensatz A und θ_{XZ} aus dem Datensatz B bekannt. Anhand dieser Werte lässt sich die Unsicherheit einschränken und ein Wertebereich für den Parameter θ_{YZ} ermitteln, in dem der wahre unbekannte Parameter liegt. (D'Orazio et al., 2006, vgl. S.100f) Dabei beeinflusst die Stärke des Zusammenhangs zwischen den gemeinsamen Variablen **X** und den spezifischen Variablen Y beziehungsweise Z den Wertebereich des Parameters θ_{XY}. Je stärker der Zusammenhang ist, desto schmaler wird der Bereich. (Rässler, 2002, vgl. S.10)
Aufgrund der Tatsache, dass es sich bei dieser Beschreibung um einen Makroansatz handelt, sei für eine weitere Vertiefung dieser Thematik beispielsweise auf D'Orazio et al. (2006, Kap. 4) oder speziell für kategoriale Variablen auf D'Orazio et al. (2006) verwiesen.

Mikroansatz mit nichtparametrischer Methode

In Endres (2019, Kap. 3.4) wird eine *Imprecise Imputation* für die Datenfusion vorgestellt, die Parallelen zu den Hot-Deck Verfahren aufweist. Die fehlenden Beobachtungen werden ebenfalls durch reale Beobachtungen aus einem anderen Datensatz fusioniert. Aus diesem Grund muss den Datensätzen erneut eine bestimmte Rolle zugewiesen werden. Ein Datensatz wird der Empfänger, einer der Spender. Für die Entscheidung gelten die in Kapitel 2.2.1 dargelegten Vor- und Nachteile, weshalb der größere Datensatz meistens die Rolle des Spenders einnimmt. Der Unterschied zu den herkömmlichen Hot-Deck Verfahren besteht darin, dass pro Einheit nicht ein Wert für die fehlende Beobachtung fusioniert wird, sondern ein Satz an möglichen Werten (Endres, 2019, vgl. S.113). Dadurch enthält der fusionierte Datensatz für die fusionierte Variable $\tilde{Z}$ pro Einheit mehrere Kategorien (Vergleich linke Abbildung 2.12). Daraus entsteht ein Unterschied zu fusionierten Datensätzen, die durch das CIA Verfahren oder Heranziehen von zusätzlichen Informationen erstellt werden. Mit dem fusionierten Datensatz können anschließend die Parameter der Verteilung nicht exakt geschätzt werden, sondern es wird eine obere und untere Grenze ermittelt. Die Berechnung der Grenzen ist anhand eines Beispiels gezeigt. In Abbildung 2.12 (links) ist ein bereits fusionierter Datensatz mit $n = 5$ Einheiten dargestellt. Für die fusionierte Variable $\tilde{Z}$ mit $k = 1, .., 3$ Kategorien ist aus der Abbildung jeweils der Satz an Kategorien zu entnehmen, die aus dem Spenderdatensatz fusioniert wurden. Der Abbildung 2.12 (rechts) sind die unteren und oberen Grenzen für die relative Häufigkeiten der Verteilung von $(\tilde{Z}, Y)$ zu entnehmen.

Y	X	$\tilde{Z}$
0	0	1|2
1	1	2|3
1	1	1|2|3
0	1	1
1	2	1|2

⇨

Wahrscheinlichkeitsverteilung von ($\tilde{Z}$,Y)		
$\tilde{Z}$/Y	0	1
1	0,2;0,4	0;0,4
2	0;0,2	0;06
3	0;0	0;0,4

Abbildung 2.12: Berechnung der oberen und unteren Grenze der Wahrscheinlichkeiten

Die unteren Grenzen p_{*kj} und oberen Grenzen p^*_{kj} der relativen Häufigkeiten werden anhand der Anzahl der Beobachtungen berechnet, in denen nur das eine Ereignis p_{*kj} auftritt oder alle Ereignisse p^*_{kj}. Für das Ereignis von $P(\tilde{Z} = 1, Y = 0)$ wird für:

- die untere Grenze p_{*10} der Quotient aus der Anzahl an Beobachtungen, die nur das eine Ereignis aufweisen (Abbildung 2.12 (links) Zeile 4), zur Gesamtanzahl *n* der Einheiten berechnet,

$$p_{*10} = \frac{1}{5} = 0,2 \tag{2.33}$$

- die obere Grenze p^*_{10} der Quotient aus der Anzahl an Beobachtungen, die mindestens das Ereignis aufweisen (Abbildung 2.12 (links) Zeilen 1 und 4), zur Gesamtanzahl *n* der Einheiten berechnet.

$$p^*_{10} = \frac{2}{5} = 0,4 \tag{2.34}$$

Damit liegen die relativen Häufigkeiten für das Ereignis von $P(\tilde{Z} = 1, Y = 0)$ zwischen 0,2 und 0,4.

In Bezug auf die Datenfusion gibt es drei verschiedene *Imprecise Imputation* Verfahren (Endres, 2019, vgl. S.112) :

- *Domain Imputation*
- *Variable-Wise Imputation*
- *Case-Wise Imputation*

Die Verfahren unterscheiden sich abhängig von der Zusammenhangsstruktur zwischen den spezifischen Variablen Y und Z zu den gemeinsamen Variablen **X**. Bei der *Domain Imputation* werden die Werte unabhängig von den gemeinsamen Variablen fusioniert. Das Verfahren weist Parallelen zu dem *Random-Hot-Deck* Verfahren ohne Spenderklassen auf. Bei den anderen beiden Verfahren werden die gemeinsamen Variablen mit einbezogen und ähneln daher dem *Random-Hot-Deck* Verfahren unter Hinzunahme von Spenderklassen.

Domain Imputation

Bei der *Domain Imputation* werden die fehlenden Beobachtungen im Empfängerdatensatz pro Einheit mit allen Merkmalsausprägungen beziehungsweise Kategorien befüllt, die die Variable im Spenderdatensatz aufweisen. Demnach enthält nach der Fusion jede Beobachtung im fusionierten Datensatz den gleichen Satz an Kategorien. (Endres, 2019, vgl. S.114). Konkret bedeutet das, dass, wenn die Variable Z im Spenderdatensatz fünf verschiedene Kategorien aufweist, im Empfängerdatensatz für jede fehlende Beobachtungen ein Satz aus den fünf verschiedenen Kategorien fusioniert wird. Somit ist gewährleistet, dass sich innerhalb des Satzes an Kategorien in jedem Fall die wahre Kategorie befindet. Der Vorteil der *Domain Imputation* ist, dass die Fusion unabhängig von den gemeinsamen Variablen ist. Somit kann sie auch verwendet werden, wenn keine gemeinsamen Variablen vorhanden sind, die einen hohen Zusammenhang zu den spezifischen Variablen aufweisen (Endres, 2019, vgl. S.114f).

Variable-Wise Imputation

Bei der *Variable-Wise Imputation* werden, in Abhängigkeit von Spenderklassen, Beobachtungen beziehungsweise ein Satz an plausiblen Werten aus dem Spenderdatensatz in den Empfängerdatensatz fusioniert (Endres, 2019, vgl. S.115). Anhand von ausgewählten gemeinsamen Variablen **X**, die eine hohe Korrelation beziehungsweise Zusammenhang zu den spezifischen Variablen Y und Z aufweisen, werden Spenderklassen gebildet (Endres, 2019, vgl. S.112). Anschließend werden alle Merkmalsausprägungen, die die zu fusionierende Variable Z innerhalb einer Spenderklasse annimmt, für die fehlenden Beobachtungen der gleichen Spenderklasse im Empfängerdatensatz fusioniert. Im Vergleich zur *Domain Imputation* besteht nach der Fusion nicht mehr jede fehlende Beobachtung aus dem gleichen Satz an Werten. Die Grenzen für die geschätzten Parameter werden somit kleiner und genauer. Jedoch ist dadurch auch nicht mehr zwangsläufig die wahre Kategorie im fusionierten Datensatz enthalten. (Endres, 2019, vgl. S.115)

Case-Wise Imputation

Treten in einem Datensatz mehrere spezifische Variablen auf, kann auch die *Case-Wise Imputation* verwendet werden. Die fehlenden Beobachtungen werden dabei als Tuple behandelt. Dementsprechend bleibt bei der Fusion die Abhängigkeit der spezifischen Variablen **Z** zueinander unverändert (Endres, 2019, vgl. S.116), da nach der Fusion pro Einheit im Tuple die Reihenfolge der spezifischen Variablen beibehalten wird. Die *Case-Wise Imputation* kann nur bei mehreren spezifischen Variablen verwendet werden und wird deshalb in dieser Arbeit nicht weiter betrachtet. Für weiterführende Informationen sei auf Endres (2019, Kap. 3.2.3) verwiesen.

2.3 Vorbereitung Datenfusion

Bevor eine Datenfusion in der Praxis durchgeführt werden kann, ist es sinnvoll, die Datensätze genauer zu analysieren und gegebenenfalls Anpassungen vorzunehmen. Die Vorbereitung der Daten auf die Fusion spielt eine wichtige Rolle für das Fusionsergebnis und hat teilweise sogar einen größeren Einfluss als die Auswahl des Fusionsverfahrens (Rässler, 2002, vgl. S.43). Eine Grundvoraussetzung für die Datenfusion ist, dass die Datensätze aus der gleichen Grundgesamtheit stammen. Dafür müssen die beiden Datensätze zeitlich, räumlich und sachlich miteinander verglichen werden (Grömping et al., 2007, vgl. S.133). Stimmen die beiden Datensätzen hinsichtlich dieser drei Kriterien überein, können die Datensätze grundsätzlich für die Fusion verwendet werden. Des Weiteren ist es wichtig, die einzelnen Variablen direkt zu betrachten und zu bewerten. Für die spezifischen Variablen ist eine Überprüfung notwendig, welche Variablen überhaupt für die Fusion in Frage kommen, da sie jeweils nur in einem Datensatz vorkommen. Dem Gegenüber ist es bei gemeinsamen Variablen wichtig, zu überprüfen, ob es sich auch wirklich um gemeinsame Variablen handelt, da mitunter inhaltlich verschiedene Variablen die gleiche (oder ähnliche) Benennung haben. Außerdem tritt die Frage auf, ob es überhaupt sinnvoll ist alle gemeinsamen Variablen für die Datenfusion zu verwenden oder ob bestimmte Eigenschaften das Fusionsergebnis negativ beeinflussen würden.

In Eurostat (2011, vgl. Kap.3) wird die Vorbereitung der Datensätze auf die Fusion in folgende drei Schritte unterteilt:

- Auswahl der Zielvariablen
- Identifikation der gemeinsamen Variablen: wird in Kapitel 2.3.1 beschrieben
- Auswahl der Matching Variablen: wird in Kapitel 2.3.2 beschrieben.

Die Zielvariablen sind die spezifischen Variablen, die nicht gemeinsam beobachtet werden und die zentrale Rollen im Zusammenhang mit der Datenfusion spielen. Wie bereits im vorherigen Kapitel 2.2 erwähnt besteht ein Ziel darin, durch die Datenfusion Informationen zu den beiden spezifischen Variablen zu erhalten. Die eindeutige Bestimmung der spezifischen Variablen ist wichtig, da anhand dieser alle weiteren Variablen ausgewählt werden, um am Ende ein gutes Fusionsergebnis zu erhalten. In den folgenden Unterkapiteln werden Voraussetzungen an die gemeinsamen Variablen sowie Analyseschritte vorgestellt, um anschließend die geforderte beste Teilmenge an gemeinsamen Variablen zu gewinnen.

2.3.1 Identifikation von gemeinsamen Variablen

Gemeinsame Variablen kommen in beiden Datensätzen vor und stehen prinzipiell zur Auswahl als Matching Variablen zur Verfügung. Um die Variablen zu unterscheiden, wird sich an der Notation von D'Orazio et al. (2006) orientiert: Variablen die in beiden Datensätzen vorhanden sind werden gemeinsame Variablen (X) genannt und die Teilmenge an gemeinsamen Variablen, die am Ende für die Fusion verwendet werden, Matching Variablen (X_M). Dabei gilt $X_M \subseteq X$. Für die weitere Verwendung müssen die gemeinsamen Variablen jeweils mit der Variable aus dem anderen Datensatz harmonieren. Nach der Analyse und der Feststellung, dass nicht alle Variablenpaare miteinander harmonieren, gibt es drei Möglichkeiten mit dieser Situation umzugehen. Entweder lassen sich die betroffenen Variablen nachträglich anpassen, stehen demnach für die Auswahl als Matching Variable weiterhin zur Verfügung, werden nicht weiter betrachtet oder die Datenfusion kann nicht mehr durchgeführt werden, weil eine zu große Anzahl an Variablen nicht miteinander harmonieren. (Eurostat, 2011, vgl. S.35)

Variablen können nicht verwendet werden, wenn sie sich bezüglich ihrer Definition oder Klassifizierung unterscheiden (Eurostat, 2011, vgl. S.35). Es ist möglich, dass Variablen auf den ersten Blick durch gleiche Benennung identisch aussehen, sie sich allerdings in ihrer Definition unterscheiden. Aus diesem Grund ist es wichtig die Definitionen und Beschreibungen der einzelnen Variablen zu überprüfen. Ein möglicher Grund für die verschiedenen Bedeutungen ist, dass die Stichproben meistens von unterschiedlichen Institutionen erhoben werden. Stimmen die Variablen inhaltlich überein, ist es insbesondere bei kategorialen Variablen notwendig, die Definition hinsichtlich ihrer Klassifizierung zu analysieren und wenn möglich anzupassen (Eurostat, 2011, vgl. S.3f). Kategoriale Variablen lassen sich in viele unterschiedliche Kategorien einteilen.

Besteht zum Beispiel die Variable „Alter“ im Datensatz A aus den Kategorien „<15“, „15-20“ und „>20“ und im Gegensatz dazu im Datensatz B aus„<15“, „15-18“, „19-20“ und „>20“, stimmen die Klassifizierungen nicht überein.

Eine Harmonisierung der Kategorien kann für kompatible Kategorien vorgenommen werden, indem beispielsweise die Kategorien „15-18“ und „19-20 “ im Datensatz B zusammengefasst werden und damit an den Datensatz A angepasst werden. Sind die Klassifizierungen der beiden Datensätze nicht kompatibel und damit nicht aneinander anpassbar, können die Variablen nicht für die Fusion verwendet werden.

Für das Fusionsergebnis ist es zudem wichtig, dass beide Datensätze für die gemeinsamen Variablen die gleiche bzw. ähnliche Verteilung aufweisen, damit sich die Datensätze nicht durch die gemeinsamen Variablen unterscheiden. Außerdem kann dadurch bewertet werden, ob die beiden Datensätze wirklich aus der gleichen Grundgesamtheit stammen. Treten viele unterschiedlich verteilte Variablen auf, ist dies ein mögliches Indiz dafür, dass die Zielpopulation der beiden Erhebungen unterschiedlich und eine Datenfusion nicht empfehlenswert ist. Treten jedoch nur bei einigen Variablen unterschiedliche Verteilungen auf, ist ein möglicher Grund die Stichprobenvariabilität und die Variablen sollten für die Fusion nicht weiter betrachtet werden. (Eurostat, 2011, vgl. S.37)

Der Grad der Ähnlich- bzw. Unähnlichkeit von Verteilungen kann durch unterschiedliche Methoden berechnet werden. Eine klassische Möglichkeit ist die Verwendung von statistischen Tests, die abhängig von ihrem Variablentyp ausgewählt werden können. Bei kategorialen Variablen bietet zum Beispiel der χ^2-Test eine Möglichkeit und bei stetigen Daten der Kolmogorov-Smirnov Test. Jedoch werden durch diese Tests die Verteilungen bereits als unähnlich angesehen, wenn nur kleine Unterschiede zwischen den Verteilung zu erkennen sind. Diese Tests sind demnach zu genau für diesen Harmonisierungsschritt und werden nicht empfohlen. (Eurostat, 2011, vgl. S. 37)

Eine andere Möglichkeit ist der empirische Ansatz, der darin besteht, die Verteilungen durch Ähnlichkeits- beziehungsweise Unähnlichkeitsmaße miteinander zu vergleichen (D'Orazio, 2013, S.161).
Ein Ähnlichkeitsmaß für den Vergleich von Verteilungen ist der Bhattacharyya-Koeffizient $BC(p_A, p_B)$, der sich aus Gleichung 2.35 aus den aufsummierten Wurzeln des Produktes der relativen Häufigkeiten p für die einzelnen Kategorien $j = 1, ..., J$ der jeweiligen Variable im Datensatz A und B ergibt. Der Wertebereich des Bhattacharyya Koeffizienten liegt zwischen 0 (Verteilungen sind ungleich) und 1 (Verteilungen sind gleich).

$$BC(p_A, p_B) = \sum_{j=1}^{J} \sqrt{p_{A,j} \cdot p_{B,j}} \tag{2.35}$$

Aus dem Bhattacharyya-Koeffizienten lässt sich ein weiteres, in der Literatur häufig verwendetes Maß, die Hellinger-Distanz $H_D(p_A, p_B)$, bestimmen 2.36. Die Hellinger-Distanz ist ein Unähnlichkeitsmaß das ebenfalls Werte zwischen 0 und 1 annehmen kann. Allerdings wird die Verteilung bei 0 als gleich und bei 1 als ungleich angesehen.

$$\begin{aligned} H_D(p_A, p_B) &= \sqrt{1 - BC(p_A, p_B)} \\ &= \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{J} (\sqrt{p_{Aj}} - \sqrt{p_{Bj}})^2} \end{aligned} \tag{2.36}$$

$\sqrt{2}$ normiert das Ergebnis, so dass alle Ergebnisse ($H_D(p_A, p_B)$) kleiner gleich 1 sind. Eine weitere Berechnungsmöglichkeit ist das Jensen-Shannon-Divergenzmaß (JS) (Lin, 1991), das auf dem Kullback-Leibler-Divergenzmaß (KL) aufbaut. Im Vergleich zur KL-Divergenz ist die Js-Divergenz jedoch ein symmetrisches ($D_{JS}(p_A, p_B) \neq D_{JS}(p_B, p_A)$) und stabileres Maß, weshalb es gegenüber kleinen Störungen der Daten unempfindlicher ist (Dudeck, 2013, S.200f). Die JS-Divergenz gibt an, wie ähnlich ($D_{JS} = 0$) oder unähnlich ($D_{JS} = 1$) sich die Verteilungen sind und ist definiert als:

$$D_{JS}(p_A, p_B) = \frac{1}{2} \sum_{j=1}^{J} p_{Aj} \cdot \log_2\left(\frac{2\ p_{Aj}}{p_{Aj} + p_{Bj}}\right) + p_{Bj} \cdot \log_2\left(\frac{2 p_{Bj}}{p_{Bj} + p_{Aj}}\right) \tag{2.37}$$

In der Literatur gibt es bisher keinen einheitlichen Grenzwert der Ähnlichkoeffizienten, bis zu dem die Verteilungen als gleich beziehungsweise ähnlich definiert sind. Jedoch für die Hellinger-Distanz geben beispielsweise Leulescu und Agafiţei (2013, vgl. S.14) und Webber und Tonkin (2013, vgl. S.13) als Faustregel ein Grenzwert von 0,05 an.

Neben den statistischen Tests und den Ähnlichkeits-/Unähnlichkeitsmaßen können Verteilungen noch anhand von Histogrammen verglichen werden. Für die jeweiligen Variablen wird für jeden Datensatz ein Histogramm erstellt. Die Verteilungen der Variablen werden als gleich bzw. ähnlich angesehen, wenn innerhalb jeder Kategorie sich bei großen Häufigkeiten diese in den beiden Datensätzen um nicht mehr als 5% voneinander unterscheiden und bei kleineren Häufigkeiten um nicht mehr als 2%. Außerdem sollte geprüft werden, ob die Form der beiden Histogramme nahezu identisch ist. (D'Orazio, 2013, S.37)

2.3.2 Auswahl der Matching Variablen

Nachdem die gemeinsamen Variablen analysiert sind und nur noch tatsächlich gemeinsame Variablen zur Verfügung stehen, können dennoch viele verschiedene Variablen auftreten. Deshalb stellt sich die Frage, ob es sinnvoll und notwendig ist alle gemeinsamen Variablen für die Verknüpfung der beiden Datensätze zu nutzen. Der Nachteil bei zu vielen Variablen besteht in einem steigenden Risko, dass die Kontingenztabellen immer spärlicher werden, das heißt, die einzelnen Zellen beinhalten nur wenige oder keine Beobachtungen. Das führt wiederum dazu, dass die Komplexität der Datenfusionsmethode steigt und die Genauigkeit der Übereinstimmungsergebnisse sinkt. (D'Orazio, 2013, S.167). Außerdem kann die bedingte Unabhängigkeit (CIA) eher gewährleistest werden, wenn die gemeinsamen Variablen eine hohe Aussagekraft zu den spezifischen Variablen Y und Z aufweisen (Rässler, 2002, vgl. S.57). Demzufolge ist es sinnvoll, ausschließlich die Variablen für die Fusion zu verwenden, die das Fusionsergebnis positiv beeinflussen. Eine Möglichkeit gemeinsame Variablen für die Fusion auszuwählen ist, zu testen wie stark der Zusammenhang zwischen den gemeinsamen Variablen und den spezifischen Variablen ist. Je höher die Aussagekraft der gemeinsamen Variablen zu den spezifischen Variablen ist, desto besser werden die Ergebnisse der Datenfusion (Grömping et al., 2007, S.127).

Klassische Methoden zur Auswahl von geeigneten Variablen beruhen auf der Berechnung von Assoziationsmaßen.

Um allerdings die Stärke des Zusammenhangs zwischen den gemeinsamen Variablen und den spezifischen Variablen zu berechnen, wird bei den klassischen Methoden eine gemeinsame Beobachtung der spezifischen Variablen Y und Z vorausgesetzt. In der Ausgangssituation der Datenfusion gibt es jedoch keine gemeinsamen Beobachtungen. Eine Konsequenz ist daher, dass die Analysen in den einzelnen Datensätzen getrennt ausgeführt werden müssen. Das Ergebnis besteht anschließend zum einen aus einer Teilmenge an Variablen, die die spezifische Variable Z besser erklären ($X_Z \subseteq X$) und zum anderen aus einer Teilmenge an Variablen, die Y besser erklären ($X_Y \subseteq X$). (D'Orazio et al., 2019, S.103)

Anschließend ist eine Auswahl der Matching Variablen vorzunehmen. Entweder können durch die Vereinigung ($X_Y \cup X_Z = X_M$) alle Variablen ausgewählt werden oder nur die Schnittmenge ($X_Y \cap X_Z = X_M$), also die Variablen, die sowohl zu Y als auch zu Z einen starken Zusammenhang aufweisen. Im Endeffekt muss aufgrund von Vorwissen und Erfahrungen die Menge X_M entsprechend eines Kompromisses aus beiden Varianten nach Gleichung 2.38 gebildet werden (D'Orazio et al., 2019, vgl. S.104).

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z \tag{2.38}$$

Der Kompromiss ist notwendig, da die Vereinigung ($X_Y \cup X_Z$) möglicherweise eine zu große Anzahl an Variablen aufweist, während die Schnittmenge $X_Y \cap X_Z$) womöglich wichtige Variablen vernachlässigt.

Insgesamt gibt es verschiedene Assoziationsmaße, die verwendet werden können, um die Stärke des Zusammenhalts zwischen Variablen auszurechnen. Die Auswahl des Verfahrens richtet sich nach dem Variablentyp (Vergleich Abbildung 2.1). Ein Überblick über verschiedene Assoziationsmaße ist bei D'Orazio et al. (2006, vgl. S.168) aufgelistet. Als Basis für die Berechnung von Zusammenhängen zwischen nominalskalierten Variablen dient häufig die Kontingenztabelle, in die die Verteilung der zu untersuchenden Variablen eingetragen ist. Für nominalskalierte Variablen (beide Variablen sind nominal) wird häufig das Cramers V berechnet, ein Maß für den statistischen Zusammenhang zweier Variablen, das auf der Pearson Chi-Quadrat (χ^2) Statistik basiert. Für die Berechnung des Cramers V wird daher der χ^2-Wert benötigt. Dieser ergibt sich nach Gleichung (2.39) aus der Summe der quadrierten Differenz zwischen beobachteten n_{ij} und erwarteten $\hat{n}_{ij}$ Häufigkeiten, geteilt durch die erwarteten Häufigkeiten, mit $i = 1,...,I$ und $j = 1,...,J$. I und J sind jeweils die Anzahl der Kategorien der Variablen, zwischen denen der Zusammenhang berechnet wird.

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}} \tag{2.39}$$

Die erwarteten Häufigkeiten ergeben sich dabei aus dem Quotienten des Produkts der Randhäufigkeiten n_{i+} und n_{+j} zur Gesamtanzahl der Beobachtungen n. Der χ^2-Wert selbst kann grundsätzlich auch als ein Zusammenhangsmaß verwendet werden, da mit stärker werdendem Zusammenhang der Wert höher ausfällt. Allerdings ist er abhängig von der Anzahl der Kategorien sowie der Gesamtanzahl an Beobachtungen, weshalb er für Vergleichszwecke nicht empfohlen wird (Kohn, 2005, vgl. S. 115).

Beim Cramers V wird dagegen der χ^2-Wert auf ein Intervall zwischen null (kein Zusammenhang) und eins (starker Zusammenhang) normiert. Dies geschieht, indem nach Gleichung 2.40 die Wurzel aus dem χ^2-Wert geteilt durch das Produkt von der Gesamtanzahl der Beobachtungen n und die kleinere Anzahl an Kategorien der Variablen ($min\{I-1, J-1\}$) gezogen wird. Aufgrund dieser Normierung ist der Wert unabhängig von der Anzahl der Merkmalsausprägungen sowie der Gesamtanzahl der Beobachtungen und lässt sich damit eindeutig interpretieren.

$$V = \sqrt{\frac{\chi^2}{n \cdot min\{I-1, J-1\}}} \tag{2.40}$$

Eine weiteres Maß, das die Stärke des Zusammenhanges zwischen zwei Variablen angibt und für die Auswahl der Matching Variablen verwendet wird, ist der Unsicherheitskoeffizient, auch Theils U genannt (D'Orazio et al., 2019, vgl. S.111). Der Unsicherheitskoeffizient gehört zu den Fehlerreduktionsmaßen, die oft als PRE-Maße (engl. proportional reduction of error) abgekürzt werden. Allgemein wird getestet, ob bei einer Modelländerung mehr Informationen über die spezifische Variable (beispielsweise Y) erhalten werden können, wenn Kenntnisse über X verfügbar sind. Insgesamt werden alle PRE-Maße nach der Gleichung 2.41 berechnet, und ermitteln um wie viel Prozent sich der Vorhersagefehler durch die Kenntnis der gemeinsamen Variable X verkleinert. (Jann, 2011, vgl. S.75)

$$\text{PRE} = \frac{E_1 - E_2}{E_1} \tag{2.41}$$

Dabei ist E_1 der Fehler der Vorhersage der spezifischen Variable Y ohne Kenntnis des Zusammenhanges zu X und E_2 der Fehler der Vorhersage von Z bei Kenntnis über den Zusammenhang mit X. Die Vorhersagefehler sind in Abhängigkeit des Variablentyps definiert. Genau wie das Cramers V kann das PRE-Maß einen Wert zwischen 0 und 1 annehmen. Eine größere Zahl weist darauf hin, dass mehr Informationen über Y vorhersagbar sind, wenn X bekannt ist. Hinsichtlich der Grenzwerte findet sich in der Literatur, wie ebenfalls für Cramers V, keine einheitliche Definition. Diese Arbeit orientiert sich an folgender Einteilung (Stephanie, 2018):

- $0 <$ PRE $< 0,1$: schwache Beziehung
- $0,1 \leq$ PRE $< 0,4$: mittlere Beziehung
- $0,4 \leq$ PRE < 1: starke Beziehung.

Ein PRE-Maß für metrisch skalierte Variablen ist das Bestimmtheitsmaß, das im TRACE-Projekt (Grömping et al., 2007, Kap. 4.5.3) für die Analyse des Zusammenhanges verwendet wird. Im Vergleich dazu ist der Unsicherheitskoeffizient für nominalskalierte Variablen definiert. Er basiert auf der Entropie der Verteilungen, welche für nominalskalierte Variablen als Streuungsmaß definiert ist (Jann, 2011, vgl. S. 50). Die Entropie $H(X)$ der Variable X ergibt sich aus der negativen Summe des Produkts der relativen Häufigkeiten p_i der Variable X und dem Zweierlogarithmus von p_i, mit $i = 1, ..., I$.

$$H(X) = -\sum_{i=1}^{I} p_i \log_2 p_i \tag{2.42}$$

Die Entropie gibt an, mit welcher Unsicherheit eine zufällig ausgewählte Beobachtung in eine bestimmte Kategorie gehört. Für den Unsicherheitskoeffizient U bedeutet das, dass dieser nach Gleichung 2.43 ermittelt, um wie viel Prozent sich die Entropie bei Y-Vorhersage $H(Y)$ zur Entropie bei Y-Vorhersage unter Berücksichtigung von X $H(Y|X)$ verkleinert (Jann, 2011, vgl. S.78f). Sind Y und X unabhängig, so ist $H(Y) = H(Y|X)$ und $U(Y|X)$ wird zu 0.

$$U(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\log_2(\frac{p_{ij}}{p_{i+}}) - \sum_{j=1}^{J} p_{+j}\log_2 p_{+j}}{-\sum_{j=1}^{J} p_{+j}\log_2 p_{+j}} \tag{2.43}$$

Ein Vorteil gegenüber den Pearson Chi-Quadrat basierten Maßen wie dem Cramers V ist, dass ein PRE-Maß von 0,4 doppelt so stark ist wie ein PRE-Maß von 0,2. Bei Cramers V gilt dies nicht, dieses kann nur als schwach, moderat oder stark interpretiert werden. (Hanneman et al., 2012, vgl. S.383)
Aus diesem Grund eignen sich PRE-Maße gut für den Vergleich.

Wie bereits in Kapitel 2.2.3 beschrieben, beeinflussen die Zusammenhänge zwischen den gemeinsamen Variablen und den spezifischen Variablen den Unsicherheitsbereich der Parameter der gemeinsamen Verteilung von (Y, Z). Die Wahrscheinlichkeit der Gültigkeit der CIA steigt des Weiteren mit Abnahme des Unsicherheitsbereichs, da unter der CIA die geschätzten Parameter immer innerhalb dieses Bereichs liegen (D'Orazio, 2013, vgl. S.121). Aus diesem Grund basiert ein anderer Ansatz für kategoriale Variablen auf der Unsicherheit. Gemeinsame Variablen werden so gewählt, dass die Unsicherheit zwischen den spezifischen Variablen am effektivsten verringert wird. Als Primärquelle für die Beschreibung der Auswahl der Matching Variablen anhand der Unsicherheit wird D'Orazio et al. (2019) verwendet. Das allgemeine Ziel ist, die Menge an gemeinsamen Variablen **X** zu identifizieren, die die Unsicherheit am stärksten minimieren. Der Ablauf zur besten Identifikation der Teilmenge an **X** Variablen ist dabei in die folgende drei Schritte unterteilt (D'Orazio, 2019, vgl. S.110):

Schritt 1 Auswahl der X_p Variable mit der kleinsten Unsicherheit $\tilde{u}_p$

Schritt 2 Berechnung der Unsicherheit $\tilde{u}_m$ für alle Möglichkeiten bei der Hinzunahme der anderen X Variablen (pro Iterationsschritt wird eine Variable hinzugefügt und vor jedem neuen Iterationsschritt wird Schritt 3 durchgeführt)

Schritt 3 Auswahl der Kombination mit der kleinsten Unsicherheit $\tilde{u}_m$.
Anschließend Überprüfung, ob Auswahl beendet werden kann. Dies trifft zu, wenn:

- der kleinste ausgewählte $\tilde{u}_m$ Wert größer ist als der im vorherigen Iterationsschritt oder
- (nur bei Verwendung von Strafterm 1 ($u_m^{(1)}$)) der Grad der Spärlichkeit nicht akzeptabel ist ($\tilde{n} \leq 1$)

Ansonsten zurück zu Schritt 2. Es beginnt ein neuer Iterationsschritt mit dem Hinzufügen von Variablen.

Die Berechnung der Unsicherheit und Spärlichkeit für die Auswahl der Matching Variablen wird im Folgenden erklärt. Im Anhang A.4 befindet sich dazu ein Fallbeispiel, welches die einzelnen Rechnungen und Schritte ausführlich beschreibt.

Die Verwendung der *Fréchet-Hoeffding-Grenzen* (engl. *Fréchet bounds*) nach Gleichung 2.44 ermöglicht es, für jede Zelle der Kontingenztabelle Y x Z, unter Kenntnis von X, die Grenzen des Unsicherheitsbereichs zu berechnen, in denen die Wahrscheinlichkeiten $p_{+jk} = Pr(Y = j, Z = k)$ liegen (D'Orazio, 2019, vgl. S.105).[7]

$$\left[\underline{p}_{+jk}, \overline{p}_{+jk}\right] = \left[\sum_i p_{i++} \max\{0, p_{j|i} + p_{k|i} - 1\}, \sum_i p_{i++} \min\{p_{j|i}, p_{k|i}\}\right] \tag{2.44}$$

Dabei wird $p_{j|i}$ aus dem Datensatz A ermittelt, $p_{k|i}$ aus B und p_{i++} durch eine Addition der Häufigkeiten beider Datensätze. Anschließend lässt sich aus den Grenzen die Gesamtunsicherheit u berechnen, die nach Gleichung 2.45 die durchschnittliche Breite der Grenzen ist (D'Orazio et al., 2019, vgl. S. 105).

$$u = \frac{1}{J \times K} \sum_{j,k} \left(\overline{p}_{+jk}, \underline{p}_{+jk}\right) \tag{2.45}$$

D'Orazio et al. (2019, vgl. S.106-109) zeigt anhand eines Beispiels allerdings den Nachteil dieser Berechnung. Mit steigender Anzahl an X Variablen wird die Unsicherheit zwar immer kleiner, jedoch tritt eine Verzerrung der Schätzung ein. D'Orazio setzt diese Verzerrung wiederum mit einer steigenden Anzahl an leeren Zellen, aufgrund von immer größer werdenden Kontingenztabellen, in Verbindung. Aus diesem Grund führt er zwei Strafterme ein, die sich auf die Genauigkeit der Schätzer und insbesondere die Spärlichkeit von Kontingenztabellen beziehen (D'Orazio, 2019, vgl. S.109). Die Berechnung der Unsicherheit ergibt sich demnach nach Gleichung 2.46 aus der Summe der allgemeinen Unsicherheit u_m und dem Strafterm g.

$$\tilde{u}_m = u_m + g \tag{2.46}$$

Die Spärlichkeit nimmt mit der Anzahl an Variablen oder an Kategorien in einer Kontingenztabelle schnell zu.

Ein Beispiel zur Verdeutlichung: Die gemeinsame Verteilung von 5 Variablen mit jeweils 5 Kategorien produziert eine Kontingenztabelle mit 3125 Zellen. Bereits das Hinzufügen einer weiteren Variable mit 5 Kategorien erweitert die Kontingenztabelle um 12500 Zellen. Bei dieser Größenordnung kann die Anzahl der Zellen schnell größer werden als die Stichprobengröße und somit zu leeren Zellen führen. In solch einer Situation wird von einer spärlichen Kontingenztabelle gesprochen.

In der Literatur gibt es allerdings keine allgemeine Regelung, ab wann Kontingenztabellen als spärlich angesehen werden.

[7] mit $i = 1, \ldots, J$ für X, $j = 1, \ldots, J$ für Y und $k = 1, \ldots, K$ für Z. I, J und K sind jeweils die Anzahl der Kategorien.

In Agresti (2002, vgl. S.391) ist die Definition allgemein gehalten und eine Kontingenztabelle wird als spärlich betrachtet, wenn diese kleine Zellenzahlen aufweist. In Agresti und Yang (1987, vgl. S.9) wird eine Kontingenztabelle als spärlich definiert, wenn das Verhältnis der Stichprobengröße zur Anzahl der Zellen relativ klein ist. Beide Quellen definieren jedoch keinen genauen Wert für „klein“.

D'Orazio beschreibt zwei verschiedene Möglichkeiten für die Berechnung des Strafterms, die sich vor allem auf die Spärlichkeit beziehen (D'Orazio et al., 2019, Kap. 5.2.4). Zum besseren Verständnis der Vorgehensweise und der Gleichungen folgt eine kurze Einführung in die Notation. Für die Auswahl der Matching Variablen wird die Teilmenge M der X Variablen gesucht, die die Unsicherheit am meisten reduziert. Daher wird die Unsicherheit für verschiedene Teilmengen der X Variablen berechnet. Insgesamt können

$$M = (2^P - 1) \tag{2.47}$$

Teilmengen erstellt werden. P ist dabei die Anzahl der verfügbaren (gemeinsamen) X Variablen. Die Kreuzklassifizierung aller gemeinsamer Variablen ergibt die Variable D_P ($D_P = X_1 \times X_2 \times ... \times X_P$) mit H_{D_P} Kategorien. Im Gegensatz dazu ist D_m die Variable, die durch die Kreuzklassifizierung einer Teilmenge gemeinsamer Variablen erhalten wird und H_{D_m} die Anzahl der Kategorien. Mit dem folgenden Beispiel lässt sich die Notation verdeutlichen. Es gibt insgesamt drei gemeinsame Variablen ($P = 3$): Geschlecht mit 2 Kategorien, Alter mit 5 und Unfallursache mit 3. Daraus ergibt sich $D_P = Geschlecht \times Alter \times Unfallursachen$ und $H_{D_P} = 2 \cdot 5 \cdot 3 = 30$. Für die Betrachtung der Teilmenge mit den X Variablen Geschlecht und Alter ist $D_m = Geschlecht \times Alter$ und $H_{D_m} = 2 \cdot 3 = 6$.

Der erste Strafterm ergibt sich nach Gleichung 2.48 durch den natürlichen Logarithmus von 1 addiert zu dem Quotienten aus der Anzahl der Kategorien der Teilmenge H_{D_m} zu der Anzahl der Kategorien aller X Variablen H_{D_P} (D'Orazio, 2019, vgl. S.109).

$$g^{(1)} = \log\left(1 + \frac{H_{D_m}}{H_{D_P}}\right) \tag{2.48}$$

Der Strafterm bezieht sich auf die Spärlichkeit der X Variablen, je mehr X Variablen, desto größer wird der Strafterm. Zusätzlich wird ein Stoppkriterium hinzugefügt, welches sich auf die Spärlichkeit der Verteilung von (X,Y) und (X,Z) bezieht. D'Orazio schlägt für diese Regelung vor, dass die Teilmenge an X Variablen akzeptabel ist, wenn die Bedingung aus Gleichung 2.49 erfüllt ist (D'Orazio, 2019, vgl. S.109). Die Spärlichkeit wird durch das Verhältnis der Stichprobengröße n zur Anzahl der Zellen der Kontingenztabelle in beiden Datensätzen berechnet und das Minimum davon ausgewählt.

$$\bar{n} = \min\left[\frac{n_A}{H_{D_m} \cdot J}, \frac{n_B}{H_{D_m} \cdot K}\right] > 1 \tag{2.49}$$

Die Anzahl der Zellen der Kontingenztabellen ergeben sich jeweils aus dem Produkt der Anzahl der Kategorien J oder K der spezifischen Variablen Y/Z und H_{D_m}. Hierbei ist anzumerken, dass die Berechnung der Spärlichkeit und vor allem die Wahl des Stoppkriteriums eine subjektive Entscheidung ist. Wie bereits erwähnt, gibt es keine einheitliche Regelung ab wann eine Kontingenztabelle als spärlich anzusehen ist.

Somit ist es durchaus möglich, für das Stoppkriterium eine andere Regelung oder eine andere Berechnung der Spärlichkeit anzuwenden. In dieser Arbeit wird sich jedoch an der Definition von D'Orazio orientiert.

Im Gegensatz dazu besteht nach Gleichung 2.50 im zweiten Strafterm nicht die Möglichkeit, die Regelung für die Spärlichkeit zu verändern. Denn diese ist bereits integriert und der Strafterm stoppt aufgrund der Bedingung $n_A > (H_{D_m} \cdot J)$ und $n_B > (H_{D_m} \cdot K)$ bei $\bar{n} \leq 1$ (D'Orazio et al., 2019, vgl. S.109).

$$g^{(2)} = \max \left[\frac{1}{(n_A - H_{D_m} \cdot J)}, \frac{1}{(n_B - H_{D_m} \cdot K)} \right] \tag{2.50}$$

$$\text{mit } n_A > (H_{D_m} \cdot J) \text{ und } n_B > (H_{D_m} \cdot K)$$

Ein Vorteil gegenüber dem ersten Strafterm ist, dass für den zweiten Strafterm kein Wissen über die Anzahl der Zellen beim Kreuzen aller verfügbaren Variablen (H_{D_Q}) vorhanden sein muss.

Anhand der am Anfang aufgelisteten Verfahrensschritte ist es möglich, am Ende die Variablen auszuwählen, die die Unsicherheit in der Kontingenztabelle von Y und Z am stärksten minimieren.

In diesem Kapitel wurden verschiedene Methoden zur Auswahl von Matching Variablen vorgestellt. Aufgrund des hohen Rechenaufwands empfehlen D'Orazio et al. (2019, vgl. S.117) die Unsicherheitsmethode bei einer geringen Anzahl gemeinsamer Variablen anzuwenden und im Gegensatz dazu bei einer größeren Variablenanzahl ein Zweischrittverfahren durchzuführen. Für das Zweischrittverfahren wird zuerst eine Teilmenge an Variablen durch die herkömmlichen Verfahren ausgewählt, um nachfolgend daraus die endgültigen Matching Variablen mit der Unsicherheitsmethode zu bestimmen. Begründet wird dieses Verfahren damit, dass die mit der Unsicherheitsmethode bestimmten Matching Variablen meistens eine Teilmenge der Variablen sind, die durch herkömmliche Verfahren ermittelt wurden. (D'Orazio et al., 2019, vgl. S.117)

2.4 Kriterien zur Bewertung der Datenfusion

Grundsätzlich basiert die Beurteilung der Datenfusion auf dem Vergleich der Eigenschaften des fusionierten Datensatzes mit den Originaldaten und deren Übereinstimmung. Dabei wurde in früheren Studien sowohl die Verteilung der Variable $\tilde{Z}$ im fusionierten Datensatz mit der Variable Z im Originaldatensatz verglichen, als auch der Zusammenhang zwischen den spezifischen Variablen Y und Z.

Beispielsweise wurden Mittelwert sowie Standardabweichung der fusionierten Variable mit denen der originalen Variablen verglichen und die Qualität der Datenfusion anhand der Differenz beurteilt (Rodgers und DeVol, 1981, vgl. S.131). Zur Bewertung des Zusammenhangs der spezifischen Variablen von fusionierten und originalen Daten wurde die Korrelation mittels Korrelationskoeffizienten berechnet und ebenfalls anhand der Differenz ausgewertet (Rodgers und DeVol, 1981, vgl. S.131). Barry (1988, vgl. S.281f) erweiterte diese Bewertungen durch den Vergleich der Regressionsschätzer auf Basis des fusionierten und des originalen Datensatzes.

Rässler (2002, Kapitel 2.5) beschäftigte sich ausführlich mit den Bewertungskriterien hinsichtlich der Datenfusion und definierte vier Validitätsebenen. Die einzelnen Ebenen beinhalten die zuvor genannten Bewertungsansätze und wurden um weitere Punkte ergänzt. Die erste Validitätsebene ist dabei am schwierigsten zu erreichen, die vierte am einfachsten. Insgesamt können mehr Analysemethoden auf den fusionierten Datensatz angewendet werden, je höher (im Sinne von: Ebene 1 ist die höchste Ebene) die erfüllte Validitätsebene ist. Allerdings können die meisten Ebenen (1-3) nur mit künstlich erzeugten Simulationsdatensätzen geprüft werden, da die eigentlich nicht vorhandenen spezifischen Variablen als Vergleich benötigt werden (Rässler, 2002, Kap. 2.5.5). Nachfolgend werden die einzelnen Validitätsebenen, die zur Bewertung der Datenfusion dienen, näher erläutert und jeweils Kennzahlen zur Überprüfung dargelegt.

2.4.1 Validitätsebenen nach Rässler

Erste Validitätsebene: Erhalt der einzelnen Werte

Die erste Ebene ist erreicht, wenn die einzelnen Werte des fusionierten Datensatzes exakt reproduziert sind und demnach mit den wahren Werten übereinstimmen (Cielebak und Rässler, 2014, vgl. S.378). Diese Ebene ist am schwierigsten zu erreichen und oft nicht so wichtig wie der Erhalt der gemeinsamen Verteilung (D'Orazio et al., 2006, vgl. S. 10). Eine hohe Trefferquote und damit eine exakte Reproduzierbarkeit der Werte ist nur dann möglich, wenn die gemeinsamen Variablen **X** die zu fusionierende Variable Y/Z exakt beschreiben. Berechnet werden kann die Ebene durch die von Rässler (2002, vgl. S.30) vorgeschlagene „Hit Rate", bei der eine Übereinstimmung zwischen einer Beobachtung im fusionierten Datensatz dem wahren Wert entspricht und als „Hit" bezeichnet wird. Die Güte der Ebene ist jedoch nicht ausschließlich von der Qualität der Datenfusion abhängig, sondern auch von der Anzahl der Merkmalsausprägungen. Bei kategorialen Daten kann die Hit Rate durchaus ein Hinweis auf die Qualität des fusionierten Datensatzes sein. Allerdings gilt dies nur bei Variablen mit wenigen Kategorien, da die Wahrscheinlichkeit für den Erhalt der exakten Werte bei einer Variable mit vielen Kategorien sehr gering und bei einem Fall mit mehreren spezifischen Variablen nahezu Null ist. Bei quantitativen Variablen ist die Wahrscheinlichkeit sogar Null und somit keine brauchbare Qualitätsmessung. (Rässler, 2002, vgl. S.30)

Zusätzlich ist das Ziel einer Datenfusion häufig nicht das Erhalten der einzelnen Werte, sondern mit den fusionierten Daten Analysen durchführen zu können. Aus diesem Grund ist vor allem der Erhalt der gemeinsamen Verteilung wichtig, da sie die statistisch relevanten Informationen enthält. Der Erhalt der einzelnen Werte alleine führt nicht unbedingt zu einer guten Reproduktion der Verteilung. (Cielebak und Rässler, 2014, vgl. S. 378)
Kiesl und Rässler (2005, vgl. S.25f) zeigen dies anhand eines Beispiels, bei dem verdeutlicht wird, dass eine hohe Trefferquote nicht gleichzusetzen ist mit einem guten Fusionsergebnis. Es handelt sich hier jedoch um ein Extrembeispiel, da die Fusion der spezifischen Variable unabhängig von den Zusammenhängen zu anderen Variablen durchgeführt wird.

Zweite Validitätsebene: Erhalt der gemeinsamen Verteilung

Die zweite Ebene kann ebenfalls nur mit Zusatzinformationen, also in Simulationsstudien, überprüft werden und definiert das eigentliche Ziel der Datenfusion. Die zweite Ebene ist erfüllt, wenn nach Gleichung 2.51 die gemeinsame Verteilung im fusionierten Datensatz $\tilde{f}_{\mathbf{X}YZ}$ der gemeinsamen wahren Verteilung $f_{\mathbf{X}YZ}$ entspricht (Rässler, 2002, vgl. S.30).

$$\tilde{f}_{\mathbf{X}YZ} = f_{\mathbf{X}YZ} \tag{2.51}$$

Ist die gemeinsame Verteilung im fusionierten Datensatz erhalten geblieben, so sind alle statistischen Analysen möglich, welche auch mit dem Originaldatensatz durchgeführt werden können (Cielebak und Rässler, 2014, vgl. S. 378). Der fusionierte Datensatz kann demnach als eine Einzelquelle (single-source Stichprobe) behandelt werden (Rässler, 2002, vgl. S.31). Allerdings kann die gemeinsame Verteilung nur reproduziert werden, wenn Annahmen, wie beispielsweise die CIA, erfüllt sind und die spezifischen Variablen Y und Z gegeben der gemeinsamen Variablen **X** unabhängig voneinander sind (Rässler, 2002, vgl. S.31).

Für die Überprüfung der zweiten Validitätsebene schlägt Rässler (2002, vgl. S.153) vor, für den Original- und den fusionierten Datensatz paarweise Kontingenztabellen mit den spezifischen Variablen Y und Z zu erstellen. Anschließend kann jeweils ein Korrelationskoeffizient (siehe beispielsweise Jann, 2011, Kap. 4.4.2) zwischen den Zellhäufigkeiten im Fusionierten und im Originaldatensatz berechnet werden oder die Zellhäufigkeiten der beiden Kontingenztabellen kann durch einen χ^2-Test auf Homogenität (siehe beispielsweise Kohn, 2005, Kap. 17.3) getestet werden. Bei der Verwendung des χ^2-Testes kann anschließend ausgewertet werden, wie oft die Nullhypothese des Testes nicht verworfen wurde und somit die Verteilungen gleich sind. Ein Nachteil bei dieser Vorgehensweise ist jedoch, dass nicht der exakte Erhalt der gemeinsamen Verteilung $f_{\mathbf{X}YZ}$ untersucht wird, sondern nur der von f_{YZ}.

Eine weitere Möglichkeit zwei Wahrscheinlichkeitsverteilungen miteinander zu vergleichen ist, wie auch schon in Kapitel 2.3.1 beschrieben, über die Hellinger-Distanz H_D möglich. Wird die Gleichung 2.36, die auf eine univariate Verteilung bezogen ist, umgestellt und auf die gemeinsame Verteilung von (X, Y, Z) übertragen, ergibt sich die Hellinger-Distanz nach Gleichung 2.52.

$$H_D(f_{XYZ}, \tilde{f}_{XYZ}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\sqrt{p_{ijk}} - \sqrt{\tilde{p}_{ijk}})^2} \tag{2.52}$$

Dabei sind p_{ijk} und $\tilde{p}_{ijk}$ die relativen Häufigkeiten jeder Zelle der Kontingenztabelle $X \times Y \times Z$ des Originaldatensatzes und $X \times Y \times \tilde{Z}$ des fusionierten Datensatzes. Durch die Begrenzung ($0 \leq H_D \leq 1$) kann die Hellinger-Distanz verschiedener Berechnungen gut miteinander verglichen werden. Für den Grenzwert, bis zu dem Wert die Verteilungen als ähnlich angesehen werden können, kann die Faustregel eines Grenzwertes von $H_D = 0,05$ verwendet werden (Leulescu und Agafiţei, 2013, vgl. S.14).

Dritte Validitätsebene: Erhalt der Korrelationsstruktur

Bei dem Erhalt der Korrelationsstruktur beziehungsweise bei qualitativen Variablen beim Erhalt der Assoziationsstruktur, geht es um den Erhalt von Zusammenhängen zwischen den nicht beobachteten Variablen (Endres, 2019, vgl. S.53). Werden die beiden ersten Validitätsebenen nicht erreicht, können mit dem fusionierten Datensatz nicht mehr alle Analysen durchgeführt werden. Jedoch ist es oft ausreichend, nach der Datenfusion eine Aussage über die Korrelation beziehungsweise Assoziation zwischen den nicht beobachteten Variablen geben zu können. Demnach wird in dieser Ebene geprüft, wie gut die Datenfusion die Zusammenhänge zwischen den spezifischen Variablen reproduziert hat. Die Korrelation wird genau dann exakt im fusionierten Datensatz wiedergegeben, wenn Y und Z im Durchschnitt bedingt unkorreliert, gegeben X sind. (Kiesl und Rässler, 2005, vgl. S.27)

Untersucht werden kann die dritte Validitätsebene durch die Abweichungen der Korrelationskoeffizienten beziehungsweise der Assoziationsmaße der Variablen aus dem fusionierten Datensatz zu denen aus dem Originaldatensatz. Rässler (2002, S.151f) verwendet dafür zum Beispiel den Spearman'sche Rangkorrelationskoeffizient für quantitative Variablen. Endres verwendet dagegen für die Untersuchung des Zusammenhanges den korrigierten Kontingenzkoeffizienten (engl. Sakoda's adjusted Pearson's C), welcher speziell für qualitative Variablen geeignet ist (Endres, 2019, zum Beispiel S.55). Der korrigierte Kontingenzkoeffizient $C^* \in [0,1]$ wird nach Gleichung 2.53 auf Basis des Chi-Quadrat-Koeffizienten χ^2 (Vergleich Gleichung 2.39) berechnet (Stocker und Steinke, 2016, vgl. S.121).

$$C^*(Y,Z) = \sqrt{\frac{\chi^2}{\chi^2+n}}\sqrt{\frac{M}{M-1}}, \qquad M = min(J,K) \tag{2.53}$$

J und K sind die Anzahl der Kategorien, der spezifischen Variablen Y und Z, und demnach M das Minimum der Kategorienanzahl. Der Vorteil des Koeffizienten ist, dass er sowohl von der Anzahl der Beobachtungen als auch von der Dimension der Kontingenztabelle unabhängig ist. Je näher der Wert des Koeffizienten an 1 liegt, umso stärker ist der Zusammenhang (Stocker und Steinke, 2016).

In Rässler (2002, vgl. S.155) und Grömping et al. (2007, vgl. S.140f) wird ein weiterer Ansatz für die Untersuchung der dritten Validitätsebene verwendet, der auf einer linearen Regressionsanalyse basiert. Es wird jeweils eine lineare Regressionsanalyse in dem fusionierten und dem originalen Datensatz durchgeführt und anschließend die Regressionskoeffizienten miteinander verglichen. Für eine weitere Vertiefung sei zum Beispiel auf Jann (2011, Kap. 6) verwiesen.

Vierte Validitätsebene: Erhalt der Randverteilungen

Die Erfüllung der vierten Validitätsebene ist die Mindestanforderung an eine Datenfusion (Cielebak und Rässler, 2014, vgl. S.379). Im Gegensatz zu den anderen Validitätsebenen kann diese auch in der praktischen Anwendung untersucht werden. Aus diesem Grund gilt eine Datenfusion häufig als erfolgreich, wenn die Güte dieser Ebene erreicht ist.

Dies ist der Fall, wenn nach Gleichung 2.54 die Randverteilung der fusionierten Variable sowie nach Gleichung 2.55 die gemeinsame Verteilung der fusionierten Variable mit den Matching Variablen im fusionierten Datensatz erhalten werden können (Rässler, 2002, S.30f).

$$\tilde{f}_Z = f_Z \tag{2.54}$$

$$\tilde{f}_{ZX} = f_{ZX} \tag{2.55}$$

Bei einer erfolgreichen Datenfusion führen die anschließenden Analysen in Bezug auf die Verteilung von $(\mathbf{X}, \tilde{Z})$ und (Z) im fusionierten Datensatz zu den gleichen Schlüssen wie die Analyse von $(\mathbf{X}, Z)$ und (Z) im Spenderdatensatz (Rässler, 2002, vgl. S.31).

Für die Untersuchung der vierten Validitätsebene werden unter anderem Signifikanztests, wie der χ^2-Test (Vergleich Gleichung 2.39), der Kolmogorov-Smirnov-Test oder der t-Test verwendet. Welcher Test geeignet ist, hängt dabei vom Variablentypen ab (Rässler, 2002, vgl. S.150f). Der t-Test vergleicht die Verteilungen anhand der Mittelwerte bei normalverteilten Stichproben und der Kolmogorov-Smirnov-Test berechnet die maximale Distanz zwischen den Verteilungen. Demnach sind beide Tests für quantitative Variablen ausgelegt und werden aus diesem Grund in der Arbeit nicht näher erläutert, für eine Vertiefung dieser Thematik sei zum Beispiel auf Jann (2011, Kap. 5.6.3) verwiesen.

Wie bereits in Bezug auf die Identifikation der gemeinsamen Variablen (Vergleich Kapitel 2.3.1) und der zweiten Validitätsebene erwähnt, können Verteilungen durch unterschiedliche Ähnlichkeits- bzw. Unähnlichkeitsmaße miteinander verglichen werden. Endres (2019, vgl. S.54) verwendet zum Beispiel die Jensen-Shannon-Divergenz (Vergleich Gleichung 2.37), um die Randverteilung der fusionierten Variable im fusionierten Datensatz mit dem Spenderdatensatz zu vergleichen. In Leulescu und Agafiţei (2013, vgl. S.20) wird speziell für die vierte Ebene die Hellinger-Distanz (Vergleich Gleichung 2.36) verwendet.

3 Aufbau der simulierten Datensätze

In dieser Arbeit bilden die Daten der Straßenverkehrsunfallstatistik aus dem Jahr 2013[8] die Datengrundlage für die Evaluierung von fusionierten Unfalldaten. Insgesamt haben bestimmte Variablen sowie die Größe der zu fusionierenden Datensätze einen Einfluss auf die Verwendung der Analysemethoden bei der Vorbereitung der Datenfusion sowie auf die Datenfusionsmethode selbst. Um anschließend die Fusion beispielsweise auch auf andere Datenbanken anzuwenden, werden die Datensätze, die aus der Straßenverkehrsunfallstatistik erstellt werden, exemplarisch an die Rahmenstruktur der GIDAS und EUSKa Datenbanken angepasst sowie an die allgemeine Ausgangssituation der Datenfusion. Der Vorgang zur Datensatzerstellung wird in drei Verfahrensschritte untergliedert:

- Aufbereitung der Straßenverkehrsunfallstatistik
- Aufteilung in zwei Datensätze
- Auswahl der spezifischen Variablen Y und Z.

Auf die genannten Verfahrensschritte wird in den folgenden Unterkapiteln näher eingegangen.

3.1 Aufbereitung der Straßenverkehrsunfallstatistik

Bei der Aufbereitung der Straßenverkehrsunfallstatistik liegt der Fokus darauf, die Datensatzstruktur sowie die Variablen an die benötigte Rahmenstruktur anzupassen. Dabei werden folgende Schritte angewendet:

- Datensatzumstellung
- Löschen von unwichtigen Variablen
- Kategorisierung von Variablen
- Löschen von Variablen, bei denen die Merkmalsausprägungen nicht passen.

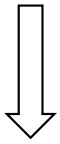
Die Merkmalsträger der Daten der Straßenverkehrsunfallstatistik sind die Unfallbeteiligten[9]. Für jeden Unfallbeteiligten sind sowohl Variablen hinsichtlich der Beteiligung aufgelistet, als auch direkt zum Unfall. Aufgrund dieser Tatsache wiederholen sich die Unfallangaben, insofern es mehr als einen Unfallbeteiligten pro Unfall gibt.

[8] FDZ der Statistischen Ämter des Bundes und der Länder, 2013.

[9] Als Unfallbeteiligte werden alle Fahrzeuge, Fußgänger sowie andere Personen erfasst, die selbst\deren Fahrzeug Schäden erlitten oder hervorgerufen haben (Statistisches Bundesamt (DESTATIS), 2019a, vgl. S.12).

Für die im Rahmen dieser Arbeit angewendete Simulation wird allerdings das Unfallszenario als Merkmalsträger benötigt, so dass im Datensatz keine Angaben zu ein und dem selben Unfall wiederholt aufgelistet sind. Um dieses Problem zu lösen, wird der Datensatz umgestellt, indem die einzelnen Angaben zu den Unfallbeteiligten jeweils an die Angaben zum Unfall durch neue Variablen angehängt werden. In Abbildung 3.1 ist dieser Vorgang schematisch dargestellt.

Unfallnr.	Angaben zum Unfall			Anzahl der Beteiligten	Nr. des Beteiligten	Angaben zum Unfallbeteiligten		
1				2	1			
1				2	2			
2				1	1			



Unfallnr.	Angaben zum Unfall			Anzahl der Beteiligten	Angaben zum 1. Unfallbeteiligten			Angaben zum 2. Unfallbeteiligten		
1				2						
2				1				-	-	-

Abbildung 3.1: Schematische Darstellung des angepassten Datensatzes (Eigene Darstellung)

Im weiteren Verlauf dieser Arbeit wird der umgestellte Datensatz der Straßenverkehrsunfallstatistik Simulationsdatensatz genannt, um Verwechslungen mit dem Originaldatensatz zu vermeiden. Anschließend beinhaltet jede Einheit „Unfallszenario“ im Simulationsdatensatz Angaben zum direkten Unfallgeschehen sowie Angaben zu den jeweiligen Beteiligten. Zusammen beschreiben sie den Unfallvorgang. Allerdings tritt nicht bei jedem Unfall die gleiche Anzahl an Unfallbeteiligten auf. Somit fehlen teilweise die Angaben zu einem weiteren Unfallbeteiligten (beispielsweise dem vierten Unfallbeteiligten). In Abbildung 3.1 ist dies durch das Symbol „-“ gekennzeichnet. Durch eine Prüfung bezüglich der Anzahl von Unfallbeteiligten zeigt sich, dass bei 70%[10] der Unfälle weniger als drei Unfallbeteiligte involviert sind und somit die Angaben fehlen. Aufgrund dieser Ergebnisse werden ausschließlich die Angaben zu den ersten beiden Unfallbeteiligten im Simulationsdatensatz beibehalten. Nachfolgend werden die leeren Merkmalsausprägungen bei den Variablen zu einem zweiten Unfallbeteiligten (treten auf, wenn es nur einen Unfallbeteiligten gibt) kategorisiert, indem die leeren Beobachtungen in die Kategorie „ohne zweiter Beteiligter“ eingeteilt werden.

Des Weiteren werden im Simulationsdatensatz die Variablen gelöscht, die inhaltlich für die Datenfusion keinen Nutzen aufweisen. Dabei handelt es sich überwiegend um zeitliche und detaillierte Ortsangaben sowie Angaben zu geschädigten Mitfahrern. Außerdem gibt es vereinzelte Variablen, die doppelt oder dreifach aufgeführt sind.

[10] FDZ der Statistischen Ämter des Bundes und der Länder, 2013.

Ein Beispiel hierfür ist die Variable Straßenzustand, der sich in „Straßenzustand 1. Kennziffer“ sowie „Straßenzustand 2. Kennziffer“ gliedert. Somit können sich pro Unfall bis zu zwei Angaben zum Straßenzustand im Datensatz befinden. Allerdings sind in über 90%[11] der Unfälle keine zweite oder dritte Angabe aufgelistet. Daher werden diese Variablen gelöscht und nur die erste Kennziffer für die weitere Analyse betrachtet. Eine Auflistung der betroffenen Variablen sowie die genaue Prozentzahl der fehlenden Einträge sind im Anhang A.5 aufgeführt.

Bis auf vereinzelte Variablen sind im Datensatz ausschließlich kategoriale Variablen enthalten, weswegen der Fokus dieser Arbeit auf der Fusion von kategorialen Daten liegt. Aus diesem Grund werden das Alter der unfallbeteiligten Personen sowie allgemein die Anzahl der Unfallbeteiligten in Kategorien eingeteilt, die sich an eine mögliche Einteilung des Statistisches Bundesamt (DESTATIS) (2019b) orientieren.

In der Straßenverkehrsunfallstatistik sind sowohl Unfälle mit Personenschaden als auch Unfälle mit Sachschaden aufgeführt. Für eine realistische Abbildung von GIDAS und EUSKa wird im Simulationsdatensatz nach Unfällen mit Personenschaden gefiltert. Eine weitere Filterung aufgrund der Datenbanken wird bezüglich der Art der Verkehrsbeteiligung vorgenommen. Von Interesse sind Unfälle mit mindestens einer PKW-Beteiligung. Im Simulationsdatensatz kann dies durch eine Filterung vorgenommen werden. Da es jedoch mit dem vorhanden Datensatz nicht möglich ist, direkt nach der PKW-Beteiligung zu filtern, kann nur grob über die Variable Fahrzeugklasse gefiltert werden. Der Simulationsdatensatz besteht abschließend aus 29 Variablen, die genaue Definition der Variablen sowie deren Kategorien sind in Anhang A.6 tabellarisch aufgelistet.

3.2 Aufteilung in zwei Datensätze

Mit den vorliegenden Daten kann noch keine Fusion durchgeführt werden, da nur ein Datensatz vorhanden ist. Um den Datensatz an die Rahmenstruktur der Datenfusion anzupassen, wird der Simulationsdatensatz in zwei Datensätze aufgeteilt.

In der Literatur werden keine Angaben zu dem Verhältnis zwischen den polizeilich erfassten Daten und den durch GIDAS erhobenen Daten gemacht. Aus diesem Grund wird das Verhältnis eigenständig durch die Verwendung von Straßenkarten, der Beschreibung des Erhebungsgebietes von GIDAS (Verkehrsunfallforschung an der TU Dresden GmbH, 2019) sowie des Unfallatlas für 2017 des statistischen Bundesamtes (Statistische Ämter des Bundes und der Länder, 2019) ermittelt. In GIDAS werden, wie in Kapitel 2.1 beschrieben, pro Jahr ca. 1000 Unfälle in dem Erhebungsgebiet Dresden erfasst. Der Abbildung 3.2 ist zu entnehmen, welche Gemeinden das Erhebungsgebiet annähernd umfasst.

[11] FDZ der Statistischen Ämter des Bundes und der Länder, 2013.

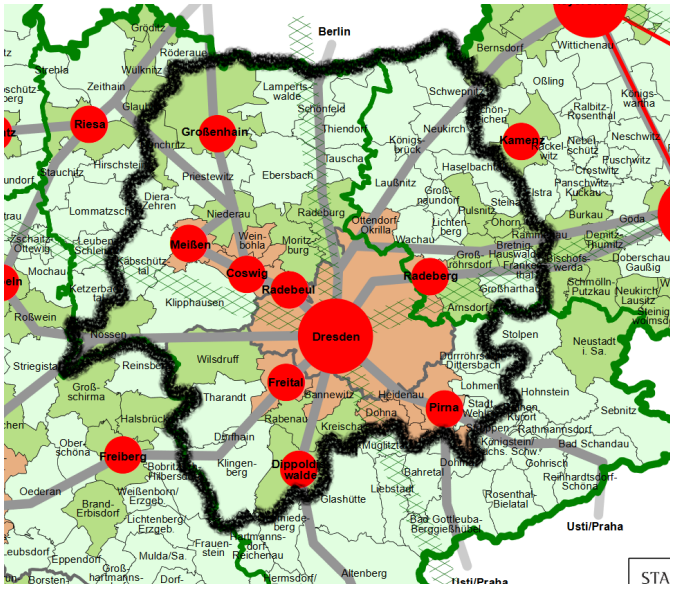


Abbildung 3.2: Erhebungsgebiet Dresden (Ausschnitt aus Karte Sächsisches Staatsministerium des Inneren (2019) mit eigenen Ergänzungen)

Anschließend werden anhand dieser Eingrenzung die Unfalldaten aus dem Unfallatlas nach den markierten Gemeinden gefiltert und die Anzahl der Unfälle ermittelt. Die durch den Unfallatlas aufgezeichneten Daten ergeben dabei, dass die Anzahl der von der Polizei registrierten Unfälle viermal so groß ist wie die Anzahl der in GIDAS erfassten Unfälle. Aus diesem Grund werden aus dem Simulationsdatensatz ohne zurücklegen zwei unabhängige Datensätze mit dem Verhältnis 1:4 gezogen. Datensatz A besteht anschließend aus 2240 und Datensatz B aus 8960 Unfällen.

3.3 Auswahl der spezifischen Variablen

Aufgrund der Tatsache, dass es sich um simulierte Datensätze handelt, ist es grundsätzlich möglich, alle Variablen als spezifische Variablen auszuwählen, da die spezifischen Variablen für die Simulation aus dem jeweiligen Datensatz raus gelöscht werden und nur theoretisch fehlen. Des Weiteren besteht die Möglichkeit, mehrere Variablen als spezifische Variablen auszuwählen. In dieser Arbeit wird jedoch die Variablenanzahl auf mehrere gemeinsame und jeweils eine spezifische Variable in den Datensätzen beschränkt. Es wird vorerst eine Gruppe von vier Variablen ausgewählt, die verschiedene Bereiche des Unfallgeschehens beschreibt. Eine Übersicht über die vier ausgewählten Variablen ist der Abbildung 3.3 zu entnehmen. Zum einen wird durch die *Ortslage* (Variablenbezeichnung: uortslage) eine Variable ausgewählt, die die Umgebung und damit die Ausgangslage beschreibt. Zum anderen fällt die Wahl auf eine Variable, die eine Aussage über das Ergebnis eines Unfalls trifft, wie beispielsweise *Kfz fahrbereit* (ufahrbkfz), die beschreibt, ob ein Fahrzeug noch fahrbereit ist. Bezüglich des direkten Unfallgeschehens werden die Variablen *Unfalltyp* (utyp1) und *Unfallart* (uart) ausgewählt. Der Unfalltyp beschreibt eine Situation, die anschließend zum Unfall führt. Der Unfall selbst wird durch die Unfallart beschrieben.

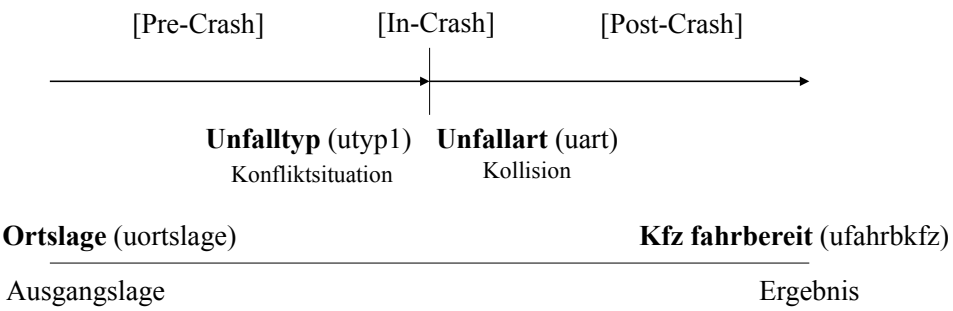


Abbildung 3.3: Zur Verfügung stehende spezifische Variablen (Eigene Darstellung)

Folgendes Beispiel soll die Unterscheidung der beiden Variablen verdeutlichen. Die Konfliktsituation ist ein in Längsrichtung fahrendes Fahrzeug, das nicht abgebogen ist sowie ein Fußgänger, der die Fahrbahn überschreitet, und folglich nicht in Längsrichtung unterwegs ist. Eingeteilt wird der Unfalltyp in die Kategorie „Überschreiten-Unfall“. Aufgrund dieser Situation kann es zu mehreren unterschiedlichen Unfällen kommen, bei denen es nicht zwangsläufig zu einer Kollision zwischen Fußgänger und Fahrzeug kommen muss. Weitere Möglichkeiten sind, dass aufgrund eines Ausweichmanövers das Fahrzeug von der Fahrbahn abkommt, oder eine Vollbremsung des Fahrzeuges zu einem Auffahrunfall mit einem nachfolgendem Fahrzeug führt. In jedem der beschriebenen Fälle wird der Unfallart eine andere Kategorie zugewiesen. (Unfallforschung der Versicherer, 2016)

Für die explizite Auswahl der spezifischen Variablen werden aus den vier Variablen Kombinationen gebildet. Die Auswahl erfolgt dabei nach einem Kreuzprinzip, damit die Variablenkombination jeweils unterschiedliche Bereiche des Unfalls abdeckt. Dafür wird eine Variable ausgewählt, die eine Eigenschaft beschreibt, die vor dem Unfall eingetroffen ist (uortslage oder utyp1) und eine Variable, die die Situation während, beziehungsweise nach dem Unfall beschreibt (uart oder ufahrbkfz) (Vergleich Abbildung 3.3). Die Kombination aus uart und ufahrbkfz wird dabei jedoch nicht weiter betrachtet, da sie keine Variable enthält, die direkt mit dem Unfallgeschehen in Verbindung steht. Demnach stehen für die Auswahl der spezifischen Variablen Y und Z die folgenden drei Kombinationen zur Auswahl:

- Y: Unfalltyp (utyp1) und Z: Unfallart (uart)
- Y: Kfz fahrbereit (ufahrbkfz) und Z: Unfalltyp (utyp1)
- Y: Ortslage (uortslage) und Z: Unfallart (uart).

Allgemein werden für die Datenfusion gemeinsame Variablen (Matching Variablen) verwendet, die einen starken Zusammenhang zu den spezifischen Variablen aufweisen (Vergleich Kapitel 2.3.2). Aus diesem Grund erfolgt die Auswahl der spezifischen Variablen nach dem Prinzip, dass es genügend gemeinsame Variablen gibt, die die spezifischen Variablen gut erklären. Folglich werden für die drei Kombinationen jeweils die Zusammenhänge zwischen den spezifischen und den gemeinsamen Variablen berechnet und die Kombination von spezifischen Variablen ausgewählt, die die meisten und aussagekräftigsten Variablen beinhalten. Als Assoziationsmaß wird der Unsicherheitskoeffizient U (Vergleich Gleichung 2.43) verwendet.

Anhand des Datensatzes A werden die Zusammenhänge zwischen den gemeinsamen Variablen und der spezifischen Variable Y berechnet, im Gegensatz dazu im Datensatz B mit der spezifischen Variable Z. In der Literatur weisen ab einem Unsicherheitskoeffizient von 0,1 die betrachteten Variablen einen mittleren Zusammenhang auf (vgl. Kapitel 2.3.2). Aus diesem Grund wird nur der Bereich $U \geq 0,1$ betrachtet und die Anzahl der Variablen in Abhängigkeit des Unsicherheitskoeffizienten aufsummiert, siehe Tabelle 3.1. Die dafür zugrunde liegenden Ergebnisse des Unsicherheitskoeffizientens für jede Kombination der spezifischen Variablen sind im Anhang A.7 aufgeführt.

Tabelle 3.1: Anzahl der gemeinsamen Variablen für jede Kombination in Bezug auf den Unsicherheitskoeffizienten[12]

	Kombination 1		Kombination 2		Kombination 3	
	utyp1 (Y)	uart (Z)	ufahrbkfz (Y)	utyp1 (Z)	uortslage (Y)	uart (Z)
$0,4 \leq U$	1	1	0	2	0	2
$0,3 \leq U < 0,4$	0	0	0	0	0	0
$0,2 \leq U < 0,3$	0	1	0	0	2	1
$0,1 \leq U < 0,2$	7	7	3	7	8	7
	$\sum 17$		$\sum 12$		$\sum 20$	

Der Tabelle 3.1 ist zu entnehmen, dass Kombination 2, aus Kfz fahrbereit (ufahrbkfz) und Unfalltyp (utyp1) die wenigsten gemeinsamen Variablen (12) mit mindestens einem mittleren Zusammenhang aufweisen. Der Grund dafür liegt an „ufahrbkfz“, da in Bezug zu dieser Variable nur drei gemeinsame Variablen einen Zusammenhang haben. Im Vergleich dazu besitzen die anderen beiden Kombinationen ein ausgeglicheneres Verhältnis der gemeinsamen Variablen. Darüber hinaus ist bei diesen beiden Kombinationen die Anzahl der Variablen höher, die einen Zusammenhang aufweisen. Kombination 1 aus Unfalltyp (utyp1) und Unfallart (uart) besitzt insgesamt 17 gemeinsame Variablen, Kombination 3 aus Ortslage (uortlsage) und Unfallart (uart) 20. Zudem besitzt die dritte Kombination mehr gemeinsame Variablen, die einen starken Zusammenhang von $U \geq 0,2$ aufweisen. Aufgrund dieser Tatsachen werden für die Simulationsdatensätze die dritte Kombination als spezifische Variablen verwendet.

Auf Grundlage der in diesem Kapitel durchgeführten Verfahrensschritte werden die beiden Simulationsdatensätze für die Datenfusion, wie in Abbildung 3.4, aufgebaut. Datensatz A besteht demnach aus $n_A = 2240$ beobachteten Unfällen mit der spezifischen Variable „uortslage“ (Y), Datensatz B aus $n_B = 8960$ Unfällen und der spezifischen Variable „uart“ (Z). Die gemeinsamen Variablen bestehen insgesamt aus 27 Variablen und sind in der mittleren Spalte abgebildet.

[12] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

	Y	$X_1 \ldots X_{27}$	Z
Datensatz A	uortslage	ustunde, uwochentag umonat, uanzbet ukategorie, uursache ucharunfst, ubesonunfst utyp1, utyp2 ustrklasse, ustrzust uaufprhind, ufahrbkfz	
Datensatz B		ugeschwbegr, ulichtverh ulichtzanl bursache_01, bursache_02 bunfallfolge_01, bunfallfolge_02 bgeschlecht_01, bgeschlecht_02 balter_c_01, balter_c_02 bfzgklasskba_01, bfzgklasskba_02	uart

Abbildung 3.4: Aufteilung der Variablen in den Simulationsdatensätzen A und B (Eigene Darstellung)

4 Datenfusion

In diesem Kapitel werden die einzelnen Schritte der Datenfusion mit den simulierten Datensätzen aus Kapitel 3 durchgeführt und in Abhängigkeit der Eigenschaften der vorhandenen Daten und Anforderungen eine Auswahl einzelner Verfahren getroffen. Der Ablauf der Datenfusion und die dafür zur Verfügung stehenden Verfahren, die in Kapitel 2 beschrieben sind, sind der Abbildung 4.1 zu entnehmen. In Abhängigkeit der Datensituation und der Ziele können für die einzelnen Schritte der Datenfusion unterschiedliche Methoden ausgewählt werden. Die für diese Arbeit ausgewählten Verfahren werden nachfolgend genauer vorgestellt und sind in Abbildung 4.1 blau hervorgehoben.

<table>
<tr><th></th><th>Ablauf Datenfusion</th><th colspan="5">Verfahren</th></tr>
<tr><td rowspan="2">Vorbereitung</td><td>Harmonisierung Vergleich Verteilung</td><td>X^2-Test</td><td>Kolmogorov-Smirnov Test</td><td>Bhattacharyya Koeffizient</td><td>Hellinger Distanz</td><td>Jensen-Shannon Divergenzmaß</td></tr>
<tr><td>Auswahl Matching Variablen</td><td>X^2-Test</td><td>Unsicherheits-koeffizient</td><td>Cramers V</td><td>Bestimmtheitsmaß</td><td>Verringerung Unsicherheit</td></tr>
<tr><td rowspan="3">Fusionsrahmen</td><td>Ansatz</td><td colspan="2">parametrisch</td><td colspan="3">nichtparametrisch</td></tr>
<tr><td>Ziel</td><td colspan="2">Mikroansatz</td><td colspan="3">Makroansatz</td></tr>
<tr><td>Datenfusion</td><td colspan="2">CIA</td><td>Hilfsinformationen</td><td colspan="2">Beibehalten der Unsicherheit</td></tr>
</table>

Abbildung 4.1: Überblick über die verschieden Verfahren in Abhängigkeit der Schritte der Datenfusion (Eigene Darstellung)

Zuerst werden in diesem Kapitel die Datensätze auf die Datenfusion vorbereitet, indem die gemeinsamen Variablen identifiziert und die Matching Variablen für die Fusion ausgewählt werden. Anschließend werden drei verschiedene Methoden für die Datenfusion durchgeführt, die auf die Rahmenstruktur der ausgewählten Punkte in Abbildung 4.1 passen.

Bei der Verwendung von parametrischen Verfahren wird davon ausgegangen, dass die Daten aus einer bestimmten Verteilung stammen. Trifft dies allerdings nicht zu, und das Verfahren wird fälschlicherweise angewendet, sind die Ergebnisse unzulässig, da die imputierten Werte „künstlich“ erzeugt und demzufolge falsch geschätzt werden (D’Orazio, 2013, vgl. S.68).

Bei nichtparametrischen Verfahren werden die imputierten Werte dagegen durch tatsächlich beobachtete Werte erzeugt und sind in deutlich mehr Situationen zulässig, da die Daten keiner bestimmten Verteilung zugeordnet werden müssen. Aus diesem Grund werden die Verfahren auch als robuste Verfahren bezeichnet (D'Orazio et al., 2006). Wird ein nichtparametrisches Verfahren verwendet, und es handelt sich eigentlich um eine bestimmte Verteilung, sind die Ergebnisse zulässig, allerdings mit einer geringeren Aussagekraft als unter der Verwendung von parametrischen Verfahren (D'Orazio, 2013, vgl. S.68). Da die Ergebnisse im Falle einer falschen Verwendung von parametrischen Verfahren unzulässig werden und darüber hinaus die nichtparametrischen Verfahren robuster sind, fällt die Entscheidung auf eine nichtparametrische Variante.

Für das Ziel der Datenfusion wird ein Mikroansatz gewählt. Zum einen ist die Voraussetzung dieser Arbeit eine Erstellung eines real fusionierten Datensatzes, zum anderen wird dieser in der Literatur häufiger verwendet. Der Grund dafür ist die anschließende leichtere Handhabung und Analyse mit einem realen Datensatz (D'Orazio et al., 2006, vgl. S.3). Des Weiteren wird aufgrund der unterschiedlich großen Datensätze eine asymmetrische Datenfusion durchgeführt, indem ausschließlich die spezifische Variable in den kleineren Datensatz fusioniert wird. Konkret für die simulierten Datensätze bedeutet das, dass der Datensatz B mit $n_B = 8960$ die Rolle des Spenders einnimmt und die spezifische Variable *uart* in den Empfängerdatensatz A mit $n_A = 2240$ fusioniert wird.

Letztendlich muss noch entschieden werden, mit welchem Ansatz das Identifikationsproblem gelöst wird. Der Ansatz durch das Heranziehen zusätzlicher Informationen wird aufgrund von fehlenden Informationen nicht verwendet. Die CIA ist eine Annahme, die bei realen Datensätzen nicht getestet werden kann und aufgrund der Datenmenge an Variablen der vorhandenen simulierten Datensätze schwierig zu testen ist. Einerseits wird der Ansatz in der Literatur kritisiert und auf falsche Ergebnisse aufgrund einer fehlerhaften Annahme hingewiesen (Barry et al., 1982, vgl. S.45). Auf der anderen Seite ist dieser Ansatz der in der Literatur am weitesten verbreitete Ansatz, der in vielen verschiedenen realen Studien verwendet wird und auch positive Ergebnisse aufweist (siehe z.B. Grömping et al., 2007 und Webber und Tonkin, 2013). Unabhängig von der Gültigkeit wird aufgrund der Popularität der Ansatz der CIA getestet, um anschließend eine Aussage in Bezug auf Unfalldaten treffen zu können. Explizit als Methode unter der CIA werden sowohl die *Distanz-Hot-Deck* Variante, als auch eine Kombination aus *Distanz-Hot-Deck* und *Random-Hot-Deck*, verwendet. Aufgrund einer fehlenden quantitativen Variable kann die Methode des *Rank-Hot-Deck* Verfahrens nicht angewendet werden.

Des Weiteren wird zusätzlich der Ansatz angewendet, der die Unsicherheit beibehält. Dafür wird die Methode der *Imprecise Imputation* verwendet. Mit diesem Ansatz wird geprüft, wie sich die Ergebnisse ohne Hinzunahme von Annahmen oder weiteren Informationen auswirken.

Alle Berechnung wurden mit der Software R (R Core Team, 2019) durchgeführt. Eine Übersicht aller verwendeten R-Pakete ist im Anhang A.1 aufgelistet.

4.1 Datensatzvorbereitung für Datenfusion

Die Vorbereitung der Datensätze spielt eine wichtige Rolle bei der Datenfusion. Eine Grundvoraussetzung ist, dass die Datensätze aus der gleichen Grundgesamtheit stammen. Für die simulierten Datensätze ist diese Anforderung erfüllt, da beide Datensätze eine Stichprobe aus dem gleichen Datensatz (Straßenverkehrsunfallstatistik) sind. Wie in Kapitel 2.3 bereits dargestellt, wird die Vorbereitung des Datensatzes in drei Schritte aufgeteilt. Der erste Schritt, die Identifizierung und Analyse der spezifischen Variablen, braucht in dieser Arbeit nicht durchgeführt werden. Es handelt sich um eine Datenfusion mit simulierten Datensätzen, bei der die spezifischen Variablen frei aus der Variablenmenge bestimmt werden können und dies bereits in Kapitel 3.3 durchgeführt wurde. Die Weiteren Schritte der Vorbereitung werden in den folgenden Unterkapiteln näher beschrieben.

4.1.1 Auswahl der gemeinsamen Variablen

In der Literatur gliedert sich dieser Schritt in zwei Bereiche auf. Zum einen werden die Variablen in Bezug auf die Definition und Klassifizierung analysiert, zum anderen hinsichtlich ihrer Verteilung. Der erste Schritt braucht für die simulierten Datensätze nicht durchgeführt werden, da aufgrund der simulierten Rahmenstruktur die 27 gemeinsamen Variablen **X** bereits in Bezug auf die Definition und Klassifizierung übereinstimmen. Für den Vergleich der Verteilungen wird in dieser Arbeit die Hellinger-Distanz (Vergleich Gleichung 2.52) als Ähnlichkeitsmaß verwendet. Der Vorteil ist, dass sie leicht zu interpretieren ist und bereits in mehreren Studien als ein Maß verwendet (zum Beispiel (Leulescu und Agafiţei, 2013, vgl. S.14) und (Castillo und Villanueva-Garcia, 2018, vgl. S.6f)) oder empfohlen wurde (D'Orazio, 2013, vgl. S.164).

Es wird jeweils die Distanz zwischen der Verteilung der Variable im Datensatz A und der Verteilung der entsprechenden Variable im Datensatz B ermittelt. Die Berechnung erfolgt mittels R durch die Verwendung der comp.prop() Funktion im Paket StatMatch (D'Orazio, 2019). Als Grenzwert für die Hellinger-Distanz wird $H_D = 0,05$ verwendet, vergleiche (Leulescu und Agafiţei, 2013, S.14). In Tabelle 4.1 sind die Ergebnisse der einzelnen Variablen aufgelistet. Auf Grundlage dieser Ergebnisse werden bis auf die markierten Variablen alle anderen als ähnlich angesehen. Die markierten Variablen liegen über dem Grenzwert von 0,05 und werden für die weitere Betrachtung somit ausgeschlossen.

Jedoch ist anzumerken, dass im Vergleich zu den Variablen mit ähnlicher Verteilung die „unähnlichen" Variablen eine großen Anzahl an Kategorien besitzen. Eine mögliche Begründung dafür ist der Nachteil der Hellinger-Distanz, dass sie die Variabilität aufgrund einer großen Anzahl an Kategorien nicht berücksichtigt (Leulescu und Agafiţei, 2013, vgl. S.14). Dies könnte Gegenstand weiterer zukünftiger Untersuchungen sein.

Tabelle 4.1: Berechnung der Hellinger-Distanz H_D für den Vergleich der Verteilungen der gemeinsamen Variablen im Datensatz A und B[13]

Name Variable X_p	Hellinger-Distanz	Name Variable X	Hellinger-Distanz
ustunde	0,035	ugeschwbegr	0,043
uwochentag	0,027	ulichtverh	0,016
umonat	0,031	ulichtzanl	0,008
uanzbet	0,028	**bursache_01**	**0,080**
ukategorie	0,005	**bursache_02**	**0,071**
uursache	0,040	bunfallfolge_01	0,023
ucharunfst	0,022	bunfallfolge_02	0,013
ubesonunfst	0,023	bgeschlecht_01	0,014
utyp 1	0,022	bgeschlecht_02	0,013
utyp 2	**0,078**	alter_c_01	0,022
ustrklasse	0,026	alter_c_01	0,028
ustrzust	0,010	**bfzgklasskba_01**	**0,065**
uaufprhind	0,027	**bfzgklasskba_02**	**0,058**
ufahrbkfz	0,015		

4.1.2 Auswahl Matching Variablen

Wie in Kapitel 2.3.2 ausführlich dargestellt, ist es nicht empfehlenswert, alle gemeinsamen Variablen als Matching-Variablen auszuwählen. D'Orazios (2019) Untersuchungen zeigen wie sinnvoll es ist, die Auswahl der Matching Variablen über die Verringerung der Unsicherheit zu berechnen, und dass dieser Ansatz jedoch bei einer großen Anzahl von gemeinsamen Variablen sehr rechenintensiv ist. Aus diesem Grund wird ein zweistufiges Verfahren angewendet, bei dem zuerst eine Teilmenge von Variablen mit einem herkömmlichen Ansatz ausgewählt und anschließend eine Analyse basierend auf der Unsicherheit durchgeführt wird. Nach dem Schritt der Identifikation der gemeinsamen Variablen stehen 22 Variablen als Auswahl für die Matching Variable zur Verfügung.

Auswahl der Teilmenge der Matching Variablen durch den klassischen Ansatz

In dieser Arbeit werden für die erste Auswahl der Matching Variablen der Unsicherheitskoeffizient nach Gleichung 2.43 verwendet. Die detaillierte Definition und Erklärung der Kennzahl ist in Kapitel 2.3.2 beschrieben. Die Auswahl des Unsicherheitskoeffizienten lässt sich dadurch begründen, dass er zu den PRE-Maßen gehört und diese bereits in der Literatur zur Analyse des Zusammenhangs bei unfallbezogenen Variablen angewendet wurde (Grömping et al., 2007, Kap. 4.5.3).

[13] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Außerdem beruht das Cramers V auf dem χ^2-Test, der wiederum nur valide ist, wenn mindestens 20% der erwarteten Häufigkeiten größer als 5 sind (Hanneman et al., 2012, S. 317). Da diese Voraussetzung nicht immer gegeben ist, wird die Verwendung des Cramers V für das weitere Vorgehen ausgeschlossen.

Die Berechnung in R erfolgt für den Unsicherheitskoeffizient über die Funktion pw.assoc() im Paket StatMatch (D'Orazio, 2019). Die Ergebnisse des Unsicherheitskoeffizienten (U) sind in Tabelle 4.2 für jede gemeinsame Variable in Bezug auf die spezifischen Variablen *uortslage* (Y) und *uart* (Z) dargestellt.

Tabelle 4.2: Unsicherheitskoeffizient U für alle möglichen Kombinationen der Variable X mit Y und Z[14]

X_p	A = „uortslage“ (Y)	B = „uart“ (Z)
	U	U
ustunde	0,0194	0,0251
uwochentag	0,0038	0,0062
umonat	0,0036	0,0091
uanzbet	0,1031	0,1588
ukategorie	0,0133	0,0156
uursache	0,0388	0,0326
ucharunfst	0,1154	0,1311
ubesonunfst	0,0237	0,0188
utyp1	0,1380	0,4102
ustrklasse	0,2946	0,0522
ustrzust	0,0148	0,0195
uaufprhind	0,0924	0,0988
ufahrbkfz	0,1024	0,0476
ugeschwbegr	0,2453	0,0280
ulichtverh	0,0044	0,0088
ulichtzanl	0,0461	0,0122
bunfallfolge_01	0,0508	0,0464
bunfallfolge_02	0,0945	0,1543
bgeschlecht_01	0,0032	0,0050
bgeschlecht_02	0,0981	0,1551
balter_c_01	0,0263	0,0205
balter_c_02	0,1100	0,1695

Es erfolgt eine Orientierung an der Einteilung aus Kapitel 2.3.2, die besagt, dass ein mittlerer Zusammenhang in Bezug auf U ab einem Wert von 0,1 gilt.

[14] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Der Unsicherheitskoeffizient identifiziert somit „ustrklasse“ als besten Prädiktor der spezifischen Variable „uortslage“ (Y), gefolgt von „ugeschwbegr“, „utyp1“, „ucharunfst“, „balter_c_02“, „uanzbet“ und „ufahrbkfz“.

In Bezug auf die Z Variable „uart“ identifiziert der Unsicherheitskoeffizient die Variable „utyp1“ als die Variable mit dem höchsten Zusammenhang. Mit deutlich kleineren Werten folgen „uanzbet“„balter_c_02“, „bgeschlecht_02“, „bunfallfolge_02“ und „ucharunfst“.

Wird die Auswahl der Matching Variablen durch den klassischen Ansatz betrachtet, bestätigen sich die in Kapitel 2.3.2 beschriebenen Problematiken. Die Auswahl der Matching Variablen durch die Schnittmenge ($X_Y \cap X_Z = X_M$) ergibt eine Teilmenge von vier Variablen $X_M =$ („utyp1“„uanzbet“, „ucharunfst“ und „balter_c_02„). Die Zahl der Variable ist damit eine angemessene Menge. Allerdings werden die beiden Variablen, die die besten Prädiktoren für „uortslage“ (Y) sind nicht mit verwendet. Im Gegenteil dazu ergibt die Vereinigungsmenge ($X_Y \cup X_Z = X_M$) eine Teilmenge von neun Variablen als Matching Variablen $X_M =$ („utyp1“, „ucharunfst“, „balter_c_02“, „ustrklasse“, „ugeschwbegr“, „uanzbet“, „ufahrbkfz“, „bgeschlecht_02“ und „bunfallfolge_02“) zurück. Ein Nachteil bei einer Auswahl von neun Matching Variablen ist, dass die gemeinsame Verteilung eine sehr große Kontingenztabelle von über 11 Millionen Zellen produziert. Die Zellenanzahl ergibt sich aus dem Produkt der Anzahl der Kategorien der Variablen:

$$7 \times 7 \times 6 \times 5 \times 20 \times 10 \times 2 \times 4 \times 5 = 11760000$$

In Bezug auf die Datensatzgrößen von 8960 und 2240 Beobachtungen kann von einer sehr spärlichen Tabelle ausgegangen werden, da die Anzahl der Zellen deutlich größer als die Beobachtungen sind und somit die Mehrheit der Zellen keine Beobachtungen aufweisen können. Wird für die Datenfusion beispielsweise ein nichtparametrisches Verfahren verwendet und die Beobachtungen der zu fusionierenden Variable anhand von Matching Variablen ausgewählt, kann es bei einer spärlichen Kontingenztabelle aufgrund der vielen leeren Zellen dazukommen, dass es keine geeignete Spender gibt.

Aufgrund des zweistufigen Auswahlverfahrens beruht die Auswahl der Matching Variablen nach dem klassischen Verfahren auf der Vereinigungsmenge der einzelnen Betrachtungen durch die Berechnung des Unsicherheitskoeffizienten, damit ausreichend Variablen zur Auswahl stehen und keine wichtigen Variablen, die einen starken Zusammenhang zu den spezifischen Variablen aufweisen, verloren gehen. Insgesamt werden die folgenden Matching Variablen für die zweite Auswahlstufe verwendet:

> $X_M =$ („utyp1“, „ucharunfst“, „uanzbet“, „balter_c_02„, „ustrklasse“, „ugeschwbegr“, „ufahrbkfz“, „bgeschlecht_02“, „bunfallfolge_02“).

Auswahl der Matching Variablen durch Unsicherheit

Nachdem eine Teilmenge an Variablen durch herkömmliche Verfahren ausgewählt wurden, wird in der zweiten Stufe untersucht, welche Variablen sich als Matching Variablen eignen, die anhand der Unsicherheitsmethode berechnet werden. Die Berechnung erfolgt in R über die Funktion selMtc.by.unc() im Paket StatMatch (D'Orazio, 2019).

Die Funktion führt die Rechnungen aus Kapitel 2.3.2 aus und berechnet die beste Teilmenge an Variablen. Das Ergebnis ist in Tabelle 4.3 zusammengefasst. Die Variablen, die die Unsicherheit in der geschätzten Kontingenztabelle von Y x Z am stärksten verringern sind unabhängig vom Strafterm *utyp1*, *ustrklasse* und *ugeschwebegr*. Die Unischerheiten werden sowohl mit dem ersten $u^{(1)}$ als auch mit dem zweiten Strafterm $u^{(2)}$ berechnet. Beide identifizieren die eben genannten Variablen als Matching Variablen, jedoch weist der zweite Strafterm insgesamt einen höheren Wert auf. Allerdings hat das keinen Einfluss auf die Auswahl der Matching Variablen. Insgesamt produziert die Verteilung der ausgewählten Matching Variablen eine Kontingenztabelle von 700 Zellen (H_{Dm}) und die Spärlichkeit in Bezug auf (X,Y) und (Z,Y) ergibt einen Wert von $\bar{n} = 1,28$, welcher demnach über dem Stoppkriterium von 1.

Tabelle 4.3: Auswahl der Matching Variablen durch die Unsicherheitsmethode[15]

Kombination von X (D_m)	Anzahl von X	H_{D_m}	u	$u^{(1)}$	$u^{(2)}$	$\bar{n}$
bedingungslos	0		0,10000000	0,10000000	0,10000000	896,00
utyp1	1	7	0,06038958	0,06039043	0,06083882	128,00
utyp1 ustrklasse	2	35	0,04207699	0,04208124	0,04253782	25,60
utyp1 ustrklasse ugeschwbegr	3	700	0,03192656	0,03201159	0,03311704	1,28

Werden die Auswahlmethoden miteinander verglichen, fällt auf, dass die Berechnung durch die Verringerung der Unsicherheit genau die Variablen ausgewählt hat, die auch bei der traditionellen Methode einen starken Zusammenhang zu den spezifischen Variablen aufweist. Für die Variable *uortslage* werden die drei Variablen mit dem stärksten Zusammenhang ausgewählt. In Bezug auf die Variable *uart* wurde ebenfalls durch *utyp1* die Variable identifiziert, die sich als bester Prädikator bei den herkömmlichen Methoden erwiesen hat. Aufgrund der Tatsache, dass die Ergebnisse der beiden Methoden gut zusammenpassen, werden für die Fusionsverfahren die folgenden drei Matching Variablen ausgewählt:

- Unfalltyp (utyp1)
- Straßenklasse (ustrklasse)
- Geschwindigkeitsbegrenzung (ugeschwbegr)

4.2 Distanz-Hot-Deck Verfahren

In Kapitel2.2.1 wurden bereits die Vor- und Nachteile der beschränkten und unbeschränkten *Distanz-Hot-Deck* Methoden erläutert. Die Randverteilung der imputierten Variable wird nach D'Orazio et al. durch das beschränkte Verfahren besser wiedergegeben. Vor allem bei einer gleichen Größe des Empfänger- und Spenderdatensatzes wird von einer perfekten Wiedergabe der Randverteilung der imputierten Variable ausgegangen. (D'Orazio et al., 2006, vgl. S. 42f)

[15] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Nachteil der beschränkten Variante ist, dass die durchschnittliche Distanz zwischen den Beobachtungen größer ist als bei der unbeschränkten Variante und somit nicht immer die ähnlichsten Beobachtungen fusioniert werden. Die Vor- und Nachteile dieser beiden Verfahren hängen insgesamt von dem Verhältnis der Datensatzgrößen ab. Bei den simulierten Datensätzen ist der Unterschied zwischen den Größen der Datensätze deutlich gegeben (1:4). Aus diesem Grund wird die These aufgestellt, dass sich die Ergebnisse der Verfahren nicht deutlich voneinander unterscheiden werden. Es ist zu erwarten, dass die Randverteilung beim beschränkten Verfahren nicht perfekt wiedergegeben wird, da nur ein kleiner Teil der Beobachtungen aus dem Spenderdatensatz fusioniert wird. Andererseits wird aufgrund der großen Spenderanzahl die durchschnittliche Distanz vermutlich kleiner und für die unbeschränkte Variante sind mit einer geringeren Anzahl von Überbenutzungen vereinzelter Spenderbeobachtungen auszugehen. Durch die aufgestellte These dominiert keine Methode mit ihren Vorteilen. Aus diesem Grund werden sowohl die beschränkte als auch die unbeschränkte Methode verwendet. Für die Lösung des Zuordnungsproblems bei der beschränkten Variante wird der in Kapitel 2.2.1 beschriebene Ungarische Algorithmus (Kuhn, 1955) verwendet. Durch die Verwendung der Ungarischen Methode findet die Fusion bei der beschränkten Variante bei jedem Durchlauf auf die gleiche Weise statt und es entsteht ein eindeutig synthetischer Datensatz. Bei der unbeschränkten Methode wird die auf die Matching Variablen bezogene ähnlichste Beobachtung fusioniert. Weisen zwei Spenderbeobachtungen beide die kleinste Distanz auf, wird zufällig eine ausgewählt. Folglich kann bei einer großen Anzahl an ähnlichen Spenderbeobachtungen das Fusionsergebnis variieren. Aus diesem Grund besteht eine Fusion der unbeschränkten Methode aus $k = 500$ Wiederholungen, um anschließend auszuwerten wie sich die Ergebnisse verändern. Für die Anzahl an Durchläufen wurde sich an anderen Studien orientiert (Gschwilm, 2017).

Auswahl des Distanzmaßes

Für die Auswahl des Distanzmaßes hat sich bei ersten Probefusionen ergeben, dass sich bei der beschränkten *Distanz-Hot-Deck* Methode die Fusionsergebnisse in Abhängigkeit der unterschiedlichen Distanzmaße aus Kapitel 2.2.1 nicht unterscheiden. Aus diesem Grund wurden die Distanzmaße anhand zweier Beispieldatensätze gesondert untersucht. Für jedes Distanzmaß wurde die Distanzmatrix berechnet und miteinander verglichen. Die Vorgehensweise der Analyse sowie die Ergebnisse sind dem Anhang A.8 zu entnehmen. Die Auswertung ergibt, dass sich die einzelnen Distanzen in den Distanzmatrizen in ihrem Wert abhängig von der Berechnung unterscheiden. Jedoch bleiben die Reihenfolgen von der kleinsten bis zur größten Distanz gleich. Demzufolge wird für die kleinste Distanz bei jedem Distanzmaß die gleiche Beobachtungskombination ausgewählt. Zurückzuführen ist das darauf, dass ausschließlich kategoriale Variablen verwendet werden und die Codierung nach D'Orazio (2019) in binäre Variablen umgewandelt werden, die wiederum numerisch angesehen werden und damit die Distanzen berechnet werden. Letztendlich wird die *Gower-Distanz* (Vergleich Gleichung 2.20) für die Datenfusion mit der *Distanz-Hot-Deck* Methode verwendet. Der Grund dafür ist, dass das Distanzmaß sowohl für qualitative als auch quantitative Variablen geeignet ist. Außerdem müssen die einzelnen Variablen nicht zusätzlich umcodiert werden.

4.2.1 Veränderung der Kategorienanzahl

Die ausgewählten Matching Variablen weisen eine unterschiedliche Anzahl an Kategorien auf. Besonders die Variable *ugeschwbegr* mit 20 Kategorien hat deutlich mehr Kategorien als die beiden anderen Variablen *utyp1*, mit sieben Kategorien, und *ustrklasse* mit fünf Kategorien. Daher trifft die Frage auf, wie die Datenfusionsergebnisse beeinflusst werden, wenn die Qualität der ursprünglichen Variablen verändert wird, indem die Anzahl der Kategorien verringert wird. Daher wird die Anzahl der Kategorien der Variable *ugeschwbegr* verringert, um zu überprüfen, ob sich bereits eine veränderte Variable auf die Ergebnisse auswirkt. Die spezifischen Variablen bleiben dabei gleich, jedoch umfasst die Geschwindigkeitsbegrenzung nach der Umwandlung einen größeren Bereich und wird demnach in Bezug auf die spezifischen Variablen ungenauer. Die 20 Kategorien der Variable *ugeschwebegr* werden nach dem Schema der Tabelle 4.4 in 5 Kategorien zusammengefasst. Diese zusammengefassten Kategorien werden in der neuen Variable *ugeschewbegr.neu* gespeichert.

Tabelle 4.4: Umwandlung der Variable ugeschwbegr in ugeschwbegr.neu durch die Veränderung der Kategorienanzahl

ugeschwbegr	ugeschwbegr.neu
005, 010, 015, 020, 025, 030, Z07, Z20, Z30	0-30
040, 050, 060	31-60
070, 080, 090, 100	61-100
110, 120, 130	101-130
keine Geschwindigkeitsbegrenzung	keine Geschwindigkeitsbegrenzung

Bevor die Datenfusion mit der veränderten Variable durchgeführt wird, werden die Vorbereitungsschritte der Datenfusion wiederholt, um zu überprüfen, wie sich die Veränderung auf die Verteilung und den Zusammenhang zu den spezifischen Variablen auswirkt. Aus diesem Grund wird für die neue Variable die Hellinger-Distanz H_D* (Vergleich Gleichung 2.36) berechnet, um die Verteilung der Variable *ugeschwbegr.neu* im Datensatz A mit der im Datensatz B zu vergleichen. In Tabelle 4.5 wird deutlich, dass sich die Verteilungen im Vergleich zu der originalen Variable ähnlicher sind und demnach die Veränderung der Variable zu einer Angleichung der Verteilung führt.

Tabelle 4.5: Vergleich der Hellinger-Distanz H_D* zwischen der veränderten Variable ugeschwebger.neu und der originalen ugeschwbegr[16]

Matching Variable	H_D*
ugeschwbegr.neu	0,0122
ugeschwbegr	0,0434

[16] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Des Weiteren ist zu erwarten, dass sich die Veränderung der Variable *ugeschwbegr* auf die Stärke des Zusammenhangs zu den spezifischen Variablen auswirkt. Aus diesem Grund wird für die Einschätzung der Veränderung sowohl die Berechnung des Zusammenhangs wiederholt durchgeführt, als auch die Unsicherheit im Vergleich zu den originalen Matching Variablen ermittelt. Die Stärke des Zusammenhangs wird für diese Einschätzung mit dem Unsicherheitskoeffizienten (Vergleich Gleichung 2.43) berechnet. In Tabelle 4.6 sind die Ergebnisse für die Variable *ugeschwbegr.neu*, im Vergleich zur ursprünglichen Variable *ugeschwbegr* aus Tabelle 4.2, aufgelistet. Im Gegensatz zur Verteilung verschlechtert sich die Stärke des Zusammenhangs sowohl für die spezifische Variable *uortslage* als auch für *uart*.

Tabelle 4.6: Vergleich des Unsicherheitskoeffizients U zwischen ugeschwbegr.neu und ugeschwbegr[17]

X	Y = uortslage	Z = uart
ugeschwbegr.neu	0,2206	0,0214
ugeschwbegr	0,2452	0,0280

Ebenfalls wird die Berechnung der Unsicherheit zusammen mit dem ersten und zweiten Strafterm durchgeführt. In Tabelle 4.7 wird deutlich, dass sich die Veränderung der Variable *ugeschwbegr.neu* nicht auf die Unsicherheit u der geschätzten Kontingenztabelle von Y x Z auswirkt. Ein Unterschied ist nur in Bezug auf die Spärlichkeit $\bar{n}$ ersichtlich, was darin begründet ist, dass die veränderte Variable nur noch ein Viertel der Kategorien aufweist.

Tabelle 4.7: Vergleich der Unsicherheit bei ugeschwbegr und ugeschwbegr.neu[17]

Kombination von X (D_m)	H_{D_m}	u	$u^{(1)}$	$u^{(2)}$	$\bar{n}$
utyp1 ustrklasse ugeschwbegr	700	0,03192656	0,03201159	0,03311704	1,28
utyp1 ustrklasse ugeschwbegr.neu	175	0,03339589	0,03348092	0,03392499	5,12

Insgesamt kann vor der Fusion zusammengefasst werden, dass eine Veränderung einer Variable hinsichtlich der Anzahl ihrer Kategorien den Zusammenhang zu anderen Variablen nicht unbedingt beeinflusst, jedoch sich die Verteilung zwischen den Datensätzen ändern kann. Trotz der geringen Veränderung bezüglich des Zusammenhangs wird das *Distanz-Hot-Deck* Verfahren zusätzlich mit der veränderten Matching Variable *ugeschwebegr.neu* durchgeführt, um den eventuellen Einfluss der Verteilungsgleichheit mit zu berücksichtigen. Dafür bleiben die Rahmenstruktur und die anderen beiden Matching Variablen gleich, es wird jeweils die beschränkte und unbeschränkte Methode (mit k=500 Durchläufen) durchgeführt.

[17] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

4.3 Random-Hot-Deck Verfahren

In den Grundlagen (Vergleich Kapitel 2.2.1) wurden zwei Varianten des *Random-Hot-Deck* Verfahrens beschrieben. Ohne die Verwendung von Spenderklassen wird die Annahme getroffen, dass die zu fusionierende Variable Z und die gemeinsamen Variablen **X** unabhängig voneinander sind. Dies kann jedoch durch die Analysen der Vorbereitung auf die Datenfusion (Vergleich Kapitel 4.1.2) ausgeschlossen werden.

Probedurchläufe der Datenfusion ergaben, dass die ermittelten Matching Variablen nicht als Variablen für die Bildung von Spenderklassen verwendet werden können, da nicht jede Spenderklasse im Spenderdatensatz eine Beobachtung enthält. Des Weiteren wurden alle möglichen Kombinationen der Matching Variablen analysiert mit dem Ergebnis, dass nur die alleinige Verwendung der Variable *utyp1* Spenderklassen bildet, die alle durch Beobachtungen gefüllt sind. Um zusätzlich eine weitere Qualitätsabweichung mit einzubeziehen, wird für die Bildung der Spenderklassen eine weitere Variable ausgewählt, die eigentlich keine optimale Voraussetzung für die Fusion aufweist, weil sie nicht als Matching Variable identifiziert wurde. Anhand der Ergebnisse für die Stärke des Zusammenhangs zu den spezifischen Variablen aus Tabelle 4.2, wird sich für die Variable *ucharunfst1* entschieden, weil sie zusätzlich die Voraussetzung erfüllt und zusammen mit der Variable *utyp1* Spenderklassen bildet, die in dem Spenderdatensatz befüllt sind.

Darüber hinaus wird trotzdem der Bezug zu den Matching Variablen beachtet, indem innerhalb jeder Spenderklasse nicht zufällig eine Beobachtung ausgewählt wird, sondern in Abhängigkeit der Matching Variablen, diejenige die am ähnlichsten ist. Somit wird bei diesem Verfahren das *Random-Hot-Deck* Verfahren mit dem *Distanz-Hot-Deck* Verfahren kombiniert.

Zusammenfassend werden für die Spenderklassen die Variablen *ucharunfst1* und *utyp1*, für die Matching Variablen die ermittelten Variablen *ustrklasse* und *ugeschwbegr* verwendet. Aus den gleichen Gründen wie beim *Distanz-Hot-Deck* Verfahren wird das *Gower-Distanzmaß* sowie eine Durchlaufanzahl von $k = 500$ verwendet.

4.4 Imprecise Imputation

Bezüglich der Fusionsverfahren unter der Annahme der CIA werden verschiedene Qualitätsstufen eingebaut. Die Datenfusion ohne Annahmen und unter Beibehalten der Unsicherheit wird ausschließlich mit den originalen simulierten Datensätzen durchgeführt. Durch die vorliegende Arbeit soll ein Überblick über den neuen Ansatz erhalten werden. Es wird getestet, wie die Ergebnisse zu interpretieren sind, wenn pro Einheit kein einzelner Wert für die fehlenden Beobachtungen fusioniert wird, sondern in Abhängigkeit des Spenderdatensatzes mehrere Werte. Wie in Kapitel 2.2.3 aufgeführt gibt es drei verschiedene Verfahren der *Imprecise Imputation*. In dieser Arbeit werden zwei davon angewendet und miteinander verglichen. Der allgemeine *Domain Ansatz* wird als Referenzfusion durchgeführt, um zu bewerten, wie sich das Fusionsergebnis ändert, wenn wie bei der *Domain Imputation* für jede Einheit alle Merkmalsträger fusioniert werden. Der Schwerpunkt liegt auf der *Variable-Wise Imputation*, die in Abhängigkeit der Spenderklassen die Merkmalsträger der Variable Z fusioniert.

Für die Spenderklassen werden die ermittelten Matching Variablen verwendet. In den simulierten Datensätzen ist jeweils eine spezifische Variable definiert. Damit erfüllen die Datensätze nicht die Anforderungen der *Case-Wise Imputation*. Aus diesem Grund wird in dieser Arbeit der Ansatz nicht verwendet.

4.5 Implementierung der verwendeten Methoden

Für die meisten in Kapitel 2.2.1 vorgestellten Methoden finden sich die Implementierungen im R-Paket StatMatch (D'Orazio, 2019). Für die Berechnung der beschränkten sowie der unbeschränkten *Distanz-Hot-Deck* Methode wird daher die im Paket enthaltene Funktion NND.hotdeck() verwendet. Ausgewählt wird als Zuordnungsalgorithmus die Ungarische Methode und als Distanzmaß die Gower-Distanz. Als Matching Variablen werden zum einen die in Kapitel 4.1.2 ausgewählten Variablen (*utyp1*, *ustrklasse* und *ugeschwbegr*) und zum anderen, für die Methode der veränderten Variable, die in Kapitel 4.2.1 beschriebene Variable *ugeschwbegr.neu* verwendet.

Für die *Random-Hot-Deck* Methode kann ebenfalls eine Funktion aus dem StatMatch-Paket genutzt werden. Die Funktion RANDwNND.hotdeck() wird zusammen mit der Gower-Distanz als Distanzmaß verwendet. Zusätzlich zu den Matching Variablen werden die Variablen für die Spenderklassen ausgewählt, die in Kapitel 4.3 bestimmt wurden. Durch die Verwendung der Funktionen NND.hotdeck() und RANDwNND.hotdeck() wird ausschließlich die Zuordnung der Beobachtungen des Empfängerdatensatzes mit den Beobachtungen aus dem Spenderdatensatz ermittelt. Aus diesem Grund wird zusätzlich die im StatMatch Paket enthaltene Funktion create.fused() verwendet, die abschließend den synthetischen Datensatz (Mikroansatz) erstellt.

Die Implementierung des zweiten Ansatzes, der *Imprecise Imputation*, findet sich im R-Paket impimp (Fink et al., 2019). Für die Fusion und Erstellung eines synthetischen Datensatzes wird die Funktion impimp() verwendet. Speziell für das Verfahren der *Variable-Wise Imputation* wird die Methode „variable_wise“ ausgewählt und für den *Domain Ansatz* die Methode „domain“. Zusätzlich werden ebenfalls die Matching Variablen aus Kapitel 4.1.2 verwendet.

5 Ergebnisse

In diesem Kapitel werden die Ergebnisse der in Kapitel 4 beschriebenen Datenfusionen aufgeführt und diskutiert. Für die nichtparametrischen Methoden unter der *CIA* werden die Ergebnisse in Abhängigkeit der Validitätsebenen nach Rässler ausgewertet. Für die *Imprecise Imputation* findet die Auswertung in Abhängigkeit der Intervallbreiten der Wahrscheinlichkeitstabelle für die jeweiligen Verteilungen statt. Abschließend werden die Ergebnisse der beiden Ansätze miteinander verglichen und diskutiert.

5.1 Kennzahlen der ersten Validitätsebene

In der Literatur wird die erste Ebene kritisch angesehen, da das Erreichen der Ebene nicht nur von der Qualität der Fusion abhängt, sondern auch von der Anzahl an Merkmalsausprägungen. Außerdem ist der Erhalt der exakten Werte nur dann gegeben, wenn der gesamte Beobachtungsvektor der fusionierten Variable pro Einheit exakt wieder gegeben wird (Rässler, 2002, vgl. S.30). In Bezug auf diese Arbeit wird eine spezifische Variable *uart* mit zehn Kategorien fusioniert, demnach muss nur ein Wert reproduziert werden und die Kategorienanzahl ist nicht zu groß. Es ist sinnvoll, die erste Ebene mit in die Auswertung einzubeziehen, um einen ersten Eindruck über die Qualität der Datenfusion zu bekommen und abschätzen zu können, ob Unterschiede zwischen den Verfahren auftreten. Die nachfolgenden Tabellen in Kapitel 5.1 bis 5.4 stellen die Verfahren wie folgt gegenüber:

- Distanz: *Distanz-Hot-Deck* Verfahren mit originaler Variable *ugeschwbegr*
- Distanz*: *Distanz-Hot-Deck* Verfahren mit veränderter Variable *ugeschwbegr.neu*
- Random: *Random-Hot-Deck* Verfahren mit Variable *ucharunfst1* für die Spenderklassen.

Für das Verfahren Distanz und Distanz* werden jeweils die Methoden beschränkt und unbeschränkt dargestellt.

Zum Ermitteln der Güte der ersten Validitätsebene wird getestet, wie gut die Werte der spezifischen Variable *uart* im fusionierten Datensatz reproduziert wurden. Dafür wird die *Hit Rate* berechnet, die den prozentualen Anteil der Übereinstimmungen zwischen den Werten der fusionierten Variable $\widetilde{uart}$ und der originalen Variable *uart* im Empfängerdatensatz bestimmt. Der Tabelle 5.1 ist zu entnehmen, dass für die verschiedenen Verfahren die Werte der *Hit Rate* eng beieinander liegen. Demnach werden für alle Verfahren durchschnittlich ca. 50% der Werte durch die Fusion exakt reproduziert.

Tabelle 5.1: Durchschnittliche HitRate und Spannweite in Prozent (gerundete Werte)[18]

Verfahren	Methode	HitRate [%]	Minimum [%]	Maximum [%]
Distanz	beschränkt (b1)	48	-	-
	unbeschränkt (u1)	50	47	52
Distanz*	beschränkt (b2)	47	-	-
	unbeschränkt (u2)	50	47	52
Random	(r)	51	48	53

Jedoch ist beim detaillierten Vergleich der Methoden die Tendenz zu erkennen, dass bei beiden Distanzverfahren die unbeschränkte Variante jeweils bessere Ergebnisse erzielt als die beschränkte. Wird des Weiteren die Streuung der unbeschränkten Methoden in die Analyse mit einbezogen, lässt sich ein weiterer Vorteil dieser Methode gegenüber der beschränkten Methode erkennen. Dies ist auch der Abbildung 5.1 zu entnehmen, in der die Verteilungen der *Hit Rate* in Form von Box-Plots dargestellt sind. Es wird ersichtlich, dass selbst die schlechteste Hit Rate (Minimum) der unbeschränkten Methode immer noch vergleichbar gute Ergebnisse wie die beschränkte Methode liefern. Eine mögliche Begründung dafür ist die im Durchschnitt nähere Distanz zwischen den Empfänger- und Spenderbeobachtungen.

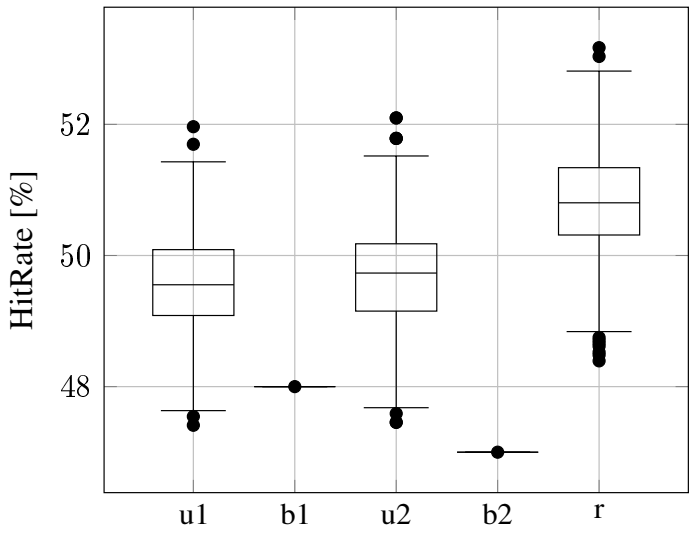


Abbildung 5.1: Streuung der Hit Rate[18]

Die Veränderung der Variable *ugeschwbegr* führt zu keiner sichtbaren Veränderung der *Hit Rate*, wie anhand der ähnlich verteilten Boxplots erkennbar ist. In Tabelle 5.1 sowie Abbildung 5.1 ist ersichtlich, dass im Vergleich zum *Distanz-Hot-Deck* Verfahren die *Random-Hot-Deck* Methode mit durchschnittlich 51% die besten Ergebnisse erzielt.

[18] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Das lässt darauf schließen, dass für die erste Ebene zusätzliche Spenderklassen eine höhere Reproduktion der fusionierten Werte erreichen, obwohl die Variablen für die Spenderklassen nicht den stärksten Zusammenhang aufweisen.

5.2 Zweite Validitätsebene

Die zweite Validitätsebene ist erreicht, wenn die gemeinsame Verteilung im fusionierten Datensatz der wahren unbeobachteten Verteilung entspricht. Die Überprüfung dieser Ebene kann nur in künstlich erzeugten Simulationsdatensätzen durchgeführt werden, in denen die wahre gemeinsame Verteilung definiert ist. In dieser Arbeit ist die wahre unbeobachtete Verteilung nicht bekannt, allerdings kann in dem besonderen Fall der simulierten Datensätze geprüft werden, inwieweit die gemeinsame Verteilung des originalen Datensatzes (Empfängerdatensatz) mit der des fusionierten Datensatzes übereinstimmt. Für die Überprüfung wird die Hellinger-Distanz (Vergleich Gleichung 2.52) verwendet. Im Vergleich zu den anderen vorgestellten Bewertungsmethoden in Kapitel 2.4.1 ist es dadurch möglich, die gesamte Verteilung $f_{\mathbf{X}YZ}$ miteinander zu vergleichen und nicht nur die der spezifischen Variablen f_{YZ}. Es wird jeweils die Distanz zwischen der Verteilung der Matching Variablen mit spezifischen Variablen ($\mathbf{X}_\mathrm{M}$, Y, Z) im Empfängerdatensatz und der Verteilung von ($\mathbf{X}_\mathrm{M}$, Y, $\tilde{Z}$) im fusionierten Datensatz berechnet. Die Berechnung erfolgt mittels der comp.prop() Funktion im R-Paket StatMatch (D'Orazio, 2019).

In Tabelle 5.2 sind die Ergebnisse dargestellt. Bei den Methoden, in denen $k = 500$ Durchläufe durchgeführt wurden, handelt es sich um Mittelwerte, die zusätzlich um die Spannweiten mit kleinstem und größtem Wert der Hellinger-Distanz ergänzt wurden. Es wird deutlich, dass alle Methoden den Grenzwert von $H_D = 0,05$ mit mindestens 30% deutlich überschreiten und damit nicht als ähnlich angesehen werden können. Somit sind nicht die gleichen statistischen Analysen möglich wie mit dem Empfängerdatensatz. Die schlechte Nachbildung ist darauf zurückzuführen, dass die CIA für die simulierten Datensätze nicht gilt und somit die Annahme fälschlicherweise getroffen wurde. Somit sind die spezifischen Variablen Y (*uortslage*) und Z (*uart*), gegeben **X** nicht voneinander unabhängig.

Tabelle 5.2: Hellinger Distanz zum Vergleich der gemeinsamen Verteilung $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{X}YZ}$[19]

Verfahren	Methode	H_D [-]	Minimum H_D [-]	Maximum H_D [-]
Distanz	beschränkt (b1)	0,36125	-	-
	unbeschränkt (u1)	0,35926	0,34010	0,37679
Distanz*	beschränkt (b2)	0,33415	-	-
	unbeschränkt (u2)	0,331247	0,31462	0,34995
Random	(r)	0,358553	0,33830	0,37805

[19] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Werden die Methoden und Varianten dennoch miteinander verglichen ohne die Einhaltung des Grenzwertes mit einzubeziehen, sind Tendenzen zu erkennen, die zu einer Veränderung der Nachbildung der gemeinsamen Verteilung führen. Die *Random Methode* hat die beste Trefferquote in der ersten Validitätsebene erreicht, weist allerdings mit einer durchschnittlichen Hellinger-Distanz von 0,359 in der zweiten Ebene nicht die besten Ergebnisse auf. Es bestätigt sich so die Aussage von Kiesl und Rässler (2005, vgl. S.25f), dass eine höhere Trefferquote nicht gleichzusetzen ist mit einem guten Erhalt der gemeinsamen Verteilung.

Werden die Werte der Hellinger-Distanz der beschränkten Methode mit den Mittelwerten der unbeschränkten Methoden verglichen, ist kein signifikanter Unterschied erkennbar. Bei beiden Methoden liegt die Abweichung bei unter 1%. Jedoch wird durch die Hinzunahme der Spannweite der Werte der Hellinger-Distanz wird ersichtlich, dass die unbeschränkte Methode sowohl schlechtere als auch bessere Ergebnisse erzielen kann. Wird dabei im Detail die Streuung der Werte mit einbezogen, siehe Abbildung 5.2, wird ersichtlich, dass die Streuung dabei symmetrisch ist und es demnach gleich wahrscheinlich ist, ob die Werte besser oder schlechter werden.

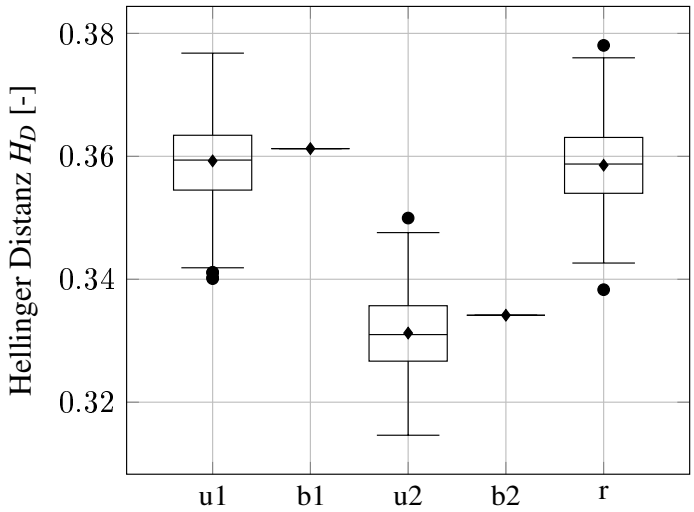


Abbildung 5.2: Streuung der Hellinger Distanz zum Vergleich der gemeinsamen Verteilung f_{XYZ} und $\tilde{f}_{XYZ}$[20]

Ein allgemeiner Unterschied ist jedoch zwischen dem *Distanz-Hot-Deck* Verfahren mit der originalen Variable und der veränderten Variable *ugeschwbegr.neu* zu erkennen. Die Hellinger-Distanzen weisen sowohl für die beschränkte als auch für die unbeschränkte Methode kleinere Werte auf, wenn *ugeschwbegr.neu* für die Fusion verwendet wird. Demnach bestätigt sich, dass eine ähnlichere Verteilung zwischen den Matching Variablen im Empfänger- und Datensatz das Fusionsergebnis hinsichtlich der gemeinsamen Verteilung positiv beeinflusst.

[20]FDZ der Statistischen Ämter des Bundes und der Länder, 2013

5.3 Dritte Validitätsebene

Die dritte Validitätsebene bezieht sich auf den Erhalt der Zusammenhangsstruktur zwischen den spezifischen Variablen. Die in Kapitel 2.4.1 vorgestellten Methoden zur Überprüfung dieser Ebene sind bis auf eine Methode vor allem für quantitative Variablen ausgelegt. Aus diesem Grund wird sich für das Verfahren in Endres (Endres, 2019, zum Beispiel S.55) entschieden und der korrigierte Kontingenzkoeffizient C^* (Vergleich Gleichung 2.53) verwendet. Für den Vergleich wird zu einem C^* zwischen den spezifischen Variablen Y und Z im originalen Datensatz (Empfängerdatensatz) berechnet und zum anderen im fusionierten Datensatz zwischen Y und $\tilde{Z}$. Die Bewertungsgrundlage bildet anschließend die Differenz zwischen den korrigierten Kontingenzkoeffizienten. Die Ergebnisse sind der Tabelle 5.3 zu entnehmen.

Tabelle 5.3: Vergleich der Zusammenhangsstruktur zwischen Y und Z im fusionierten und originalen Datensatz anhand des korrigierten Kontingenzkoeffizienten C^*[21]

Verfahren	Methode	C^* Originaldatensatz [-]	$\tilde{C}^*$ fusionierte Datensatz [-]	Differenz [-]
Distanz	beschränkt (b1)	0,57720	0,51570	0,06150
	unbeschränkt (u1)	0,57720	0,52797	0,04923
Distanz*	beschränkt (b2)	0,57720	0,52028	0,05692
	unbeschränkt (u2)	0,57720	0,53051	0,04669
Random	(r)	0,57720	0,52468	0,05252

Bei allen Methoden wird die Stärke des Zusammenhangs zwischen den Variablen durch die Datenfusion abgeschwächt. In Bezug auf den gemessenen C^* im Originaldatensatz, weicht mit 11% der $\tilde{C}^*$ durch die beschränkte *Distanz-Hot-Deck* Methode mit der originalen Variable am meisten ab. Die besten Ergebnisse erzielen im Durchschnitt die unbeschränkten Methoden, die durch die Fusion eine Abweichung von 8 und 9% erzeugen. Die Abweichung durch die Random Methode liegt zwischen denen der beiden zuvor genannten Methoden. In Bezug auf die Veränderung der Variable *ugeschwbegr* ist eine Verbesserung durch die Verwendung der Variable *ugeschwbegr.neu* zu erkennen. Dies lässt den Rückschluss zu, dass sich die Verteilungsgleichheit der Matching Variablen positiv auf das Erreichen der dritten Ebene auswirkt.

In der Literatur gibt es keine allgemeine Definition, ab welcher Abweichung nicht mehr von einer ähnlichen Zusammenhangsstruktur ausgegangen werden kann. Wird die Auswertung von Endres (2019, vgl. S.70f) hinzugezogen, sind alle Zusammenhangstrukturen im fusionierten Datensatz reproduziert. In Barry (1988, vgl. S.281) wird ein Korrelationskoeffizient für die Messung des Zusammenhangs berechnet. Dabei weichen die Werte bis zu 33% vom originalen Koeffizienten ab und er definiert die Zusammenhangsstruktur als nicht reproduziert. Jedoch bezieht er zusätzlich das Kreuzproduktverhältnis mit ein, bei denen die Abweichungen bei ungefähr 80% liegen.

[21] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Insgesamt scheinen daher die Zusammenhänge zwischen den spezifischen Variablen für alle Methoden gut erhalten zu sein und Analysen bezüglich des Zusammenhangs zwischen den spezifischen Variablen führen mit den fusionierten Datensätzen zu gleichen Ergebnissen.

5.4 Vierte Validitätsebene

Die Überprüfung der vierten Validitätsebene ist die einzige, die mit realen Datensätzen durchzuführen ist, da für die Überprüfung keine zusätzlichen Informationen notwendig sind. Die Güte der Ebene ist erreicht, wenn f_Z und f_{ZX} aus dem Spenderdatensatz reproduziert werden können. Für die Überprüfung der vierten Validitätsebene wird ebenfalls die Hellinger-Distanz (Vergleich Gleichung 2.36) verwendet, weil diese bereits in mehreren Studien für die Überprüfung dieser Ebene verwendet wurde (zum Beispiel Webber und Tonkin, 2013, S. 21f). Die Berechnung erfolgt mittels der comp.prop() Funktion im R-Paket StatMatch (D'Orazio, 2019).

Für die Überprüfung des Erhalts von f_Z wird jeweils die Hellinger Distanz H_D zwischen der Verteilung der Variable Z im Spenderdatensatz und der Variable $\tilde{Z}$ im fusionierten Datensatz berechnet. Die Ergebnisse sind der Tabelle 5.4 zu entnehmen.

Tabelle 5.4: Hellinger-Distanz zum Vergleich von f_Z und $\tilde{f}_Z$[22]

Verfahren	Methode	$H_D(f_Z, \tilde{f}_Z)$ [-]	Minimum H_D [-]	Maximum H_D[-]
Distanz	beschränkt (b1)	0,0230	-	-
	unbeschränkt (u1)	0,0252	0,0123	0,0460
Distanz*	beschränkt (b2)	0,0211	-	-
	unbeschränkt (u2)	0,0253	0,0139	0,0408
Random	(r)	0,0251	0,0104	0,0414

Aus den aufgeführten Ergebnissen wird ersichtlich, dass die Randverteilung der Variable Z im fusionierten Datensatz für alle Methoden und Varianten nachgebildet werden konnte. Auch unter Hinzunahme der maximalen Distanzen liegt kein Wert über dem Grenzwert von $H_D = 0,05$.

Werden die Verfahren miteinander verglichen, in denen die Fusion mehrmals durchgeführt wurde, ist kein Unterschied zwischen den durchschnittlichen Werten der Hellinger-Distanzen zu erkennen. Dies wird auch deutlich unter Hinzunahme der Streuung der Werte (Vergleich Abbildung 5.3).

[22] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

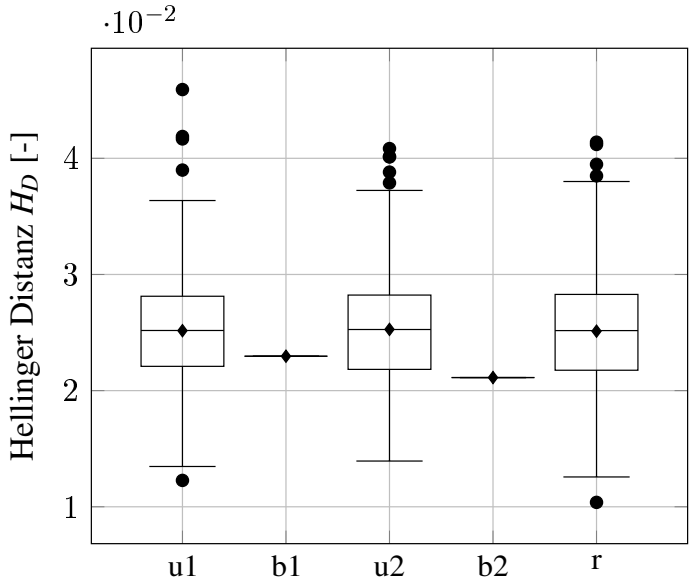


Abbildung 5.3: Streuung der Hellinger-Distanz zum Vergleich von f_Z und $\tilde{f}_Z$[23]

Demnach beeinflusst weder die veränderte Variable *ugeschwbegr.neu*, noch die zusätzliche Variable *ucharunfst1* die Reproduktion der Randverteilung. Werden im Vergleich dazu die Methoden der *Distanz-Hot-Deck* Verfahren miteinander verglichen, wird die Randverteilung im Durchschnitt durch die beschränkte Methode besser reproduziert. Jedoch kann kein signifikanter Unterschied zwischen den beschränkten und unbeschränkten Methoden festgestellt werden, da die Abweichungen nur bei 0,0022 (Distanz) sowie 0,0042 (Distanz*) liegen. Wird die Streuung der unbeschränkten Methoden mit einbezogen (Vergleich Abbildung 5.3), können bei diesen Methoden die Ergebnisse sowohl besser (Wert der Hellinger-Distanz wird kleiner) als auch schlechter (Hellinger-Distanz wird größer) als der Durchschnittswert werden. Demnach beeinflusst die Streuung der Hellinger-Distanz der unbeschränkten Methode zur durchschnittlichen Hellinger-Distanz der beschränkten Methode die Abweichung. Die in Kapitel 4.2 aufgestellte These, dass aufgrund des großen Unterschiedes zwischen den Datensätzen die Ergebnisse der Randverteilung der zu fusionierenden Variable sich unter der Verwendung der beschränkten und unbeschränkten Methode nicht deutlich voneinander unterscheiden, kann somit nicht eindeutig bestätigt werden.

Zusätzlich wird in der vierten Ebene für die Überprüfung des Erhalts von f_{ZX} jeweils die Hellinger Distanz H_D zwischen f_{ZX_p} mit $p = 1, 2, 3$ für alle drei Matching Variablen im Spenderdatensatz zu $\tilde{f}_{ZX_p}$ im fusionierten Datensatz berechnet. In Tabelle 5.5 sind die durchschnittlichen Werte der Hellinger-Distanz für die verschiedenen Matching Variablen dargestellt.

[23] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Tabelle 5.5: Hellinger-Distanz zum Vergleich der Randverteilung von der fusionierten Variable (Z = uart) und der jeweiligen Matching Variable (X_1 = utyp, X_2 = ustrklasse, X_3 = ugeschwbegr/ugeschwbegr.neu)[24]

Verfahren	Methode	ZX_1	ZX_2	ZX_3
Distanz	beschränkt (b1)	0,0559	0,0646	0,0930
	unbeschränkt (u1)	0,0636	0,0642	0,1049
Distanz*	beschränkt (b2)	0,0554	0,0653	**0,0446**
	unbeschränkt (u2)	0,0637	0,0647	0,0578
Random	(r)	0,0743	0,0743	0,1086

Werden die Ergebnisse mit dem Grenzwert der Hellinger-Distanz von 0,05 verglichen, wird ersichtlich, dass nur unter dem beschränkten *Distanz-Hot-Deck* Ansatz mit *ugeschwbegr.neu* (Distanz*) die Verteilungen als gleich angesehen werden können. Jedoch ist der Tabelle ebenfalls zu entnehmen, dass die meisten anderen Werte der Hellinger-Distanz in Bezug auf die Matching Variablen *utyp1* und *ustrklasse* nur geringfügig höhere Werte als 0,05 aufweisen. In Anbetracht, dass der Grenzwert von $H_D = 0,05$ kein statistisch festgelegter Wert ist, sondern teilweise in der Literatur als Faustregel verwendet wird, kann bei Werten wie zum Beispiel 0,0559 nicht eindeutig die Aussage getroffen werden, dass anschließend die Analysen von (X, Z) und (X, $\tilde{Z}$) nicht doch zu den selben Schlüssen führen können.

Allgemein wird ersichtlich, dass die Werte der Hellinger-Distanzen unter der *Random-Hot-Deck* Datenfusion am größten sind und somit die gemeinsame Verteilung der fusionierten Variable zu den Matching Variablen am schlechtesten nachgebildet werden konnte. Dies war auch so zu erwarten, da die Distanzen bezüglich der Matching Variablen innerhalb der Spenderklassen berechnet wurden und somit bereits eine Beeinflussung der Variable *ucharunfst1* stattgefunden hat. Demnach standen nicht immer die Beobachtungen mit der kleinsten Distanz zur Verfügung. Dies hat sich negativ auf die Nachbildung der gemeinsamen Verteilung der spezifischen Variablen zu den Matching Variablen ausgewirkt.

Des Weiteren werden die Verfahren mit der originalen Variable *ugeschwbegr* in Abhängigkeit der Matching Variablen verglichen. Dabei unterscheiden sich die Werte der Hellinger Distanz der ersten beiden Matching Variablen (*utyp1* und *ustrklasse*) von der dritten (*ugeschewbegr*). Die gemeinsame Verteilung der ersten beiden Matching Variablen lassen sich besser im fusionierten Datensatz nachbilden als die der dritten Variable. Werden die Ergebnisse aus der Vorbereitung (Kapitel 4.1) für die Datenfusion mit einbezogen, können Zusammenhänge abgeleitet werden. Die Verteilungen der Variablen *utyp1* und *ustrklasse* sind in den beiden Datensätzen A und B ähnlicher, als die der Variable *ugeschwbegr* (Vergleich Tabelle 4.1). Außerdem weist auch die Variable *ugeschwbegr* den schlechtesten Zusammenhang zu der spezifischen Variable Z auf (Vergleich Tabelle 4.2). Demnach bestätigt sich, dass die Eigenschaften der Matching Variablen einen Einfluss auf das Fusionsergebnis haben.

[24] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Werden die Verfahren des *Distanz-Hot-Deck* Verfahrens miteinander verglichen, ist eine eindeutige Verbesserung des Fusionsergebnis durch die veränderte Variable *ugeschwbegr* zu erkennen. Beispielsweise kann für die beschränkte Fusion nach Änderung der Kategorienanzahl die gemeinsame Verteilung im fusionierten Datensatz so gut nachgebildet werden, dass die Verteilungen als gleich angesehen werden, da $H_D = 0,0446$ kleiner als der Grenzwert von 0,05 ist. Demnach bestätigt sich, dass das Fusionsergebnis positiv beeinflusst wird, wenn die Verteilung zwischen den Variablen im Empfänger- und Spenderdatensatz gleich sind. Bestärkt wird das durch die Hinzunahme der Ergebnisse der zweiten Matching Variable *ustrklasse*. Der Zusammenhang zu den spezifischen Variablen ist bei der Variable *ustrklasse* noch immer stärker, allerdings ist durch die Veränderung der Variable *ugeschwbegr* die Ähnlichkeit der Verteilung besser geworden, demnach wurde f_{ZX_3} besser nachgebildet. Die Werte der Hellinger-Distanz bezüglich der ersten und zweiten Matching Variable verändern sich nach der Veränderung der Variable *ugeschwbegr* nicht.

Werden wiederum die Methoden des *Distanz-Hot-Deck* Verfahrens aus Tabelle 5.5 miteinander verglichen, fällt auf, dass bei der zweiten Matching Variable *ustrklasse* die durchschnittlichen Werte der Hellinger-Distanz bei der beschränkten Fusion mit der unbeschränkten übereinstimmen und beide Verfahren zu einem ähnlichen Ergebnis führen. Bei der ersten Variable *utyp1* sowie der dritten Variable *ugeschwbegr* beziehungsweise *ugeschwbegr.neu* wird allerdings durch die beschränkte Datenfusion ein besseres Ergebnis erzielt. Vor allem bei der dritten Matching Variable wird dies deutlich, wenn zusätzlich die Verteilung der Werte der Hellinger-Distanz mit einbezogen werden. Der Abbildung 5.4 ist zu entnehmen, dass bei der unbeschränkten Methode auch die besten Werte der Hellinger-Distanz (Minimum) immer noch vergleichbare Ergebnisse wie die beschränkte Methode liefern. Die Box-Plots mit den Verteilungen der Hellinger-Distanzen für die anderen Variablen weisen keine Besonderheiten auf und sind daher im Anhang A.9 dargestellt.

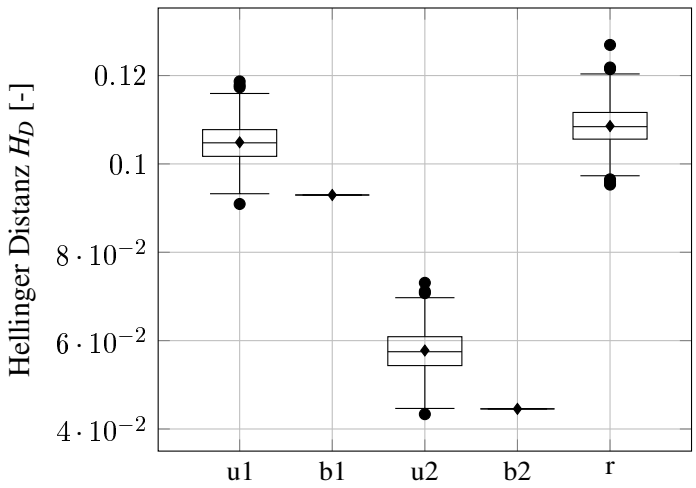


Abbildung 5.4: Streuung der Hellinger-Distanz zum Vergleich von f_{ZX_3} und $\tilde{f}_{ZX_3}$[25]

[25] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

5.5 Qualität Imprecise Imputation

Für die Qualität der *Imprecise Imputation* werden wie bei den Validitätsebenen nach Rässler die Randverteilungen sowie die gemeinsame Verteilung im fusionierten Datensatz mit dem originalen Datensatz (Empfängerdatensatz) verglichen. Jedoch können aufgrund der intervallgeschätzten Wahrscheinlichkeiten nicht die gleichen Berechnungen durchgeführt werden. Für die Qualitätseinschätzung der fusionierten Datensätze durch die *Imprecise Imputation* werden aus diesem Grund zwei Bewertungsschritte verwendet, die in Endres (2019, vgl. S.123ff) vorgeschlagen werden.

Im **ersten Bewertungsschritt** wird überprüft, ob die Parameter der Verteilung aus dem originalen Datensatz (Empfängerdatensatz) innerhalb der oberen und unteren Grenzen der geschätzten Parameter aus dem fusionierten Datensatz liegen. Daran kann ermittelt werden, wie gut die wahren Werte getroffen wurden. Als **zweiter Bewertungsschritt** wird die durchschnittliche Breite der Intervalle berechnet, die damit den mittleren Abstand zwischen der oberen und unteren Grenze angibt. Eine Intervallbreite von 0 bedeutet eine genaue Schätzung. Wird eine asymmetrische Datenfusion durchgeführt, und demnach nur die spezifische Variable in den Empfängerdatensatz fusioniert, kann die Breite maximal 1 betragen. Wird allerdings eine symmetrische Datenfusion $A \cup B$ durchgeführt, verändert sich die Breite in Abhängigkeit des Verhältnisses zwischen Datensatz A und B. Bei $n_A = n_B$ liegt der maximale Wert bei 0,5. (Endres, 2019, vgl. S.123)
Letztendlich müssen die beiden Vergleiche wiederum miteinander in Beziehung gesetzt werden, da eine durchschnittliche kleine Intervallbreite auf eine genaue Schätzung hinweist, jedoch bei einer kleinen Anzahl von getroffen Intervallbreiten wiederum auf eine falsche „genaue“ Schätzung.

Geprüft werden die beiden Schritte sowohl für die Randverteilung der fusionierten Variable (Z) sowie für die Verteilung zwischen der fusionierten Variable mit den Matching Variablen (Z, X_p) und der gemeinsamen Verteilung von ($\mathbf{X}$, Y, Z). Für die Berechnung der oberen und unteren Grenze wird die Funktion impest() aus dem R-Package impimp (Fink et al., 2019) verwendet. Mit der Funktion kann für ein Ereignis die untere und obere Wahrscheinlichkeit ausgerechnet werden, dass dieses Ereignis eintritt. Für die Bewertung des fusionierten Datensatz bedeutet das, dass in Abhängigkeit der Verteilung für jede Zelle der Wahrscheinlichkeitstabelle die obere und untere Grenze der relativen Häufigkeit berechnet wird.

Erster Bewertungsschritt: Übereinstimmung der Parameter

Im ersten Schritt wird die prozentuale Häufigkeit berechnet, ob die wahren Parameter der Verteilungen aus dem Empfängerdatensatz innerhalb der Intervalle der Verteilungen aus dem fusionierten Datensatz liegen. Für die Randverteilung der fusionierten Variable (Z) sowie der Verteilung zwischen der Matching Variablen mit der fusionierten Variable (Z, X_p) sind die Ergebnisse in den ersten vier Spalten der Tabelle 5.6 sowohl für die *Variable-Wise* als auch die *Domain Methode* dargestellt. Die letzte Spalte beinhaltet die Ergebnisse für die gemeinsame Verteilung von ($\mathbf{X}_M$, Y, Z). Wie zu erwarten war, sind bei der *Domain Imputation* alle Intervalle getroffen worden, da für jede fehlende Beobachtung im Empfängerdatensatz ein Satz aus allen zehn Kategorien der spezifischen Variable *uart* fusioniert wurde.

Tabelle 5.6: Anteil in %, wie häufig die Parameter der originalen Verteilung (Empfängerdatensatz) im geschätzten Intervall der fusionierten Datei liegen[26]

Methode	Z	ZX_1	ZX_2	ZX_3	$ZY\mathbf{X}_M$
variable wise	100	100	100	99	99,57857
domain	100	100	100	100	100

Unter der Verwendung der *Variable-Wise Imputation* konnte bei der Verteilung der Variable Z mit der dritten Matching Variable *ugeschwbegr* keine 100%ige Übereinstimmung erreicht werden, genauso auch wie bei der gemeinsamen Verteilung. Beim Vergleich zwischen den beiden Verteilungen ist für die gemeinsame Verteilung die Trefferquote wiederum höher. Begründen lässt sich das unter anderem mit der Spärlichkeit der Wahrscheinlichkeitstabelle der gemeinsamen Verteilung. Die gemeinsame Verteilung erzeugt mehr Zellen (14000) als es Unfälle (2240) im Empfängerdatensatz gibt. Demnach hatten die meisten Einträge in der Tabelle den Wert 0, was zur einer Erhöhung der Trefferquote führte. Der Einfluss der geringeren Ähnlichkeit zwischen der Variable *ugeschwbegr* in beiden Datensätzen kann eine mögliche Begründung dafür sein, dass im Vergleich zu den anderen Randverteilungen (Z, X_1) und (Z, X_2) eine Übereinstimmung von 100% nicht erreicht wurde.

Zweiter Bewertungsschritt: Durchschnittliche Intervallbreite

In Tabelle 5.7 sind für die *Domain Imputation* und *Variable-Wise Imputation* die durchschnittliche Breite sowie zusätzlich der Median[27] für die unterschiedlichen Verteilungen aufgelistet. Bis auf die Verteilung der fusionierten Variable Z mit der zweiten Matching Variable *ustrklasse* ergeben die *Variable-Wise Imputation* kleinere Intervalle als bei der *Domain Imputation*.

Tabelle 5.7: Arithmetischer Mittelwert und Median der Intervallbreiten im fusionierten Datensatz für Variable-Wise und Domain Imputation[26]

		Z	ZX_1	ZX_2	ZX_3	$ZY\mathbf{X}_M$
domain	Mittelwert	1	0,2	0,005	0,14286	0,00071
	Median	1	0,18214	0,00290	0,15536	0
		Z	ZX_1	ZX_2	ZX_3	$ZY\mathbf{X}_M$
variable-wise	Mittelwert	0,67384	0,13477	0,03370	0,09626	0,00048
	Median	0,75357	0,10558	0,00090	0,07611	0

[26] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

[27] Aufgrund einer Geheimhaltungsvereinbarung mit dem Forschungsdatenzentrum Dresden wurden keine Box-Plots erstellt, die die gesamte Verteilung der Breiten für die einzelnen Verteilungen darstellen.

Werden die Intervallbreiten der Verteilungen zwischen der Variable Z und den Matching Variablen miteinander vergleichen, erzielt sowohl unter der *Variable-Wise Imputation* als auch unter der *Domain Imputation* die zweite Matching Variable *ustrklasse* die kleinsten Intervalle und die der ersten Variable *utyp1* die größten Intervalle. Daraus lässt sich demnach nicht ableiten, dass ein besserer Zusammenhang zwischen den Matching Variablen und spezifischen Variablen zu einem kleineren Intervall und damit genaueren Schätzung führt. Auch spiegelt sich die Ähnlichkeit der Matching Variablen zwischen den Datensätzen A und B nicht in den Intervallbreiten wieder. Begründen lässt sich das womöglich damit, das insgesamt alle Matching Variablen einen stärkeren Zusammenhang aufweisen und die Unterschiede zueinander nicht groß genug sind. Dies lässt den Schluss zu, dass die Ursache woanders begründet liegt, was es in weiterführender Forschung zu untersuchen gilt. Auch spiegelt sich die Ähnlichkeit der Matching Variablen zwischen den Datensätzen A und B nicht in den Intervallbreiten wieder.

Die Intervallbreiten der gemeinsamen Verteilung deuten auf eine genaue Schätzung der Intervalle hin, da diese unter beiden Fusionsansätzen Mittelwerte nahe dem Wert 0 aufweisen. Dies wird ebenfalls durch den Median bestätigt, die Hälfte der Intervallbreiten werden genau geschätzt. In diesem Kontext ist aber auch der Hinweis aus Schritt 1 mit zu beachten, dass die erzeugte Kontingenztabelle größer als die Datensatzgröße ist und demnach die meisten Einträge in der Tabelle keine Werte aufweisen und dadurch die Schätzung positiv beeinflusst wird.

5.6 Vergleich und Auswertung der Ergebnisse

5.6.1 Fazit der Ergebnisse der fusionierten Datensätze unter der Annahme von CIA

In der zweiten Validitätsebene weichen die Werte der Hellinger-Distanz mindestens 30% von dem Grenzwert von 0,05 ab. Somit kann die Datenfusion für keines der Verfahren als erfolgreich angesehen werden, in dem Sinne, dass alle statistischen Tests auf den fusionierten Datensatz angewendet werden können.

Werden allerdings Abstriche gemacht und das Ziel sind nicht nur multivariate Analysen (Validitätsebene 2), bei denen alle Variablen in Verbindung gesetzt werden, sondern bivariate Analysen, bei denen die Korrelation zwischen zwei Variablen untersucht wird, so können durch das Erfüllen der dritten Ebene die Fusionen als Erfolgreich betrachtet werden. Der Vergleich der Zusammenhänge zwischen den spezifischen Variablen *uortslage* und *uart* in den fusionierten Datensätzen gegenüber denen im originalen Datensatz liefern vielversprechende Ergebnisse. Damit bestätigen sich die in Grömping et al. (2007, vgl. S.141) erzielten Ergebnisse, dass durch eine Datenfusion von Unfalldatenbanken bivariate Beziehungen erhalten bleiben können. Aus diesem Grund ist es sinnvoll, die fusionierten Datensätze hinsichtlich des Zusammenhangs zwischen den spezifischen Variablen in weiterführender Forschung zu untersuchen.

Da in realen Datensätzen ausschließlich die vierte Validitätsebene geprüft werden kann, werden die ermittelten Ergebnisse hinsichtlich der Aussagekraft dieser Validitätsebene überprüft:

Ebene 1: Die Ergebnisse der vierten Validitätsebenen geben keinen Hinweis hinsichtlich der Erfüllung der ersten Ebene. Des Weiteren bestätigt sich im Allgemeinen, dass die erste Stufe der Validitätsebene keinen Hinweis auf den Erhalt der gemeinsamen Verteilung gibt.

Ebene 2: In Bezug auf den in der Literatur als Faustregel verwendeten Grenzwert für die Hellinger-Distanz von 0,05, kann die vierte Validitätsebene mindestens für das unbeschränkte *Distanz-Hot-Deck* Verfahren mit der veränderten Variable *ugeschwbegr.neu* (Distanz*) als erfüllt angesehen werden. Da dies im Widerspruch zu den Ergebnissen der zweiten Validitätsebene steht, zeigt sich, dass das Erfüllen der vierten Ebene keinen eindeutigen Hinweis über das Zutreffen der zweiten Validitätsebene liefert.

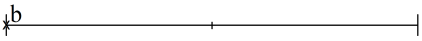
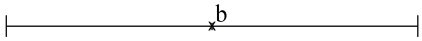
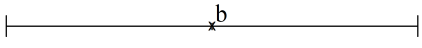
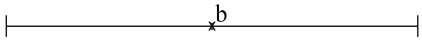
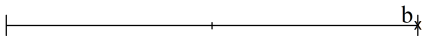
Ebene 3: Bezüglich dieser Ebene ist keine eindeutige Aussage zu treffen, da die dritte Ebene für alle Verfahren und Methoden erfüllt ist, jedoch die vierte Ebene keine eindeutigen Aussagen erzielt und die Ergebnisse nahe dem Grenzwert von 0,05 liegen.

Aus den genannten Gründen kann allgemein zusammengefasst werden, dass es kritisch ist, die vierte Ebene als Qualitätsmerkmal für eine erfolgreiche Fusion anzusehen.

Werden die einzelnen Verfahren und Einflüsse auf die Datenfusionsmethoden miteinander verglichen, können zumindest bestimmte Rückschlüsse getroffen werden, dass unter gewissen Voraussetzungen die Fusion bessere Ergebnisse erzielt. In der Literatur wird bei herkömmlichen Hot-Deck Verfahren in einem Datensatz diskutiert, inwieweit ein Kompromiss zwischen der beschränkten und unbeschränkten Methode eingegangen werden muss (Joenssen, 2015, vgl. S. 111f). Auch bei den in dieser Arbeit ermittelten Ergebnissen ist keine eindeutige Aussage zu treffen. Der ersten und zweiten Spalte der Tabelle 5.8 ist die Bewertung für die unbeschränkte und beschränkte Datenfusion in Abhängigkeit der Durchschnittswerte zu entnehmen.

Insgesamt bestätigen die Ergebnisse die Aussage von D'Orazio et al. (2006, vgl. S.42f), dass der beschränkte Ansatz im Durchschnitt die Randverteilung der Variable Z ($\tilde{f}_Z$) im fusionierten Datensatz besser wiedergibt. Werden weiterhin die durchschnittlichen Ergebnisse betrachtet, wird daraus ersichtlich, dass die unbeschränkte Methode für die Ebenen 1 bis 3 bessere Ergebnisse erzielt und für die beschränkte Methode beide Varianten der Ebene 4 ($\tilde{f}_Z$ und $\tilde{f}_{ZX_p}$). Jedoch muss bei der unbeschränkten Methode die Streuung aufgrund der 500 Durchläufe in der Auswertung mit einbezogen werden. Ausschließlich die Validitätsebene 1 erzielt bessere Ergebnisse für die unbeschränkte Methode. Hier erreicht selbst die schlechteste *Hit Rate* der unbeschränkten Methode vergleichbar gute Ergebnisse wie die beschränkte Methode. Im Gegensatz dazu ist die beschränkte Methode für die Ebene 4 ($\tilde{f}_{ZX_3}$) die geeignetste. Die beschränkte Methode liefert immer so gute Ergebnisse wie die bestmögliche Hellinger-Distanz unter der Verwendung der unbeschränkten Methode.

Tabelle 5.8: Bewertung des unbeschränkten und beschränkten Ansatzes in Abhängigkeit der Validitätsebenen

Validitätsebene	beschränkt	unbeschränkt	Streuung unbeschränkt (u) zu beschränkt (b) schlecht Durchschnitt gut
Ebene 1	-	+	
Ebene 2	-	+	
Ebene 3	-	+	
Ebene 4: $\tilde{f}_Z$	+	-	
Ebene 4: $\tilde{f}_{ZX_1}$ & $\tilde{f}_{ZX_2}$	+	-	
Ebene 4: $\tilde{f}_{ZX_3}$	+	-	

Für Ebene 2 und 4 ($\tilde{f}_Z$, $\tilde{f}_{ZX_1}$ und $\tilde{f}_{ZX_2}$) ist eine eindeutige Tendenz nicht zu erkennen. Hier kann die Streuung sowohl positiven als auch negativen Einfluss nehmen, ohne im Durchschnitt zu besseren Ergebnissen zu führen. Auf Grund der stabileren und somit sichereren Ergebnisse wird jedoch eine Empfehlung für die beschränkte Methode ausgesprochen.

Für die Verwendung von Matching Variablen von unterschiedlicher Qualität kann eine eindeutige Aussage getroffen werden. In jeder Validitätsebene hat das *Distanz-Hot-Deck* Verfahren unter der Verwendung der veränderten Variable *ugeschwbegr.neu* bessere Ergebnisse erzielt als mit der ursprünglichen Variable *ugeschwbegr*. Demnach bestätigt sich der in der Literatur verwendete Vorbereitungsschritt in Bezug auf die Überprüfung der Verteilung der einzelnen Variablen zwischen beiden Datensätzen. Allerdings ist in diesem Zusammenhang der in der Literatur verwendete Grenzwert der Hellinger-Distanz von 0,05 kritisch zu hinterfragen. Bereits die ursprüngliche Variable *ugeschwbegr* ist als ähnlich anzusehen. Jedoch weist diese im Vergleich zur veränderten Variable *ugeschwbegr.neu* schlechtere Ergebnisse auf. In der Literatur wird vor allem über die hohe Aussagekraft der gemeinsamen Variablen gesprochen (D'Orazio et al. (2006, Kap. 6.2) und Rässler (2002, vgl. S.117)), jedoch hat sich bestätigt, dass es mindestens ebenso wichtig ist, dass die gemeinsamen Variablen die gleiche Verteilung in den Datensätzen A und B aufweisen.

5.6.2 Fazit der Ergebnisse der fusionierten Datensätze unter Beibehalten der Unsicherheit

Insgesamt kann für die Datenfusion unter der Verwendung des nicht parametrischen Mikroansatzes mit Beibehalten der Unsicherheit festgestellt werden, dass mit einer Wahrscheinlichkeit von über 99% für jede Verteilung die ermittelten Wahrscheinlichkeitsintervalle den originalen Wert beinhalten. Somit werden in weniger als 1% falsche Schlüsse gezogen. Jedoch kann eine Einschätzung von Zusammenhängen zwischen den Variablen nie eindeutig getroffen werden, sondern ausschließlich in Abhängigkeit der Intervalle.

In Bezug darauf, kann die *Variable-Wise Imputation* der *Domain Imputation* vorgezogen werden, da die durchschnittlichen Intervallbreiten kleiner sind. Demnach erlaubt es diese Methode, genauere Schlüsse zu ziehen.

Für die Verwendung von Matching Variablen mit unterschiedlicher Qualität können bei der Betrachtung der Verteilung dieser Variablen zu der fusionierten Variable *uart* keine kleineren Intervallbreiten, und somit Verbesserungen, in Abhängigkeit einer besseren Qualität der Matching Variablen festgestellt werden. Damit steht das Ergebnis im Widerspruch zu den Ergebnissen aus der Literatur (Endres, 2019, vgl. S.125) und sollte in weiterführenden Untersuchungen genauer analysiert werden.

5.6.3 Vergleich

Die Datenfusion unter der Annahme der bedingten Unabhängigkeit (*CIA*) erzielt Datensätze, bei denen die Auswertungsmöglichkeiten einfacher und präziser anzuwenden sind. Jedoch zeigen die Ergebnisse, dass ausschließlich bivariate Zusammenhänge nachgebildet werden können, nicht jedoch die gemeinsame Verteilung in den fusionierten Datensätzen. Somit würden darauf aufbauende multivariate Analysen zu falschen Schlüssen führen. Im Gegensatz dazu lassen sich mittels *Beibehalten der Unsicherheit* (Imprecise Imputation) die Verteilungen im fusionierten Datensatz besser nachbilden, jedoch kann eine punktuelle Aussage über die Zusammenhänge der Variablen nur unter Berücksichtigung der jeweiligen Wahrscheinlichkeitsintervalle getroffen werden.

Tabelle 5.9 gibt abschließend einen Gesamtüberblick über die Verfahren *CIA* und *Beibehalten der Unsicherheit* und ermöglicht eine Qualitätseinschätzung hinsichtlich der wichtigen Kriterien.

Tabelle 5.9: Qualitätseinschätzung für CIA und Imprecise Imputation (+: positiv, O: neutral, -: negativ, /: nicht zutreffend)

Bewertungskriterien	*CIA*	*Beibehalten der Unsicherheit*
beschränkte Datenfusion	+	/
unbeschränkte Datenfusion	O	/
Zusammenhangsstärke zwischen den Matching Variablen und spezifischen Variablen	+	O
Verteilungsgleichheit der Matching Variablen	+	O
Umgang mit fusionierten Datensätzen (Punktschätzung)	+	-
Genauigkeit der Ergebnisse multivariat	-	+
Genauigkeit der Ergebnisse bivariat	+	+

6 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurden zwei unterschiedliche Datensätze auf Basis der Straßenverkehrsunfallstatistik konstruiert und anschließend mit unterschiedlichen Qualitätsstufen und Datenfusionsverfahren fusioniert. Das Ziel bestand darin, die Qualität der Reproduktion des fusionierten Datensatzes gegenüber des Originaldatensatzes auszuwerten.

Für die Erstellung der simulierten Datensätze wurde der Ausgangsdatensatz der Straßenverkehrsunfallstatistik in drei Schritten an die Rahmenstruktur der Datenbanken GIDAS und EUSKa sowie an die allgemeine Datenfusion angepasst. Nach einer allgemeinen Aufbereitung (Umstrukturierung, Filterung und Löschung von Variablen) im ersten Schritt wurden im zweiten Schritt aus dem Ausgangsdatensatz zwei unabhängige Datensätze mit einem zuvor ermittelten Größenverhältnis von 1:4 gebildet. Anschließend erfolgte im dritten Schritt die Auswahl der zwei spezifischen Variablen. Dafür wurde anhand von zuvor festgelegten Kriterien die Kombination an Variablen ausgewählt, die zu den meisten anderen Variablen einen starken Zusammenhang aufwies. Abschließend setzte sich die Datengrundlage für die Datenfusion aus 27 gemeinsamen Variablen in zwei simulierten Datensätzen (Datensatz A mit 2240 Unfälle und der spezifischen Variable *uortslage* sowie Datensatz B mit 8960 Unfälle und der Variable *uart*) zusammen.

Aufbauend auf dieser Grundlage erfolgte eine Vorbereitung der simulierten Datensätze für die Datenfusion, indem die gemeinsamen Variablen hinsichtlich ihrer Verteilung in beiden Datensätzen analysiert wurden. Daraus ergab sich, dass bis auf vier Variablen alle anderen eine ähnliche Verteilung aufwiesen und für die Auswahl der Matching Variable in Frage kamen. Für die endgültige Auswahl wurde ein zweistufiges Verfahren durchgeführt. Zunächst wurde anhand des Unsicherheitskoeffizienten die Vereinigungsmenge an gemeinsamen Variablen bestimmt, die in Bezug auf die spezifischen Variablen in beiden Datensätzen die stärksten Zusammenhänge aufwies. Dies führte zu neun Variablen, aus denen anschließend durch die Unsicherheitsmethode die endgültigen drei Matching Variablen (*utyp1*, *ustrklasse* und *ugeschwbegr*) bestimmt wurden, die die Unsicherheit zwischen den spezifischen Variablen am effektivsten verringern.

Aufgrund der Datensituation und dem Ziel einen Datensatz zu erhalten, wurde die Datenfusion unter dem Mikroansatz mit nichtparametrischer Methode durchgeführt. Dabei wurde die spezifische Variable *uart* aus dem Spenderdatensatz in den Empfängerdatensatz fusioniert. Um unterschiedliche Fusionsverfahren und Qualitätsstufen in Abhängigkeit der Qualität zu bewerten, wurde zum einen die Datenfusion unter der bedingten Unabhängigkeit (CIA) und zum anderen unter Beibehalten der Unsicherheit durchgeführt.

Unter der CIA wurde sowohl das beschränkte als auch das unbeschränkte Distanz-Hot-Deck Verfahren durchgeführt und miteinander verglichen. Außerdem wurden zwei zusätzliche Qualitätsstufen eingefügt.

Dafür wurde untersucht, inwieweit die Fusion beeinflusst wird, wenn die Qualität einer Matching Variable durch eine Änderung der Kategorienanzahl beeinflusst wird. Eine Voruntersuchung ergab, dass sich die Verteilungen der Variable angeglichen haben. Dementsprechend wurde zusätzlich untersucht, wie sich das Fusionsergebnis verändert, wenn eine der Matching Variablen eine verbesserte Verteilungsgleichheit aufweist. Außerdem wurde anhand einer Kombination aus Distanz-Hot-Deck und Random-Hot-Deck Verfahren untersucht, inwieweit das Fusionsergebnis beeinflusst wird, wenn eine Variable für die Spenderklassen ausgewählt wird, die keinen starken Zusammenhang zu den spezifischen Variablen aufweist. Die vier Validitätsebenen von Rässler bildeten die Bewertungsgrundlage für die Datenfusion unter der CIA.

Die Untersuchungen ergaben, dass keine der untersuchten Verfahren und Methoden die zweite Validitätsebene erfüllte. Somit konnte grundsätzlich keines der Verfahren einen fusionierten Datensatz erzeugen, der den Originaldatensatz exakt reproduzieren konnte. Somit können die fusionierten Datensätze nicht ohne Einschränkungen für statistische Analysen verwendet werden. Allerdings ergibt sich durch das Erfüllen der dritten Ebene, dass die spezifischen Variablen *uortslage* und *uart* in den fusionierten Datensätzen die gleiche Zusammenhangsstruktur wie im originalen Datensatz aufweisen. Somit führen bivariate Analysen, bei denen die Korrelation zwischen zwei Variablen untersucht wird, zu den gleichen Ergebnissen. Es sei an dieser Stelle darauf hingewiesen, dass in der Literatur keine einheitliche Definition existiert, ab welcher Differenz zwischen den Zusammenhangsmaßen von einem ähnlichen Zusammenhang ausgegangen werden kann. Aus diesem Grund ist es sinnvoll, die fusionierten Datensätze hinsichtlich des Zusammenhangs zwischen den spezifischen Variablen in weiterführender Forschung zu untersuchen. In der Literatur wird häufig die vierte Validitätsebene als Mindestanforderung genannt, da unter realen Anwendungen der Datenfusion ausschließlich diese Ebene auf Gültigkeit getestet werden kann. Aufgrund der gezeigten Widersprüche beim Erfüllen der zweiten und dritten Ebene zeigt sich jedoch, dass nicht allein durch die vierte Validitätsebene eine Aussage über das Erfüllen der Datenfusion getroffen werden kann.

Beim Vergleich der beschränkten und unbeschränkten Distanz-Hot-Deck Methode, erwies sich die beschränkte als die empfehlenswerte Methode, da sie stabilere und damit sicherere Ergebnisse erzielte. Bei der Random-Hot-Deck Methode mit verminderter Qualität der Variable veränderte sich die Qualität in Bezug auf die Datenfusion ohne Verwendung der Spenderklasse nicht. In der Literatur wird vor allem über die notwendige hohe Aussagekraft der gemeinsamen Variablen gesprochen, dies bestätigte sich ebenfalls durch die Untersuchungen. Des Weiteren hat sich zusätzlich gezeigt, dass es mindestens genau so wichtig ist, dass die gemeinsamen Variablen die gleiche Verteilung in den Datensätzen aufweisen. Die verbesserte Verteilungsgleichheit durch die veränderte Anzahl an Kategorien einer Matching Variable erzielte für jede Validitätsebene bessere Ergebnisse. Allerdings ist in diesem Zusammenhang der in der Literatur verwendete Grenzwert der Hellinger-Distanz von 0,05 kritisch zu hinterfragen. Bereits die ursprüngliche Variable konnte als ähnlich angesehen werden, obwohl sie im Vergleich zur veränderten Variable schlechtere Ergebnisse aufweist. Eine nähere Beleuchtung des Einflusses des Grenzwerts im Rahmen weiterer Forschung scheint sinnvoll.

Unter dem Ansatz des Beibehaltens der Unsicherheit wurde die *Domain* und *Variable-Wise Imputation* der *Imprecise Imputation* durchgeführt.

Dabei ging hervor, dass die *Variable-Wise Imputation* der *Domain Imputation* vorgezogen werden sollte, da die durchschnittlichen Intervallbreiten kleiner sind und dadurch genauere Schlüsse ermöglicht werden. Die Ergebnisse erzielten keine Verbesserung in Abhängigkeit eines stärkeren Zusammenhangs der Matching Variablen zu den spezifischen Variablen. Damit steht das Ergebnis im Widerspruch zu den Ergebnissen aus der Literatur und sollte in weiterführenden Untersuchungen genauer analysiert werden, indem beispielsweise größere Unterschiede der Zusammenhänge zwischen den Matching Variablen und spezifischen Variablen erlaubt werden.

Aus den Ergebnissen dieser Arbeit lässt sich zusammenfassen, dass die Datenfusion generell unter der bedingten Unabhängigkeitsannahme eine unsichere Methode ist. Bevor eine repräsentative Aussage über Zusammenhänge von fusionierten Variablen für die Bildung von Szenarien für den Testszenarienkatalog getroffen werden kann, ist darauf aufbauende Forschung notwendig. Nach dem Stand dieser Arbeit können fusionierte Daten maximal als Unterstützung verwendet werden, um bivariate Zusammenhänge abzuleiten. Jedoch ist dafür eine sorgfältige Auswahl der gemeinsamen Variablen hinsichtlich ihrer Verteilung in beiden Datensätzen sowie der Stärke des Zusammenhangs zu den spezifischen Variablen notwendig.

Durch die *Imprecise Imputation* konnte ein erster Eindruck gewonnen werden, der gezeigt hat, dass die Verteilungen im fusionierten Datensatz gut nachgebildet werden konnten. Jedoch kann eine Aussage über die Zusammenhänge von Variablen nur unter Berücksichtigung der jeweiligen Wahrscheinlichkeitsintervalle getroffen werden. Daher ist es auch bei diesem Ansatz wichtig, dass in weiterführender Forschung ermittelt wird, inwieweit in Abhängigkeit der Wahrscheinlichkeitsintervalle eine repräsentative Aussage für die Generierung von Szenarien für die Bewertung von automatisierten Fahrfunktionen möglich ist.

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc. ISBN: 978-0-471-36093-3. Verfügbar unter: http://www.gbv.de/dms/ilmenau/toc/349821267.PDF.

Agresti, A. und Yang, M.-C. (1987). „An Empirical Investigation of Some Effects of Sparseness in Contingency Tables“. In: *Computational Statistics & Data Analysis* 5.1, S. 9–21. ISSN: 0167-9473. Verfügbar unter: http://www.sciencedirect.com/science/article/pii/016794738790003X [Zugriff am: 24. 10. 2019].

Barry, J. T. (1988). „An Investigation of Statistical Matching“. In: *Journal of Applied Statistics* 15.3, S. 275–283. ISSN: 0266-4763. Verfügbar unter: https://doi.org/10.1080/02664768800000038 [Zugriff am: 01. 12. 2019].

Barry, R. A., Stewart, W. H. und Turner, J. S. (1982). *An Empirical Evaluation of Statistical Matching Methodologies*. Working Paper 82-804. Verfügbar unter: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1180&context=business_workingpapers [Zugriff am: 14. 11. 2019].

Bönninger, J. (2018). „Fahrszenen Beherrschen – Mit Sicherheit SePIA – Halbzeitpräsentation“. Präsentation (Dresden).

Brunner, H. und Georgi, A. (2003). „Drei Jahre Verkehrsunfallforschung an Der TU Dresden“. In: *ATZ - Automobiltechnische Zeitschrift* 105.2, S. 166–171.

Castillo, P. R. del und Villanueva-Garcia, J. (2018). *Linking Tax Morale and Personal Income Tax in Spain*. 7218. CESifo Group Munich. Verfügbar unter: https://ideas.repec.org/p/ces/ceswps/_7218.html [Zugriff am: 25. 10. 2019].

Cielebak, J. und Rässler, S. (2014). „Data Fusion, Record Linkage und Data Mining“. In: *Handbuch Methoden der empirischen Sozialforschung*. Hrsg. von N. Baur und J. Blasius. Wiesbaden: Springer VS, S. 367–382. ISBN: 978-3-531-17809-7.

D'Orazio, M. (2019). *StatMatch: Statistical Matching or Data Fusion*. Version 1.3.0. Verfügbar unter: https://CRAN.R-project.org/package=StatMatch [Zugriff am: 01. 08. 2019].

D'Orazio, M., Di Zio, M. und Scanu, M. (2006). *Statistical Matching: Theory and Practice*. England: John Wiley & Sons Ltd. ISBN: 978-0-470-02353-2.

D'Orazio, M., Di Zio, M. und Scanu, M. (2019). „Auxiliary Variable Selection in a Statistical Matching Problem". In: *Analysis of Integrated Data*. Hrsg. von L. Zhang und R. L. Chambers. Boca Raton, Florida: CRC Press, S. 101–120. ISBN: 978-1-4987-2798-3.

D'Orazio, M. (2013). „Statistical Matching: Methodological Issues and Practice with R-StatMatch". Präsentation (Vitoria-Gasteiz). Verfügbar unter: `http://www.eustat.eus/productosServicios/datos/Seminario_55_Slides.pdf` [Zugriff am: 17. 09. 2019].

D'Orazio, M., Di Zio, M. und Scanu, M. (2006). „Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints". In: *Journal of Official Statistics* 22.1, S. 137–157.

De Waal, T. (2015). *Statistical Matching: Experimental Results and Future Research Questions*. CBS Discussion Paper 19. The Hague, Heerlen, Bonaire: Statistics Netherlands.

Dudeck, S. G. (2013). *Kamerabasierte In-situ-Überwachung gepulster Laserschweißprozesse*. Bd. 6. Forschungsberichte aus der Industriellen Informationstechnik. Karlsruhe: KIT Scientific Publishing. ISBN: 978-3-7315-0019-3. Verfügbar unter: `http://dx.doi.org/10.5445/KSP/1000034572` [Zugriff am: 26. 10. 2019].

Endres, E.-M. (2019). „Statistical Matching Meets Probabilistic Graphical Models". PhDThesis. Ludwig-Maximilians-Universität München.

Eurostat (2011). *Report of WP2. Recommendations on the Use of Methodologies for the Integration of Surveys and Administrative Data*. ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data.

Fahrmeir, L. et al. (2016). *Statistik: Der Weg Zur Datenanalyse*. 8. überarbeitete und ergänzte Auflage. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-662-50372-0. Verfügbar unter: `https://doi.org/10.1007/978-3-662-50372-0_1` [Zugriff am: 05. 11. 2019].

FDZ der Statistischen Ämter des Bundes und der Länder (2013). *Straßenverkehrsunfallstatistik 2013, Eigene Berechnungen*. Verfügbar unter: `http://dx.doi.org/10.21242/46241.2013.00.00.1.1.0`.

Fink, P., Endres, E. und Schmoll, M. (2019). *impimp: Imprecise Imputation for Statistical Matching*. Version 0.3.1. Verfügbar unter: `https://CRAN.R-project.org/package=impimp` [Zugriff am: 10. 09. 2019].

Grömping, U., Pfeiffer, M. und Stock, W. (2007). *Deliverable 7.1 Statistical Methods for Improving the Usability of Existing Accident Databases*. Project No. 027763—TRACE.

Gschwilm, V. (2017). „Repräsentativität im Statistischen Matching. Bewertung der Datenqualität fusionierter Datensätze im Rahmen einer Simulationsstudie". Magisterarb. LMU München.

Handl, A. und Kuhlenkasper, T. (2017). *Multivariate Analysemethoden: Theorie Und Praxis Mit R*. 3., wesentl. überarb. Aufl. Statistik Und Ihre Anwendungen. Berlin, Heidelberg: Springer Spektrum. ISBN: 978-3-662-54754-0. Verfügbar unter: https://doi.org/10.1007/978-3-662-54754-0_4 [Zugriff am: 06. 10. 2019].

Hanneman, R. A., Kposowa, A. J. und Riddle, M. D. (2012). *Basic Statistics for Social Research*. 1. Aufl. San Francisco, CA: Jossey-Bass. ISBN: 978-0-470-58798-0.

Harrell, F. E. J. (2019). *Hmisc: Harrell Miscellaneous*. Version 4.2-0. Verfügbar unter: https://CRAN.R-project.org/package=Hmisc [Zugriff am: 30. 07. 2019].

Hautzinger, H., Pfeiffer, M. und Schmidt, J. (2006). *Hochrechnung von Daten aus Erhebungen am Unfallort: Bericht zum Forschungsprojekt 82.221/2002*. Berichte der Bundesanstalt für Strassenwesen, Fahrzeugtechnik, Heft 59. Bremerhaven: Wirtschaftsverl. NW, Verl. für neue Wissenschaft. ISBN: 978-3-86509-505-3.

Jann, B. (2011). *Einführung in Die Statistik*. Bearb. von A. Mohr. 2., bearb. Aufl. Hand- Und Lehrbücher Der Sozialwissenschaften. München [u.a.]: Oldenbourg Wissenschaftsverlag GmbH. ISBN: 978-3-486-71092-2. Verfügbar unter: http://www.degruyter.com/doi/book/10.1524/9783486710922.

Joenssen, D. (2015). „Hot-Deck-Verfahren Zur Imputation Fehlender Daten: Auswirkungen Des Donor Limits“. Technische Universität Ilmenau. Verfügbar unter: https://books.google.de/books?id=OyrTjwEACAAJ.

Johannsen, H. (2013). *Unfallmechanik und Unfallrekonstruktion*. 3., überarbeitete Auflage. Wiesbaden: Springer Fachmedien Wiesbaden. ISBN: 978-3-658-01593-0 978-3-658-01594-7.

Kadane, J. B. (1978). „Some Statistical Problems in Merging Data Files“. In: *Compendium of Tax Research*, S. 159–171.

Kaufmann, H. und Pape, H. (1996). „Kapitel 9. Clusteranalyse“. In: *Multivariate Statistische Verfahren*. Hrsg. von L. Fahrmeir, A. Hamerle und G. Tutz. 2. überarb. Aufl. Berlin, Boston: De Gruyter, S. 437–536. ISBN: 978-3-11-081602-0. Verfügbar unter: https://www.degruyter.com/view/books/9783110816020/9783110816020-011/9783110816020-011.xml [Zugriff am: 14. 11. 2019].

Kiesl, H. und Rässler, S. (2005). „Techniken und Einsatzgebiete von Datenintegration und Datenfusion“. In: *Datenfusion und Datenintegration: 6. wissenschaftliche Tagung*. Hrsg. von C. König, M. Stahl und E. Wiegand. Bd. 10. Bonn: Informationszentrum Sozialwissenschaften, S. 17–32. ISBN: 3-8206-0148-1.

Kohn, W. (2005). *Statistik: Datenanalyse Und Wahrscheinlichkeitsrechnung*. Statistik Und Ihre Anwendungen. Berlin Heidelberg: Springer-Verlag. ISBN: 978-3-540-21677-3. Verfügbar unter: https://www.springer.com/de/book/9783540216773 [Zugriff am: 18. 11. 2019].

Kraftfahrt-Bundesamt (2019). *Verzeichnis Zur Systematisierung von Kraftfahrzeugen Und Ihren Anhängern*. Flensburg. Verfügbar unter: https://www.kba.de/SharedDocs/Publikationen/DE/Statistik/SV/sv1_2019_08_pdf.pdf?__blob=publicationFile&v=6 [Zugriff am: 22. 11. 2019].

Kuhn, H. W. (1955). „The Hungarian method for the assignment problem“. In: *Naval Research Logistics Quarterly* 2.1-2, S. 83–97. ISSN: 1931-9193. Verfügbar unter: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109 [Zugriff am: 22. 10. 2019].

Leulescu, A. und Agafiţei, M. (2013). *Statistical Matching: A Model Based Approach for Data Integration*. Luxembourg: Publications Office of the European Union. ISBN: 978-92-79-30355-5.

Lin, J. (Jan. 1991). „Divergence Measures Based on the Shannon Entropy“. In: *IEEE Transactions on Information Theory* 37.1, S. 145–151. ISSN: 0018-9448, 1557-9654.

Meinfelder, F. (2013). „Datenfusion: Theoretische Implikationen Und Praktische Umsetzung“. In: *Weiterentwicklung Der Amtlichen Haushaltsstatistiken*. Hrsg. von T. Riede, N. Ott und S. Bechthold. Berlin: SCIVERO, S. 83–98. ISBN: 978-3-944417-02-8.

Meyer, D. und Buchta, C. (2019). *Proxy: Distance and Similarity Measures*. Version 0.4-23. Verfügbar unter: https://CRAN.R-project.org/package=proxy.

Prokop, G., Hannawald, L. und Köbe, M. (2017). „Szenarien-Basierte Plattform Zur Inspektion Automatisierter Fahrfunktionen - Eine Bewertungsmethodik Zur Inspektion Automatisierter Fahrfunktionen“. In: Symposium für Unfallforschung und Sicherheit im Straßenverkehr der ADAC Stiftung. Verfügbar unter: https://www.researchgate.net/publication/316645327 [Zugriff am: 06. 07. 2019].

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Version 3.6.1. R Foundation for Statistical Computing. Vienna, Austria. Verfügbar unter: https://www.R-project.org [Zugriff am: 29. 07. 2019].

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer Science & Business. ISBN: 978-0-387-95516-2.

Rodgers, W. L. (1984). „An Evaluation of Statistical Matching“. In: *Journal of Business & Economic Statistics* 2.1, S. 91–102.

Rodgers, W. L. und DeVol, E. B. (1981). „An Evaluation of Statistical Matching“. In: *Proceedings of the Survey Research Methods Section, American Statistical Association*, S. 128–132. Verfügbar unter: http://www.asasrms.org/Proceedings/y1981f.html [Zugriff am: 14. 11. 2019].

Rubin, D. B. (1976). „Inference and Missing Data“. In: *Biometrika* 63.3, S. 581–592.

Sächsisches Staatsministerium des Inneren (2019). *Landesentwicklungsplan 2013: Raumstruktur*. Verfügbar unter: https://www.landesentwicklung.sachsen.de/download/Landesentwicklung/karte01-raumstruktur.pdf [Zugriff am: 12. 08. 2019].

Sims, C. A. (1972). „Comments“. In: *Annals of Economic and Social Measurement* 1.3, S. 343–345. Verfügbar unter: https://www.nber.org/books/aesm72-3 [Zugriff am: 20. 11. 2019].

Singh, A. et al. (1993). „Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption“. In: *Survey Methodology* 19.1, S. 59–79. Verfügbar unter: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199300114475 [Zugriff am: 03. 11. 2019].

Statistische Ämter des Bundes und der Länder (2019). *Unfallatlas 2017 | Kartenanwendung*. Verfügbar unter: https://unfallatlas.statistikportal.de/ [Zugriff am: 18. 07. 2019].

Statistisches Bundesamt (DESTATIS) (2017). *Qualitätsbericht: Statistik der Straßenverkehrsunfälle*. Verfügbar unter: https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Verkehrsunfaelle/einfuehrung.html [Zugriff am: 18. 11. 2019].

Statistisches Bundesamt (DESTATIS) (2019a). *Fachserie 8 Reihe 7: Verkehrsunfälle 2018*.

Statistisches Bundesamt (DESTATIS) (2019b). *Verkehrsunfälle Zeitreihen 2018*. Wiesbaden.

Stephanie (2018). *Proportional Reduction in Error (PRE Test)*. Verfügbar unter: https://www.statisticshowto.datasciencecentral.com/proportional-reduction-in-error-pre-test/ [Zugriff am: 19. 11. 2019].

Stocker, T. C. und Steinke, I. (2016). *Statistik, Grundlagen und Methodik*. Berlin, Boston: Walter de Gruyter GmbH. ISBN: 978-3-11-035388-4.

The Institute of Statistical Mathematics (2018). *catdap: Categorical Data Analysis Program Package*. Version 1.3.4. Verfügbar unter: https://CRAN.R-project.org/package=catdap [Zugriff am: 03. 09. 2019].

Unfallforschung der Versicherer (2016). *Unfalltypen-Katalog: Leitfaden Zur Bestimmung Des Unfalltyps*. Berlin. Verfügbar unter: https://udv.de/sites/default/files/tx_udvpublications/unfalltypen-katalog_udv_web_2.pdf [Zugriff am: 06. 12. 2019].

Verkehrsunfallforschung an der TU Dresden GmbH (2019). *Entstehungsgeschichte von GIDAS*. Verfügbar unter: https://www.gidas.org/ueber-gidas/gidas-historie/ [Zugriff am: 12. 08. 2019].

Webber, D. und Tonkin, R. (2013). *Statistical Matching of EU-SILC and the Household Budget Survey to Compare Poverty Estimates Using Income, Expenditures and Material Deprivation.* Eurostat - Methodologies and Working Papers. OCLC: 960405079. Luxembourg: Publications Office. Verfügbar unter: http://dx.publications.europa.eu/10.2785/4151 [Zugriff am: 26. 11. 2019].

Wickham, H. und RStudio (2019). *forcats: Tools for Working with Categorical Variables (Factors)*. Version 0.4.0. Verfügbar unter: `https://CRAN.R-project.org/package=forcats` [Zugriff am: 30. 07. 2019].

Wickham, H. et al. (2019). *dplyr: A Grammar of Data Manipulation*. Version 0.8.3. Verfügbar unter: `https://CRAN.R-project.org/package=dplyr` [Zugriff am: 30. 07. 2019].

Anhang

A.1 Übersicht verwendeter R-Pakete

Alle Berechnungen in dieser Arbeit wurden mit R (R Core Team, 2019) durchgeführt. Verwendet wurden dabei die folgend aufgelisteten R-Pakete:

- catdap, Version 1.3.4 (The Institute of Statistical Mathematics, 2018)
- dplyr, Version 0.8.3 (Wickham et al., 2019)
- forcarts, Version 0.4.0 (Wickham und RStudio, 2019)
- Hmisc, Version, 4.2-0 (Harrell, 2019)
- impimp, Version 0.3.1 (Fink et al., 2019)
- proxy, Version 0.4-23 (Meyer und Buchta, 2019)
- StatMatch, Version 1.3.0 (D'Orazio, 2019)

A.2 Ursachen für Missing Data

Die drei von Rubin (1976) definierten Ausfallmechanismen beschreiben die Gründe, warum Werte in Datensätzen fehlen. Für die Lösung des Missing-Data Problems spielt die Betrachtung des Ausfallmechanismus eine wichtige Rolle, da in Abhängigkeit des Mechanismus bestimmte Einschränkungen bezüglich der Lösungsmethode auftreten. Die Erklärung der Ausfallmechanismen erfolgt nach Cielebak und Rässler (2014, S.373) und Joenssen (2015, Kap. 2.3).

Missing Completely at Random (MCAR)

Im Falle eines MCAR-Mechanimus hängen die fehlenden Werte in einem Datensatz weder von der Variable mit den fehlenden (unbeobachteten) Werten, noch von den vorhandenen (beobachteten) Werten der anderen Variablen ab. Beispielhaft wird ein Datensatz mit den Variablen Alter und Geschlecht betrachtet, wobei die Werte der Altersangabe unvollständig sind. Bei MCAR wird angenommen, dass das Fehlen der Werte weder mit dem Alter, noch mit dem Geschlecht zusammenhängt. Der Datenausfall ist somit rein zufällig und das Fehlen der Werte steht in keiner Beziehung zu den Daten. Insgesamt wird der Datenausfall durch äußere Einflüsse oder Fehler beeinflusst. Im genannten Beispiel wären Fehler bei der Dateneingabe oder Unlesbarkeit mögliche Ursachen.

Missing Random (MAR)

Ist das Auftreten der fehlenden Werte zwar unabhänging von der Variable selbst, aber abhänging von den beobachteten Werten der anderen Variablen, so wird ein MAR-Mechanismus angenommen. Wird der Beispieldatensatz mit den Variablen Alter und Geschlecht wiederholt betrachtet, fehlen die Werte in Abhängigkeit vom Geschlecht, da hier zum Beispiel davon ausgegangen wird, dass die weiblichen Befragten die Altersangabe eher verweigern als die männlichen Befragten.

Missing Not at Random (MNAR)

Im Gegensatz zum MAR hängt beim MNAR der Datenausfall direkt von der Variable mit den fehlenden Werten ab. Demnach ist in Bezug auf den Beispieldatensatz die Wahrscheinlichkeit eines fehlenden Wertes abhängig vom Alter selbst, zum Beispiel durch fehlende Angaben zum Alter bei älteren Personen.

A.3 Ungarische Methode

Für die Erklärung der ungarischen Methode zur Lösung des Zuordnungsproblem bei dem beschränkten Distanz-Hot-Deck Verfahren, werden zwei Datensätze aus dem Beispieldatensatz „iris“, der in R zur Verfügung steht, mit jeweils 6 Beobachtungen erstellt. Anschließend wird mit dem Distanzmaß „Manhattan“ eine Distanzmatrix berechnet, die jeweils die Distanzen zwischen den Beobachtungen angibt. Die Distanzmatrix, die als Ausgangslage für das Zuordnungsproblem dient, ist der Abbildung A.1 zu entnehmen.

d_{ij}		Datensatz B					
		b_1	b_2	b_3	b_4	b_5	b_6
Datensatz A	a_1	0,6	0,2	1,3	0,6	0,5	0,4
	a_2	0,7	0,5	0,6	0,1	1,2	0,5
	a_3	0,3	0,5	0,6	0,3	1,2	0,3
	a_4	0,3	0,7	0,4	0,3	1,4	0,5
	a_5	0,6	0,2	1,3	0,6	0,5	0,4
	a_6	1,3	0,9	2,0	1,3	0,2	1,1

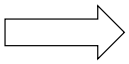
Abbildung A.1: Distanzmatrix

Das Ziel der Ungarischen Methode besteht darin, die Beobachtungen der beiden Datensätze so zuzuordnen, dass die Gesamtdistanz ein Minimum aufweist.

Reduzierung der Distanzmatrix

Im ersten Schritt wird die Distanzmatrix in eine reduzierte Matrix umgewandelt, damit jede Spalte und Zeile mindestens eine Null aufweist. Die Zeilenreduktion ist in Abbildung A.2 dargestellt. Dafür wird für jede Zeile in der Datenmatrix der Minimalwert ermittelt und dieser dann von allen Zahlen in der jeweiligen Zeile subtrahiert. Der Einfachheit halber werden zuvor die Werte aus der Distanzmatrix (Abbildung A.1) mit dem Wert 10 multipliziert, um mit ganzen Zahlen rechnen zu können. In Zeile a_1 ist die 2 der kleinste Werte, deshalb wird in dieser Zeile der Wert 2 von allen Zahlen abgezogen und es entsteht an der zweiten Stelle in a_1 eine Null.

	b_1	b_2	b_3	b_4	b_5	b_6	Min Zeilen
a_1	6	2	13	6	5	4	2
a_2	7	5	6	1	12	5	1
a_3	3	5	6	3	12	3	3
a_4	3	7	4	3	14	5	3
a_5	6	2	13	6	5	4	2
a_6	13	9	20	13	2	11	2

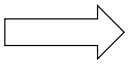


	b_1	b_2	b_3	b_4	b_5	b_6
a_1	4	0	11	4	3	2
a_2	6	4	5	0	11	4
a_3	0	2	3	0	9	0
a_4	0	4	1	0	11	2
a_5	4	0	11	4	3	2
a_6	11	7	18	11	0	9

Abbildung A.2: Zeilenreduktion

Nach der Zeilenreduktion wird die Spaltenreduktion durchgeführt (Vergleich Abbildung A.3). Bei diesem Schritt werden pro Spalte die Minimalwerte ermittel und alle Spaltenwerte um ihr jeweiliges minimales Spaltenelement reduziert. Anschließend enthält die reduzierte Matrix (Vergleich Abbildung A.3 rechts) in jeder Zeile und Spalte mindestens einmal den Wert Null, die damit pro Spalte und Zeile die niedrigsten Distanzen kennzeichnen.

	b_1	b_2	b_3	b_4	b_5	b_6
a_1	4	0	11	4	3	2
a_2	6	4	5	0	11	4
a_3	0	2	3	0	9	0
a_4	0	4	1	0	11	2
a_5	4	0	11	4	3	2
a_6	11	7	18	11	0	9
Min Spalten	0	0	1	0	0	0



	b_1	b_2	b_3	b_4	b_5	b_6
a_1	4	0	10	4	3	2
a_2	6	4	4	0	11	4
a_3	0	2	2	0	9	0
a_4	0	4	0	0	11	2
a_5	4	0	10	4	3	2
a_6	11	7	17	11	0	9

Abbildung A.3: Spaltenreduktion

Zuordnung und Finden der optimalen Lösung

Im zweiten Schritt der ungarischen Methode findet die Zuordnung statt. Es werden die Nullen ermittelt, die unabhängig sind und dementsprechend genau einer Zeile und Spalte zugeordnet werden können. Abbildung A.4 zeigt das Vorgehen der Zuordnung, bei dem wie folgt verfahren wird.

1. Alle Zeilen der reduzierten Matrix von oben nach unten durchgehen und entweder
   - bei **einer** einzelnen Null in der Zeile die Null markieren und die Spalte durchstreichen oder
   - bei **mehreren**, **keiner** oder ausschließlich **durchgestrichenen** Null/Nullen in der Zeile zur nächsten Zeile wechseln

2. Alle Spalten der reduzierten Matrix von links nach rechts durchgehen und entweder
   - bei **einer** einzelnen Null in der Spalte die Null markieren und die Zeile durchstreichen oder
   - bei **mehreren**, **keiner** oder ausschließlich **durchgestrichenen** Null/Nullen in der Spalte zur nächsten Spalte wechseln

Gegebenenfalls den 1. und 2. Schritt wiederholen, damit am Ende alle Nullen durchgestrichen sind. Anschließend wird die Anzahl der markierten Nullen mit der Anzahl der Zeilen verglichen. Stimmen die Werte überein, gibt es ein eindeutiges Ergebnis und die markierten Nullen geben an, welche Zeile (Beobachtungen Datensatz A) mit welcher Spalte (Beobachtungen Datensatz B) miteinander verbunden werden. Die zugehörige Distanz wird dabei für die jeweilige Kombination der Distanzmatrix (Vergleich Abbildung A.1) entnommen.

Stimmen die Werte nicht überein (wie in Abbildung A.4 , markierte Nullen = 5 und Zeilen = 6) wird eine neue Matrix erstellt. Dafür wird das Minimum der nicht durchgestrichenen Werten ermittelt.

	b_1	b_2	b_3	b_4	b_5	b_6
a_1	4	0 [1]	10	4	3	2
a_2	6	4	4	0 [2]	11	4
a_3	0	2	2	0	9	0 [5]
a_4	0	4	0 [4]	0	11	2
a_5	4	0	10	4	3	2
a_6	11	7	17	11	0 [3]	9

Abbildung A.4: Zuordnung 1

In diesem Beispiel ist das der Wert 2. Dieser wird dann bei den nicht durchgestrichenen Werten abgezogen und bei den doppelt durchgestrichenen Werten addiert. Anschließend erhält man die in Abbildung A.5 dargestellte Matrix und fängt erneut an, die unabhängigen Nullen zu identifizieren, indem die Schritte 1 und 2 durchgeführt werden.

	b_1	b_2	b_3	b_4	b_5	b_6
a_1	2	0 [1]	8	4	3	0
a_2	4	4	2	0 [2]	11	2
a_3	0 [6]	4	2	2	11	0
a_4	0	6	0 [5]	2	13	2
a_5	2	0	8	4	3	0 [3]
a_6	9	7	15	11	0 [4]	7

Abbildung A.5: Zuordnung 2

Nach der Identifizierung der Nullen wird erneut die Anzahl der markierten Nullen mit der Zeilenanzahl verglichen. Dieses Mal stimmt die Anzahl der markierten Nullen und die Zeilenanzahl überein und die optimale Zuordnung konnte gefunden werden. Die Zuordnung lautet:

- a_1 mit b_2: Distanz 0,2
- a_2 mit b_4: Distanz 0,1

- a_3 mit b_1: Distanz 0,3
- a_4 mit b_3: Distanz 0,4
- a_5 mit b_6: Distanz 0,4
- a_6 mit b_5: Distanz 0,2

Anhand dieser Aufzählung wird deutlich, dass die Gesamtdistanz minimal ist, aber nicht jede Beobachtungen ihren ähnlichsten Partner zugeteilt bekommen hat. Beispielweise ist die vierte Beobachtung im Datensatz A (a_4) mit der dritten Beobachtung im Datensatz B (b_3) verbunden und weisen eine Distanz von 0,4 auf. Beobachtung 1 (b_1) wäre allerdings mit einer Distanz von 0,3 ähnlicher.

A.4 Anhang Beispielrechnung

A.4.1 Auswahl der Matching Variable durch Unsicherheit

Ausgangssituation:

- Datensatz A: $n_A = 30$ Beobachtungen, spezifische Variable Y mit $J = 2$ Kategorien $\rightarrow P(Y = j)$ für $j = 1, ..., J$
- Datensatz B: $n_B = 20$ Beobachtungen, spezifische Variable Z mit $K = 2$ Kategorien $\rightarrow P(Z = k)$ für $j = 1, ..., K$
- gemeinsame Variablen $\mathbf{X} = (X_1, X_2, X_3)$ (in Datensatz A und Datensatz B enthalten), X1 mit 3 Kategorien, X2 und X3 mit 2 Kategorien

Variablendefinition:

- P ist die Anzahl der verfügbaren X Variablen $\rightarrow P = 3$
- $M = (2^P - 1)$ Gesamtanzahl an Teilmengen die kreiert werden können $\rightarrow M = (2^3 - 1) = 7$
- H_p Anzahl der Kategorien der Variable $X_p \rightarrow$ für Variable $X_1 = X1$ ist $H_1 = 3$
- D_m Variable, die durch die Kreuzklassifizierung einer Teilmenge der gemeinsamen X Variablen entsteht $\rightarrow$ es gibt insgesamt $M = 7$ verschiedene D_m Variablen, eine davon ist $D_m = X1 \times X2$
- H_{D_m} Anzahl der Kategorien der Variable $D_m \rightarrow$ für Beispiel $D_m = X1 \times X2$ ist $H_{D_m} = 3 \times 2 = 6$
- D_P enthält alle P Variablen (letzte Teilmenge) $\rightarrow D_P = X1 \times X2 \times X3$
- H_{D_P} Anzahl der Kategorien der Variable $D_P \rightarrow H_{D_P} = 3 \times 2 \times 2 = 12$

Die Wahl der besten Kombination der X Variablen, die die Unsicherheit am meisten verringert, wird durch 3 Schritte ermittelt.

Schritt 1: Auswahl der X Variable, die die niedrigste Unsicherheit u aufweist

Die gesamte Unsicherheit (ohne Strafterm) wird durch die durchschnittliche Breite der Fréchet Grenzen berechnet. Aus diesem Grund werden als erstes die Fréchet-Grenzen für alle drei X Variablen anhand der Gleichung A.1 mit $p_{hjk} = Pr(X = h, Y = j, Z = k)$ für $h = 1, ..., H, j = 1, ..., J, k = 1, ..., Z$ berechnet.

$$\left[\underline{p}_{+jk}, \overline{p}_{+jk}\right] = \left[\sum_h p_{h++} \max\{0, p_{j|h} + p_{k|h} - 1\}, \sum_h p_{h++} \min\{p_{j|h}, p_{k|h}\}\right] \tag{A.1}$$

Für die Berechnung der Grenzen werden die bedingten relativen Häufigkeiten für Y gegeben X ($p_{j|h}$), für Z gegeben X ($p_{j|h}$) sowie die relativen Randhäufigkeiten X (p_{h++}) benötigt. Die bedingten relativen Häufigkeiten für Y gegeben X ergeben sich aus der Kontingenztabelle X x Y, die mit den Daten aus dem Datensatz A erstellt wird und in Abbildung A.6 a) für X_1 dargestellt ist.

Wird jedes Zeilenelement durch die Zeilensumme (Randhäufigkeit X_1) geteilt, entsteht die bedingte Kontingenztabelle Y|X_1 (Vergleich Abbildung A.6 b)) mit den bedingten relativen Häufigkeiten für Y gegeben X_1. Nach dem gleichen Verfahren entstehen auch die bedingten relativen Häufigkeiten für Z gegeben X_1 (vergleiche Abbildung A.6 e)), nur anhand der Kontingenztabelle X_1 x Z (Vergleich Abbildung A.6 d)) mit den Daten aus dem Datensatz B. Die relativen Randhäufigkeiten für X_1 entstehen aus den Daten beider Datensätzen und sind der Abbildung A.6 c) zu entnehmen.

a) $X_1 \times Y$

X_1\Y	1	2	Σ
1	8	5	13
2	7	1	8
3	4	5	9
Σ	19	11	30

b) Y|X_1

X_1\Y	1	2	Σ
1	$\frac{8}{13} = 0{,}62$	0,38	1
2	0,875	0,125	1
3	0,444	0,556	1

c) X_1

X_1	
1	$\frac{13+7}{30+20} = 0{,}40$
2	$\frac{8+8}{30+20} = 0{,}32$
3	$\frac{9+5}{30+20} = 0{,}28$

d) $X_1 \times Z$

X_1\Z	1	2	Σ
1	2	5	7
2	4	4	8
3	3	2	5
Σ	9	11	20

e) Z|X_1

X_1\Y	1	2	Σ
1	$\frac{2}{7} = 0{,}29$	0,71	1
2	0,50	0,50	1
3	0,60	0,40	1

Abbildung A.6: Kontingenztabellen

Nachfolgend wird zum Verständnis die Berechnung für $X_1 = X1$ und der Kategorie $j = 1$ der Y Variable und der Kategorie $k = 2$ der Z-Variable genauer beschrieben. Die Berechnung der oberen und unteren Grenze sind in den Gleichungen A.2 und A.3 dargestellt. Die Werte werden den Kontingenztabellen aus Abbildung A.6 entnommen und sind rot markiert.

Berechnung der unteren Grenze:

$$\begin{aligned}
\underline{p}_{+12} &= \sum_{h=1}^{3} p_{h++} \max\{0; p_{j|h} + p_{k|h} - 1\} \\
&= p_{1++} \max\{0; p_{1|1} + p_{2|1} - 1\} + p_{2++} \max\{0; p_{1|2} + p_{2|2} - 1\} \\
&\quad + p_{3++} \max\{0; p_{1|3} + p_{2|3} - 1\} \\
&= 0,4 \max\{0; 0,62 + 0,71 - 1\} + 0,32 \max\{0; 0,875 + 0,5 - 1\} \\
&\quad + 0,28 \max\{0; 0,444 + 0,4 - 1\} \\
&= 0,252
\end{aligned} \tag{A.2}$$

Berechnung der oberen Grenze:

$$\begin{aligned}
\overline{p}_{+12} &= \sum_{h=1}^{3} p_{h++} \min\{p_{j|h}; p_{k|h}\} \\
&= p_{1++} \min\{p_{1|1}; p_{2|1}\} + p_{2++} \min\{p_{1|2}; p_{2|2}\} + p_{3++} \min\{p_{1|3}; p_{2|3}\} \\
&= 0,4 \min\{0,62; 0,71\} + 0,32 \min\{0,875; 0,5\} + 0,28 \min\{0,444; 0,4\} \\
&= 0,52
\end{aligned} \tag{A.3}$$

Bei gegebener X_1 liegt demnach die Wahrscheinlichkeit $p_{+12} = Pr(Y = 1, Z = 2)$ im Intervall $\left[\underline{p}_{+12}, \overline{p}_{+12}\right] = [0,252; 0,52]$. Die Berechnung der restlichen Grenzen für die anderen Kategorien von Y und Z sowie für die beiden anderen X_q Variablen (X_2 und X_3) finden mit der Funktion Frechet.bounds.cat() aus dem Paket StatMatch (D'Orazio, 2019) statt. Die Ergebnisse sind in Abbildung A.7 aufgelistet. Die rot markierten Werte entsprechen der schriftlich ausgerechneten oberen und unteren Grenze aus Gleichung A.2 und A.3. Anhand der Fréchet Grenzen kann die Gesamtunsicherheit u_q berechnet werden. In Gleichung A.4 wird die Rechnung beispielhaft für die Gesamtunsicherheit u_1 mit X_1 durchgeführt, indem der Quotient aus der Summe der Intervallbreiten der Fréchet-Grenzen und dem Produkt der Kategorieanzahl der spezifischen Variablen gebildet wird.

$$\begin{aligned}
u_1 &= \frac{1}{J \times K} \sum_{k=1}^{K} \sum_{j=1}^{J} \left(\overline{p}_{+jk}, \underline{p}_{+jk}\right) \\
&= \frac{1}{2 \times 2}((0,40 - 0,13) + (0,31 - 0,04) + (0,52 - 0,25) + (0,31 - 0,04)) = 0,27
\end{aligned} \tag{A.4}$$

X1

	Y	Z	untere Grenze	obere Grenze	Intervallbreite
1	1	1	0,1324	0,3987	0,2663
2	2	1	0,0436	0,3098	0,2663
3	**1**	**2**	**0,2519**	**0,5182**	**0,2663**
4	2	2	0,0396	0,3058	0,2663

X2

	Y	Z	untere Grenze	obere Grenze	Intervallbreite
1	1	1	0,1228	0,4501	0,3273
2	2	1	0,0000	0,3273	0,3273
3	1	2	0,1812	0,5085	0,3273
4	2	2	0,0414	0,3687	0,3273

X3

	Y	Z	untere Grenze	obere Grenze	Intervallbreite
1	1	1	0,0620	0,4444	0,3824
2	2	1	0,0000	0,3824	0,3824
3	1	2	0,1732	0,5556	0,3824
4	2	2	0,0000	0,3824	0,3824

Abbildung A.7: Fréchet-Grenzen

In Tabelle A.1 sind die Unsicherheiten für die beiden anderen Variablen aufgelistet. Anhand dieser Tabelle kann nun die Variable mit der kleinsten Gesamtunsicherheit ausgewählt werden. Für dieses Fallbeispiel ist das die Variable X1 mit einer Unsicherheit von $u_1 = 0,266$.

Tabelle A.1: Schritt 1: Auswahl der X Variable mit der kleinsten Gesamtunsicherheit

	X1	X2	X3
Anzahl Kategorien (H_q)	3	2	2
Gesamtunsicherheit (u_q)	0,266	0,327	0,382
Reihenfolge	1	2	3

Schritt 2: Berechnung der Unsicherheit für alle Möglichkeiten bei Hinzunahme der anderen X Variablen

Unter Berücksichtigung des Strafterms (in dem Beispiel werden beide Strafterme berechnet) wird die Unsicherheit für die Kombination der ausgewählten Variable X1 mit den beiden anderen X Variablen berechnet. Die Gesamtunsicherheit wird wie in Schritt 1 berechnet und wird nicht nochmal speziell dargestellt. Die Berechnung der Unsicherheit mit Strafterm 1 für die Kombination X1 x X2 lautet:

$$\tilde{u}_m^{(1)} = u_m + g^{(1)} = u_m + \log\left(1 + \frac{H_{D_m}}{H_{D_Q}}\right) = 0,215 + \log\left(1 + \frac{6}{3 \times 2 \times 2}\right) = 0,621. \tag{A.5}$$

Für die gleiche Kombination der X Variablen wird die Unsicherheit mit dem zweiten Strafterm wie folgt berechnet:

$$\tilde{u}_m^{(2)} = u_m + g^{(2)} = u_m + \max\left[\frac{1}{(n_A - H_{D_m} \times J)}, \frac{1}{(n_B - H_{D_m} \times K)}\right] \tag{A.6}$$

$$= 0,215 + \max\left[\frac{1}{(30 - 6 \times 2)}, \frac{1}{(20 - 6 \times 2)}\right] = 0,34. \tag{A.7}$$

Für die Gleichung gilt die Bedingung $n_A > (H_{D_m} \times J)$ und $n_B > (H_{D_m} \times K)$, diese ist bei der Rechnung mit $30 > (6 \times 2)$ und $20 > (6 \times 2)$ gegeben. Der Tabelle A.2 sind die Werte für die Unsicherheit der weiteren Kombination X1 x X3 zu entnehmen.

Tabelle A.2: Schritt 2: Berechnung der Gesamtunsicherheit

m	D_m	H_{D_m}	u_m	$\bar{u}_m^{(1)}$	$\bar{u}_m^{(2)}$
0	uneingeschränkt	-	0,367	0,367	0,367
1	X1	3	0,266	0,489	0,337
2	X1 x X2	6	0,215	0,621	0,340
3	X1 x X3	6	0,210	0,615	0,335

Schritt 3: Auswahl Kombination mit kleinster Unsicherheit und Überprüfung der Spärlichkeit

Nachdem der zweite Schritt beendet ist und für jede mögliche Kombination die Gesamtunsicherheit inklusive Strafterm berechnet wurde, wird die Kombination mit der kleinsten Unsicherheit ausgewählt. Der Tabelle A.2 zu entnehmen ist, dass sowohl unter Verwendung des Strafterms 1 als auch des Strafterms 2 die Kombination X1 x X3 die niedrigsten Werte aufweist. Aus diesem Grund wird mit dieser Kombination weiter gerechnet und geprüft, ob die Unsicherheit kleiner ist als die Unsicherheit nur mit der Variable X1 oder uneingeschränkt (ohne Hinzunahme eine X Variable) sowie ob bei der Verwendung des Strafterms 1 der Grad der Spärlichkeit noch akzeptabel ist.

Der Tabelle A.3 ist zu entnehmen, dass für den Strafterm 1 die erste Bedingung nicht erfüllt ist und die Unsicherheit am niedrigsten bei der uneingeschränkten Variante ist.

Tabelle A.3: Schritt 3.1: Berechnung der Gesamtunsicherheit

m	D_m	H_{D_m}	$\bar{n}$	u_m	$\bar{u}_m^{(1)}$	$\bar{u}_m^{(2)}$
0	uneingeschränkt	-	10	0,367	0,367	0,367
1	X1	3	3,333	0,266	0,489	0,337
2	X1 x X3	6	1,667	0,210	0,615	0,335

Aus diesem Grund braucht die Spärlichkeit nicht mehr geprüft werden und unter Verwendung des Strafterms 1 ist das Ergebnis, dass die Unsicherheit am stärksten verringert wird, wenn keine X Variable hinzugezogen wird. Damit aber die Rechnung der Spärlichkeit einmal beispielhaft gezeigt wird, wird die Spärlichkeit für die Kombination X1 x X3 durch die Gleichung A.8 berechnet. Mit einem Wert von 1,667 liegt der Wert noch über dem Stoppkriterium und in Bezug auf die Spärlichkeit wäre die Kombination der X Variablen akzeptabel.

$$\bar{n} = \min\left[\frac{n_A}{H_{D_m} \times J}, \frac{n_B}{H_{D_m} \times K}\right] = \min\left[\frac{30}{6 \times 2}, \frac{20}{6 \times 2}\right] = 1,667 \tag{A.8}$$

Für den Strafterm 2 ist allerdings die erste Bedingung erfüllt und die Gesamtunsicherheit der Kombination X1 x X3 ist am niedrigsten (Vergleich Tabelle A.3).

Aus diesem Grund beginnt ein **neuer Iterationsschritt (Schritt 2)** und es wird die Unsicherheit inklusive Strafterm 2 für alle Möglichkeiten unter Hinzunahme der X Variablen berechnet. In diesem Beispiel bleibt nur noch die Variable X2 übrig und daher gibt es nur eine Möglichkeit, für die die Gesamtunsicherheit berechnet wird. Das Vorgehen ist wie zuvor in Schritt 1 erklärt. Zuerst werden die Fréchet-Grenzen für die Kombination X1 x X3 x X2 berechnet. Das Ergebnis ist in Abbildung A.8 dargestellt.

X1 x X2 x X3

	Y	Z	untere Grenze	obere Grenze	Intervallbreite
1	1	1	0,1764	0,3317	0,1553
2	2	1	0,0500	0,2053	0,1553
3	1	2	0,1919	0,3472	0,1553
4	2	2	0,1044	0,2597	0,1553

Abbildung A.8: Fréchet-Grenzen

Aus den Fréchet-Grenzen wird anschließend die Gesamtunsicherheit und der zweite Strafterm berechnet. Beide Ergebnisse sind im Vergleich zu den Werten der vorherigen Kombinationen in Abbildung A.4 dargestellt.

Tabelle A.4: Schritt 3.2: Berechnung der Gesamtunsicherheit

m	D_m	H_{D_m}	u_m	$\bar{u}_m^{(2)}$
0	uneingeschränkt	-	0,367	0,367
1	X1	3	0,266	0,337
2	X1 x X3	6	0,210	0,335
3	X1 x X3 x X2	12	0,155	1,155

Die Gesamtunsicherheit wird nach Gleichung A.9 berechnet und anschließend die Gesamtunsicherheit inklusive Strafterm 2 durch die Gleichung A.10.

$$\begin{aligned} u_m &= \frac{1}{J \times K} \sum_{k=1}^{K} \sum_{j=1}^{J} \left(\overline{p}_{+jk}, \underline{p}_{+jk} \right) \\ &= \frac{1}{2 \times 2} ((0,332 - 0,176) + (0,205 - 0,050) + (0,347 - 0,192) + (0,260 - 0,104)) \\ &= 0,155 \end{aligned} \tag{A.9}$$

$$\begin{aligned} \tilde{u}_m^{(2)} &= d_m + g^{(2)} = d_m + \max \left[\frac{1}{(n_A - H_{D_m} \times J)}, \frac{1}{(n_B - H_{D_m} \times K)} \right] \\ &= 0,155 + \max \left[\frac{1}{(30 - 12 \times 2)}, \frac{1}{(20 - 12 \times 2)} \right] = 1,155 \end{aligned} \tag{A.10}$$

Für Gleichung A.10 gilt die Bedingung $n_A > (H_{D_m} \times J)$ und $n_B > (H_{D_m} \times K)$. Diese ist für $n_B > (H_{D_m} \times K)$ nicht gegeben, da die Anzahl der Beobachtungen im Datensatz kleiner ist als das Produkt der Kategorien der Kombi X1 x X3 x X2 und der Anzahl der Kategorien der spezifischen Variable Z ($20 \not> (12 \times 2)$). Aus diesem Grund ist die Kombination von X1 x X3 x X2 ungültig und die Überprüfung der Kombination beendet. Das Ergebnis für die Auswahl der Matching Variablen unter Verwendung des zweiten Strafterms ergibt, dass die Kombination aus X1 und X3 die Unsicherheit am stärksten verringert.

A.5 Variablen mit mehreren Kennziffern

Der Tabelle A.5 ist die Anzahl der Unfälle und der prozentuale Anteil zu entnehmen, in denen keine Angaben zu den aufgezählten Variablen eingetragen wurde. Bei den aufgelisteten Zahlenwerten handelt es sich aus Datenschutzgründen um auf die Hunderterstelle gerundete Werte. Die Variablen wurden aus dem Datensatz entnommen, nachdem der Datensatz hinsichtlich der Unfälle mit Personenschaden und PKW-Beteiligung gefiltert wurde. Der Datensatz besteht aus $n = 224200$ beobachteten Unfällen. Auf Grund der Rundung der Werte fehlen bei der dritten Kennziffer der Variable ubesonunfst3 (Besonderheit der Unfallstelle) 100% der Einträge, was wiederum in Wirklichkeit nicht der Fall ist.

Tabelle A.5: Anzahl der Unfälle, die keine Angaben bezüglich der zweiten und dritten Kennziffer einer Variable aufweisen[28]

Variable	Anzahl der Unfälle	Anzahl der Unfälle in Prozent [%]
uursache2	223500	99,69
ucharunfst2	211600	94,38
ucharunfst3	223300	99,60
ubesonunfst2	224000	99,91
ubesonunfst3	224200	100,00
ustrzust2	222500	99,24

[28] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

A.6 Variablenbezeichnungen

Der Tabelle A.6 sind die Variablen sowie ihre Bedeutung und Kategorien zu entnehmen, die in den Datensätzen A und B vorkommen. Die Variablen und Kategorien stammen größtenteils aus der Straßenverkehrsunfallstatistik[29] und unterscheiden sich aufgrund der Datensatzumstellung hinsichtlich der Variablen für die Unfallbeteiligten. Die Kategorien der Kraftfahrzeugklassen sind aufgrund ihres Umfangs nicht in der Tabelle aufgelistet. Eine Auflistung mit Erklärung befindet sich in Kraftfahrt-Bundesamt (2019).

Tabelle A.6: Auflistung der Variablen mit Bedeutung und Kategorien in den simulierten Datensätze (angelehnt an Straßenverkehrsunfallstatistik)

Variable	Bedeutung	Kategorien (Merkmalsausprägungen)
umonat	Monat	1-12
ustunde	Stunde	0-23
uwochentag	Wochentag	1 = Sonntag 2 = Montag 3 = Dienstag 4 = Mittwoch 5 = Donnerstag 6 = Freitag 7 = Samstag
uart	Unfallart	1 = Zusammenstoß mit anfahrendem/anhaltendem/ ruhendem Fahrzeug 2 = Zusammenstoß mit vorausfahrendem/wartendem Fahrzeug 3 = Zusammenstoß mit seitlich in gleicher Richtung fahrendem Fahrzeug 4 = Zusammenstoß mit entgegenkommendem Fahrzeug 5 = Zusammenstoß mit einbiegendem/kreuzendem Fahrzeug 6 = Zusammenstoß zwischen Fahrzeug und Fußgänger 7 = Aufprall auf Fahrbahnhindernis 8 = Abkommen von Fahrbahn nach rechts 9 = Abkommen von Fahrbahn nach links 0 = Unfall anderer Art
ugeschwbegr	Geschwindigkeits-begrenzung	005 = 5 km/h 010 = 10 km/h 015 = 15 km/h 020 = 20 km/h 025 = 25 km/h

[29] FDZ der Statistischen Ämter des Bundes und der Länder, 2013.

		030 = 30 km/h 040 = 40 km/h 050 = 50 km/h 060 = 60 km/h 070 = 70 km/h 080 = 80 km/h 090 = 90 km/h 100 = 100 km/h 110 = 110 km/h 120 = 120 km/h 130 = 130 km/h Z07 = Fußgängerzone/verkehrsberuhigter Bereich (Zone Schrittgeschwindigkeit) Z20 = Zone 20 km/h Z30 = Zone 30 km/h [leer] = ohne Geschwindigkeitsbegrenzung
ulichtverh	Lichtverhältnisse	0 = Tageslicht 1 = Dämmerung 2 = Dunkelheit
ustrzust	Straßenzustand	0 = trocken 1 = nass/feucht 2 = winterglatt 5 = schlüpfrig (durch Öl, Dung, Laub o.ä.)
uaufprhind	Aufprall auf Hindernis	0 = Baum 1 = Mast 2 = Widerlager 3 = Schutzplanke 4 = sonstiges Hindernis 5 = kein Aufprall
uursache	Unfallursache	70 = Verunreinigung durch ausgeflossenes Öl 71 = Andere Verunreinigungen durch Straßenbenutzer 72 = Schnee, Eis 73 = Regen 74 = Andere Einflüsse (u.a. Laub, angeschwemmter Lehm) 75 = Spurrillen, im Zusammenhang mit Regen, Schnee und Eis 76 = Anderer Zustand der Straße 77 = Mangelhafte Beleuchtung der Straße 78 = Nicht ordnungsgemäßer Zustand der Verkehrszeichen oder -einrichtungen

		79 = Mangelhafte Sicherung von Bahnübergängen 80 = Nebel 81 = Starken Regen, Hagel, Schneegestöber usw. 82 = Blendende Sonne 83 = Seitenwind 84 = Unwetter oder sonstige Witterungseinflüsse 85 = Nicht oder unzureichend gesicherte Arbeitsstelle auf der Fahrbahn 86 = Wild auf der Fahrbahn 87 = Anderes Tier auf der Fahrbahn 88 = Sonstiges Hindernis auf der Fahrbahn (ausgenommen haltende, parkende oder liegengebliebende Fahrzeuge oder unzureichend gesicherte Unfallstellen) 89 = Sonstige Ursachen [leer] = Ursachen durch Unfallbeteiligte
ufahrbkfz	min. 1 Kfz nicht mehr fahrbereit	1 = mindestens 1 Kfz nicht mehr fahrbereit [leer] = alle beteiligten Kfz noch fahrbereit
uortslage	Ortslage	1 = innerorts 2 = außerorts
ukategorie	Unfallkategorie	1 = Unfall mit Getöteten 2 = Unfall mit Schwerverletzten 3 = Unfall mit Leichtverletzten
utyp1	Unfalltyp	1 = Fahrunfall 2 = Abbiegeunfall 3 = Einbiegen- /Kreuzen-Unfall 4 = Unfall durch ruhenden Verkehr 5 = Überschreitenunfall 6 = Unfall im Längsverkehr 7 = sonstiger Unfall
utyp2	Untergliederung Unfalltyp	- keine eindeutige Zuordnung möglich -
ustrklasse	Straßenklasse	1 = Autobahn 2 = Bundesstraße 3 = Landesstraße 4 = Kreisstraße 5 = Gemeinde- oder sonstige Straße
ucharunfst	Charakteristik der Unfallstelle	1 = Kreuzung 2 = Einmündung 3 = Grundstücksein-/ausfahrt 4 = Steigung

		5 = Gefälle 6 = Kurve [leer] = ohne
ubesonunfst	Besonderheit der Unfallstelle	2 = schienengleicher Wegübergang 3 = Fußgängerüberweg (Zebrastreifen) 4 = Fußgängerfurt 5 = Haltestelle 6 = Arbeitsstelle 7 = verkehrsberuhigter Bereich [leer] = ohne
ulichtzanl	Lichtzeichenanlage	8 = Lichtzeichenanlage in Betrieb 9 = Lichtzeichenanlage außer Betrieb [] = keine Lichtzeichenanlage
uanzbet	Anzahl der Unfall-beteiligten	1 = 1 Beteiligter 2 = 2 Beteiligte 3 = 3 Beteiligte 4 = 4 Beteiligte 5 = 5 Beteiligte 6 = 6 Beteiligte 7 = 7 Beteiligte 8 = 8 Beteiligte 9 = 9-14 Beteiligte 15 = 15 bis 99 Beteiligte
bursache1_01	Unfallursache des 1. Beteiligten	- keine eindeutige Zuordnung möglich -
balter_c_01	Alter des 1. Beteiligten	1 = 0-14 2 = 15-17 3 = 18-24 4 = 25-64 5 = 65 und älter [] = unbekannt, keine Person
bgeschlecht_01	Geschlecht des 1. Beteiligten	1 = männlich 2 = weiblich [] = ohne
bunfallfolge_01	Unfallfolge beim 1. Beteiligten	1 = getötet 2 = schwer verletzt 3 = leicht verletzt [] = ohne

bfzgklasskba_01	Fahrzeugklasse beim 1. Beteiligten	- zu viele Kategorien[30] -
bursache1_02	Unfallursache des 2. Beteiligten	- keine eindeutige Zuordnung möglich -
balter_c_02	Alter des 2. Beteiligten	1 = 0-14 2 = 15-17 3 = 18-24 4 = 25-64 5 = 65 und älter [] = unbekannt, keine Person kein 2. Beteiligter = kein 2. Beteiligter
bgeschlecht_02	Geschlecht des 2. Beteiligten	1 = männlich 2 = weiblich [] = ohne kein 2. Beteiligter = kein 2. Beteiligter
bunfallfolge_02	Unfallfolge beim 2. Beteiligten	1 = getötet 2 = schwer verletzt 3 = leicht verletzt [] = ohne kein 2. Beteiligter = kein 2. Beteiligter
bfzgklasskba_02	Fahrzeugklasse beim 2. Beteiligten	- zu viele Kategorien[30]-

[30] Kraftfahrt-Bundesamt, 2019.

A.7 Auswertung der spezifischen Variablen

Die in Tabelle A.7, A.8 und A.9 dargestellten Werte bilden die Datengrundlage für Tabelle 3.1.

Tabelle A.7: Berechnung der Stärke des Zusammenhangs zwischen den gemeinsamen Variablen X und der spezifischen Variablen utyp1 (Unfalltyp) und uart (Unfallart) durch den Unsicherheitskoeffizienten U[31]

utyp1 (Y)		uart (Z)	
Variable X	U_{utyp1}	Variable X	U_{uart}
bursache1_01	0,5515	bursache1_01	0,4443
bfzgklasskba_02	0,1874	bfzgklasskba_02	0,2006
bursache1_02	0,1845	bursache1_02	0,1773
balter_c_02	0,1732	balter_c_02	0,1695
		uanzbet	0,1588
ucharunfst1	0,1651	bgeschlecht_02	0,1551
bgeschlecht_02	0,1569	bunfallfolge_02	0,1543
bunfallfolge_02	0,1508	ucharunfst1	0,1311
utyp2	0,1408	utyp2	0,1006
uaufprhind	0,0823	uaufprhind	0,0988
bunfallfolge_01	0,0541	ustrklasse	0,0522
uortslage	0,0491	uortslage	0,0506
uursache1	0,0461	ufahrbkfz	0,0476
ustrklasse	0,0416	bunfallfolge_01	0,0464
ugeschwbegr	0,0388	uursache1	0,0326
ustunde	0,0386	ugeschwbegr	0,0280
ufahrbkfz	0,0381	bfzgklasskba_01	0,0275
bfzgklasskba_01	0,0366	ustunde	0,0251
balter_c_01	0,0274	balter_c_01	0,0205
ustrzust1	0,0269	ustrzust1	0,0195
ubesonunfst1	0,0194	ubesonunfst1	0,0188
ulichtzanl	0,0192	ukategorie	0,0156
umonat	0,0163	ulichtzanl	0,0122
ukategorie	0,0105	umonat	0,0091
uwochentag	0,0104	ulichtverh	0,0088
ulichtverh	0,0094	uwochentag	0,0062
bgeschlecht_01	0,0030	bgeschlecht_01	0,0050

[31] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Tabelle A.8: Berechnung der Stärke des Zusammenhangs zwischen den gemeinsamen Variablen X und der spezifischen Variablen ufahrbkfz (Kfz fahrbereit) und utyp1 (Unfalltyp) durch den Unsicherheitskoeffizienten U[32]

ufahrbkfz (Y)		utyp1 (Z)	
Variable X	$U_{ufahrbkfz}$	Variable X	U_{utyp1}
bfzgklasskba_02	0,1805	bursache1_01	0,5504
uart	0,1455	uart	0,4621
bursache1_01	0,1262	ucharunfst1	0,1724
bursache1_02	0,0977	bfzgklasskba_02	0,1624
uortslage	0,0925	bursache1_02	0,1581
uaufprhind	0,0911	balter_c_02	0,1478
balter_c_02	0,0864	bunfallfolge_02	0,1350
bgeschlecht_02	0,0641	utyp2	0,1343
bunfallfolge_02	0,0624	bgeschlecht_02	0,1336
bunfallfolge_01	0,0619	uaufprhind	0,0770
ustrklasse	0,0572	ustrklasse	0,0512
utyp2	0,0543	uortslage	0,0490
ugeschwbegr	0,0470	bunfallfolge_01	0,0463
bfzgklasskba_01	0,0410	uursache1	0,0390
ustunde	0,0362	bfzgklasskba_01	0,0295
ucharunfst1	0,0349	ugeschwbegr	0,0278
balter_c_01	0,0305	ulichtzanl	0,0269
uursache1	0,0212	ustrzust1	0,0246
ubesonunfst1	0,0166	ustunde	0,0222
ulichtverh	0,0132	balter_c_01	0,0220
ustrzust1	0,0105	ubesonunfst1	0,0188
ukategorie	0,0095	ukategorie	0,0113
uwochentag	0,0084	umonat	0,0091
umonat	0,0057	ulichtverh	0,0086
ulichtzanl	0,0046	uwochentag	0,0050
bgeschlecht_01	0,0001	bgeschlecht_01	0,0041

[32] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

Tabelle A.9: Berechnung der Stärke des Zusammenhangs zwischen den gemeinsamen Variablen X und der spezifischen Variablen uortslage (Ortslage) und uart (Unfallart) durch den Unsicherheitskoeffizienten U[33]

uortslage (Y)		uart (Z)	
Variable X	U	Variable X	U
ustrklasse	0,2946	bursache1_01	0,4443
ugeschwbegr	0,2453	utyp1	0,4102
bursache1_01	0,1688	bfzgklasskba_02	0,2006
bfzgklasskba_02	0,1561	bursache1_02	0,1773
utyp1	0,1380	balter_c_02	0,1695
ucharunfst1	0,1154	uanzbet	0,1588
bursache1_02	0,1154	bgeschlecht_02	0,1551
balter_c_02	0,1100	bunfallfolge_02	0,1543
uanzbet	0,1031	ucharunfst1	0,1311
ufahrbkfz	0,1024	utyp2	0,1006
bgeschlecht_02	0,0981	uaufprhind	0,0988
bunfallfolge_02	0,0945	ustrklasse	0,0522
uaufprhind	0,0924	ufahrbkfz	0,0476
utyp2	0,0560	bunfallfolge_01	0,0464
bunfallfolge_01	0,0508	uursache1	0,0326
ulichtzanl	0,0461	ugeschwbegr	0,0280
uursache1	0,0388	bfzgklasskba_01	0,0275
balter_c_01	0,0263	ustunde	0,0251
ubesonunfst1	0,0237	balter_c_01	0,0205
bfzgklasskba_01	0,0236	ustrzust1	0,0195
ustunde	0,0194	ubesonunfst1	0,0188
ustrzust1	0,0148	ukategorie	0,0156
ukategorie	0,0133	ulichtzanl	0,0122
ulichtverh	0,0044	umonat	0,0091
uwochentag	0,0038	ulichtverh	0,0088
umonat	0,0036	uwochentag	0,0062
bgeschlecht_01	0,0032	bgeschlecht_01	0,0050

[33] FDZ der Statistischen Ämter des Bundes und der Länder, 2013

A.8 Auswahl Distanzmaß

Für die Auswahl des Distanzmaßes werden zwei Beispieldatensätze erstellt, die ebenfalls wie die simulierten Datensätze aus kategorialen Variablen bestehen. Beide Beispieldatensätze bestehen jeweils aus vier gleichen Variablen $\mathbf{X} = (X_1, X_2, X_3, X_4)$ und jeweils einer Datensatzgröße von $n_{1/2} = 10$. Anschließend wird für jedes Distanzmaß die Distanzmatrix berechnet, die die Distanzen zwischen den Beobachtungen der beiden Datensätze beinhaltet. Unter der Verwendung von Distanzmaßen für quantitative Variablen wurden die Variablen zuvor mit der im R-Paket StatMatch (D'Orazio, 2019) enthaltene Funktion fact2dummy() in Dummy Variablen umgewandelt. Für die Berechnung der Distanzmatrizen wird die Funktion dist() aus dem R-Packet Proxy (Meyer und Buchta, 2019) verwendet. Die Funktion berechnet bereits die Ähnlichkeitskoeffizienten nach Gleichung 2.15 in Distanzmaße um. Die Tabellen A.10 - A.16 beinhalten jeweils die Distanzmatrizen für die jeweiligen Distanzmaße aus Kapitel 2.2.1.

Tabelle A.10: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung der Gower-Distanz

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	1.00	0.25	0.75	0.75	1.00	0.75	0.75	0.50	0.75	1.00
n_{1_2}	0.50	1.00	0.50	0.75	0.50	0.50	0.50	1.00	1.00	0.50
n_{1_3}	1.00	0.75	1.00	0.75	0.75	0.50	1.00	0.25	0.25	0.75
n_{1_4}	0.75	0.75	0.75	0.50	0.75	1.00	0.75	1.00	1.00	0.75
n_{1_5}	1.00	0.75	1.00	0.75	0.75	0.50	1.00	0.25	0.00	0.75
n_{1_6}	0.75	0.75	1.00	1.00	1.00	0.75	0.75	0.50	0.50	0.50
n_{1_7}	1.00	0.50	1.00	0.75	1.00	0.75	1.00	0.50	0.75	1.00
n_{1_8}	0.50	0.50	0.75	0.50	0.75	1.00	0.75	1.00	1.00	0.50
n_{1_9}	0.50	1.00	0.50	0.75	0.50	0.50	0.50	1.00	1.00	0.50
$n_{1_{10}}$	0.50	0.75	0.75	0.50	0.75	1.00	0.75	1.00	1.00	0.75

Tabelle A.11: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung der Manhattan-Distanz

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	8	2	6	6	8	6	6	4	6	8
n_{1_2}	4	8	4	6	4	4	4	8	8	4
n_{1_3}	8	6	8	6	6	4	8	2	2	6
n_{1_4}	6	6	6	4	6	8	6	8	8	6
n_{1_5}	8	6	8	6	6	4	8	2	0	6
n_{1_6}	6	6	8	8	8	6	6	4	4	4
n_{1_7}	8	4	8	6	8	6	8	4	6	8
n_{1_8}	4	4	6	4	6	8	6	8	8	4
n_{1_9}	4	8	4	6	4	4	4	8	8	4
$n_{1_{10}}$	4	6	6	4	6	8	6	8	8	6

Tabelle A.12: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung der Euklidischen Distanz

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	2,8284	1,4142	2,4495	2,4495	2,8284	2,4495	2,4495	2,0000	2,4495	2,8284
n_{1_2}	2,0000	2,8284	2,0000	2,4495	2,0000	2,0000	2,0000	2,8284	2,8284	2,0000
n_{1_3}	2,8284	2,4495	2,8284	2,4495	2,4495	2,0000	2,8284	1,4142	1,4142	2,4495
n_{1_4}	2,4495	2,4495	2,4495	2,0000	2,4495	2,8284	2,4495	2,8284	2,8284	2,4495
n_{1_5}	2,8284	2,4495	2,8284	2,4495	2,4495	2,0000	2,8284	1,4142	0,0000	2,4495
n_{1_6}	2,4495	2,4495	2,8284	2,8284	2,8284	2,4495	2,4495	2,0000	2,0000	2,0000
n_{1_7}	2,8284	2,0000	2,8284	2,4495	2,8284	2,4495	2,8284	2,0000	2,4495	2,8284
n_{1_8}	2,0000	2,0000	2,4495	2,0000	2,4495	2,8284	2,4495	2,8284	2,8284	2,0000
n_{1_9}	2,0000	2,8284	2,0000	2,4495	2,0000	2,0000	2,0000	2,8284	2,8284	2,0000
$n_{1_{10}}$	2,0000	2,4495	2,4495	2,0000	2,4495	2,8284	2,4495	2,8284	2,8284	2,4495

Tabelle A.13: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung des umgerechneten Simple-Matching-Koeffizienten

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	0,4211	0,1053	0,3158	0,3158	0,4211	0,3158	0,3158	0,2105	0,3158	0,4211
n_{1_2}	0,2105	0,4211	0,2105	0,3158	0,2105	0,2105	0,2105	0,4211	0,4211	0,2105
n_{1_3}	0,4211	0,3158	0,4211	0,3158	0,3158	0,2105	0,4211	0,1053	0,1053	0,3158
n_{1_4}	0,3158	0,3158	0,3158	0,2105	0,3158	0,4211	0,3158	0,4211	0,4211	0,3158
n_{1_5}	0,4211	0,3158	0,4211	0,3158	0,3158	0,2105	0,4211	0,1053	0,0000	0,3158
n_{1_6}	0,3158	0,3158	0,4211	0,4211	0,4211	0,3158	0,3158	0,2105	0,2105	0,2105
n_{1_7}	0,4211	0,2105	0,4211	0,3158	0,4211	0,3158	0,4211	0,2105	0,3158	0,4211
n_{1_8}	0,2105	0,2105	0,3158	0,2105	0,3158	0,4211	0,3158	0,4211	0,4211	0,2105
n_{1_9}	0,2105	0,4211	0,2105	0,3158	0,2105	0,2105	0,2105	0,4211	0,4211	0,2105
$n_{1_{10}}$	0,2105	0,3158	0,3158	0,2105	0,3158	0,4211	0,3158	0,4211	0,4211	0,3158

Tabelle A.14: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung des umgerechneten Rogers & Tanimoto-Koeffizienten

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	0,5926	0,1905	0,4800	0,4800	0,5926	0,4800	0,4800	0,3478	0,4800	0,5926
n_{1_2}	0,3478	0,5926	0,3478	0,4800	0,3478	0,3478	0,3478	0,5926	0,5926	0,3478
n_{1_3}	0,5926	0,4800	0,5926	0,4800	0,4800	0,3478	0,5926	0,1905	0,1905	0,4800
n_{1_4}	0,4800	0,4800	0,4800	0,3478	0,4800	0,5926	0,4800	0,5926	0,5926	0,4800
n_{1_5}	0,5926	0,4800	0,5926	0,4800	0,4800	0,3478	0,5926	0,1905	0,0000	0,4800
n_{1_6}	0,4800	0,4800	0,5926	0,5926	0,5926	0,4800	0,4800	0,3478	0,3478	0,3478
n_{1_7}	0,5926	0,3478	0,5926	0,4800	0,5926	0,4800	0,5926	0,3478	0,4800	0,5926
n_{1_8}	0,3478	0,3478	0,4800	0,3478	0,4800	0,5926	0,4800	0,5926	0,5926	0,3478
n_{1_9}	0,3478	0,5926	0,3478	0,4800	0,3478	0,3478	0,3478	0,5926	0,5926	0,3478
$n_{1_{10}}$	0,3478	0,4800	0,4800	0,3478	0,4800	0,5926	0,4800	0,5926	0,5926	0,4800

Tabelle A.15: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung des umgerechneten Jaccard-Koeffizienten

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	1,0000	0,4000	0,8571	0,8571	1,0000	0,8571	0,8571	0,6667	0,8571	1,0000
n_{1_2}	0,6667	1,0000	0,6667	0,8571	0,6667	0,6667	0,6667	1,0000	1,0000	0,6667
n_{1_3}	1,0000	0,8571	1,0000	0,8571	0,8571	0,6667	1,0000	0,4000	0,4000	0,8571
n_{1_4}	0,8571	0,8571	0,8571	0,6667	0,8571	1,0000	0,8571	1,0000	1,0000	0,8571
n_{1_5}	1,0000	0,8571	1,0000	0,8571	0,8571	0,6667	1,0000	0,4000	0,0000	0,8571
n_{1_6}	0,8571	0,8571	1,0000	1,0000	1,0000	0,8571	0,8571	0,6667	0,6667	0,6667
n_{1_7}	1,0000	0,6667	1,0000	0,8571	1,0000	0,8571	1,0000	0,6667	0,8571	1,0000
n_{1_8}	0,6667	0,6667	0,8571	0,6667	0,8571	1,0000	0,8571	1,0000	1,0000	0,6667
n_{1_9}	0,6667	1,0000	0,6667	0,8571	0,6667	0,6667	0,6667	1,0000	1,0000	0,6667
$n_{1_{10}}$	0,6667	0,8571	0,8571	0,6667	0,8571	1,0000	0,8571	1,0000	1,0000	0,8571

Tabelle A.16: Berechnung der Distanzen zwischen den Beispieldatensätzen 1 und 2, unter der Verwendung des umgerechneten Dice-Koeffizienten

	n_{2_1}	n_{2_2}	n_{2_3}	n_{2_4}	n_{2_5}	n_{2_6}	n_{2_7}	n_{2_8}	n_{2_9}	$n_{2_{10}}$
n_{1_1}	1,00	0,25	0,75	0,75	1,00	0,75	0,75	0,50	0,75	1,00
n_{1_2}	0,50	1,00	0,50	0,75	0,50	0,50	0,50	1,00	1,00	0,50
n_{1_3}	1,00	0,75	1,00	0,75	0,75	0,50	1,00	0,25	0,25	0,75
n_{1_4}	0,75	0,75	0,75	0,50	0,75	1,00	0,75	1,00	1,00	0,75
n_{1_5}	1,00	0,75	1,00	0,75	0,75	0,50	1,00	0,25	0,00	0,75
n_{1_6}	0,75	0,75	1,00	1,00	1,00	0,75	0,75	0,50	0,50	0,50
n_{1_7}	1,00	0,50	1,00	0,75	1,00	0,75	1,00	0,50	0,75	1,00
n_{1_8}	0,50	0,50	0,75	0,50	0,75	1,00	0,75	1,00	1,00	0,50
n_{1_9}	0,50	1,00	0,50	0,75	0,50	0,50	0,50	1,00	1,00	0,50
$n_{1_{10}}$	0,50	0,75	0,75	0,50	0,75	1,00	0,75	1,00	1,00	0,75

A.9 Ergebnisse

Den Abbildungen A.9 und A.10 sind die Streuungen der Hellinger-Distanzen H_D für die Verfahren aus Kapitel 5 zu entnehmen.

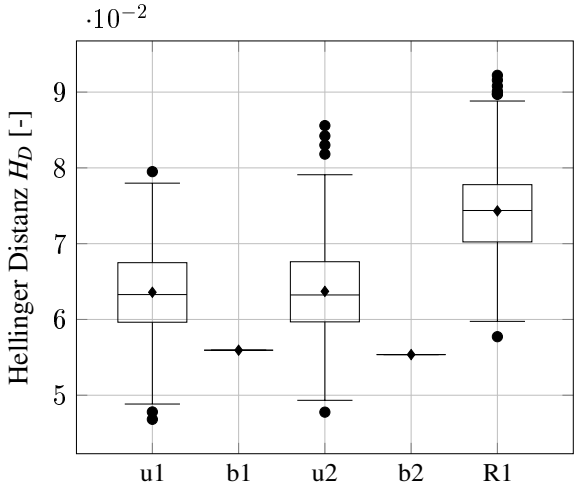


Abbildung A.9: Hellinger-Distanz zum Vergleich von f_{ZX_1} und $\tilde{f}_{ZX_1}$ [34]

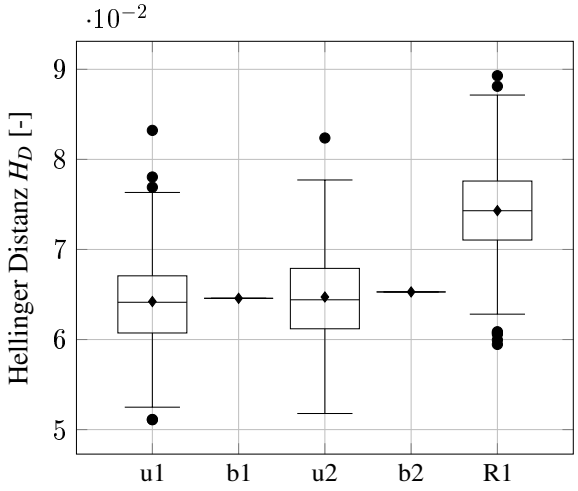


Abbildung A.10: Hellinger-Distanz zum Vergleich von f_{ZX_2} und $\tilde{f}_{ZX_2}$ [34]

[34] FDZ der Statistischen Ämter des Bundes und der Länder, 2013