LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Studienabschlussarbeiten
Fakultät für Mathematik, Informatik und Statistik

Gschwilm, Veronika:

Repräsentativität im Statistischen Matching Bewertung der Datenqualität fusionierter Datensätze im Rahmen einer Simulationsstudie

Masterarbeit, Wintersemester 2017

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.40348

Masterarbeit

Repräsentativität im Statistischen Matching

Bewertung der Datenqualität fusionierter Datensätze im Rahmen einer Simulationsstudie

Veronika Gschwilm

Ludwig-Maximilians-Universität München
Institut für Statistik

8. Februar 2017

Betreuer

Prof. Dr. Thomas Augustin

Eva Endres M.Sc.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

(Ort, Datum)	(Veronika Gschwilm)

Zusammenfassung

In der folgenden Arbeit wird die Qualität von Datensätzen , die durch das Statistische Matching fusioniert wurden, mittels einer Simulationsstudie bewertet. Bei dieser Methode werden für gewöhnlich ähnliche Individuen anhand gemeinsam beobachteter Merkmale identifiziert und ihre Beobachtungen im Anschluss verbunden.
Dabei bietet die Arbeit zunächst eine genaue Einführung in die Theorie des Statistischen Matchings, in dem unter anderem die Problematik des Verfahrens aufgezeigt wird. Für das auftretende Matching-Problem werden gängige Lösungsansätze vorgestellt. Einen häufig verwendeten Ansatz stellt die Annahme der bedingten Unabhängigkeit dar, für den verschiedene Methoden dargestellt werden.
Im weiteren Verlauf der Arbeit werden verschiedene Konzepte zur Bewertung der Qualität fusionierter Datensätze diskutiert. Als Bewertungsgrundlage für die Simulationsstudie werden daraus die vier Validitätsstufen von Rässler (2002) gewählt und verschiedene Kennzahlen zur Messung der einzelnen Stufen entwickelt.
In der Simulationsstudie werden anschließend zwei multivariat normalverteilte Datensätze fusioniert. Dabei wird der Einfluss verschiedener Matching-Methoden und -Szenarien auf die Qualität des fusionierten Datensatzes untersucht. Auch die Auswirkungen einer unterschiedlich hohen Explanatory Power der gemeinsamen Variablen wird analysiert. Einen weiteren Schwerpunkt dieser Simulationsstudie stellen die Auswirkungen einer Verletzung der bedingten Unabhängigkeitsannahme dar.
Die Ergebnisse der Simulationsstudie werden anschließend mit früheren Evaluationsstudien verglichen.

Notation

Generell bezeichnen lateinische sowie griechische Kleinbuchstaben Skalare, sind diese fett gedruckt Vektoren. Fett gedruckte Großbuchstaben stehen generell für Matrizen. Schätzer werden mit einem Zirkumflex ($\wedge$) gekennzeichnet, während fusionierte Beobachtungen bzw. Parameter fusionierter Datensätze mit einer Tilde ($\sim$) markiert werden.

$\mathbf{X} = (X_1, ..., X_p, ..., X_P)$	Gemeinsam erhobene Zufallsvariablen
$\mathbf{Y} = (Y_1, ..., Y_q, ..., Y_Q)$	Nur in A erhobene Zufallsvariablen
$\mathbf{Z} = (Z_1, ..., Z_r, ..., Z_R)$	Nur in B erhobene Zufallsvariablen
$A \in \mathbb{R}^{n_A \times (P+Q)}$	Datensatz $(\mathbf{Y}, \mathbf{Z})$, meist Empfängerdatensatz
$B \in \mathbb{R}^{n_B \times (P+R)}$	Datensatz $(\mathbf{X}, \mathbf{Z})$, meist Spenderdatensatz
$A \cup B$	Vereinigung der Datensätze A und B. $\mathbf{Y}$ ist nur in A und $\mathbf{Z}$ nur in B erhoben
$(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$	vollständiger wahrer Datensatz
$f(\cdot)$	Dichtefunktion einer Wahrscheinlichkeitsverteilung
$F(\cdot)$	Verteilungsfunktion einer Wahrscheinlichkeitsvtlg.
$\boldsymbol{\theta}$	(mehrere) unbekannte Parameter einer Wahrscheinlichkeitsvtlg.
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Normalverteilung mit Mittelwertsvektor $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$
$\boldsymbol{\mu} = (\mu_{X_1}, ..., \mu_{Z_R})^T$	Mittelwertsvektor von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$
$\boldsymbol{\Sigma_{XYZ}} \in \mathbb{R}^{(P+Q+R)\times(P+Q+R)}$	Kovarianzmatrix von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, einzelne Elemente werden mit σ bezeichnet
$\boldsymbol{R_{XYZ}} \in \mathbb{R}^{(P+Q+R)\times(P+Q+R)}$	Korrelationsmatrix von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, einzelne Elemente werden mit r bezeichnet
α	Intercept
$\boldsymbol{\beta} = (\beta_{X_1}, ..., \beta_{X_P})^T$	Vektor der Regressionskoeffizienten von $\mathbf{X}$

Inhaltsverzeichnis

1 Einführung

Die Qualität einer statistischen Analyse hängt zu einem Großteil von der Qualität der untersuchten Daten ab. Eine gute statistische Analyse beginnt demnach also bereits bei der Datenerhebung, deren Ziel das Messen möglichst optimaler Daten ist.
Dies ist allerdings in vielen Forschungsgebieten mit erheblichen Schwierigkeiten verbunden: Besonders für die Datenerhebung durch Personenbefragungen, wie es vor allem in den Sozialwissenschaften oder der Marktforschung durchgeführt wird, ist eine hohe Datenqualität meist nur unter hohem zeitlichen und finanziellen Aufwand zu gewährleisten. Viele Fragestellungen lassen sich zudem nur durch die Berücksichtigung vieler verschiedener Variablen beantworten. In diesem Fall wird die oben genannte Problematik zusätzlich verstärkt, da mit der Dauer einer Befragung die Qualität der Antworten sinkt und gleichzeitig die Zahl der Nonresponses zunimmt (vgl. D'Orazio et al., 2006, S. 1).
Einen alternativen Ansatz zur aufwendigen Erhebung neuer Daten bietet die Fusionierung bereits vorhandener Datenquellen. Ein mögliches Vorgehen stellt dabei das Statistische Matching dar, bei welchem zwei oder mehrere Datensätze so zusammengefügt werden, dass jedem Individuum einer Stichprobe Beobachtungen aus anderen Erhebungen hinzugefügt werden. Mit den so zusammengefügten Daten ist anschließend eine gemeinsame Analyse von Variablen möglich, die nicht zusammen untersucht wurden. Abbildung 1.1 zeigt ein Anwendungsbeispiel aus Kiesl und Rässler (2005).
Durch die Verbindung der Daten eines Konsumenten- und eines Fernsehpanels kann eine gemeinsame Analyse von Verbraucher- und Fernsehverhalten durchgeführt werden, obwohl diese Variablen in unterschiedlichen Stichproben untersucht wurden. Anhand der in beiden Daten erhobenen Variablen, wie Alter und Geschlecht, werden möglichst ähnliche Personen definiert, deren Beobachtungen anschließend verbunden werden.

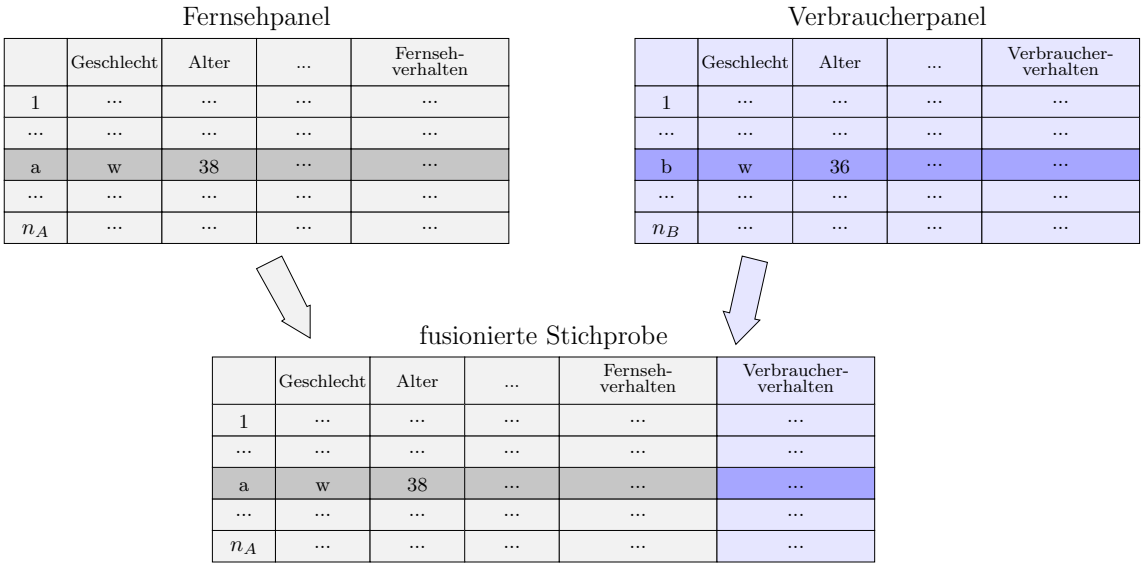

Abbildung 1.1: Vorgang der Fusionierung zweier Datensätze anhand der demographischen Variablen Alter und Geschlecht. In dieser Darstellung wurde die Hot Deck Distance Methode verwendet. Grafik und Beispiel nach Kiesl und Rässler (2005).

Da so die gewünschten Analysen ohne eine erneute Datenerhebung durchgeführt werden können, wirkt diese Methodik besonders in Forschungsgebieten, für die qualitativ hochwertige Daten nur schwer gewonnen werden können, sehr attraktiv. In den letzten Jahren wurde beispielsweise vermehrt auf dem Gebiet der Verbindung administrativer Daten geforscht (siehe z. B. Webber und Tonkin, 2013). Auch die Verbindung administrativer Daten mit Survey-Daten stellt eine Anwendungsmöglichkeit dar (vgl. z. B. Künn, 2015). So können möglicherweise Antworten auf komplexe Fragestellungen gefunden werden, wie sie z. B. in der politikrelevanten Forschung bearbeitet werden. Allerdings stellt das Statistische Matching nur dann eine mögliche Anwendung dar, wenn die daraus resultierenden Daten ein festgelegtes Qualitätslevel erreichen.
In dieser Arbeit wird die Repräsentativität fusionierter Datensätze daher mithilfe einer Simulationsstudie untersucht. Dazu wird in **Kapitel zwei** zunächst eine allgemeine Einführung in die Theorie des Statistischen Matchings gegeben, bei der zum einen die Problemstellung und die getroffenen Annahmen dieses Verfahrens und zum anderen passende Matching-Methoden vorgestellt werden.
Anschließend werden in **Kapitel drei** drei verschiedene Bewertungskriterien zur Beurteilung der Qualität fusionierter Datensätze vorgestellt. Es wird dabei detailliert

auf die vier Validitätsstufen von Rässler (2002, Kap. 2.5) eingegangen, welche die Bewertungsgrundlage für die in der Arbeit durchgeführte Simulationsstudie darstellen. Gleichzeitig werden Kennzahlen zur Messung dieser Stufen festgelegt.
Abschließend werden in **Kapitel vier** das Design der durchgeführten Methode näher beschrieben und die Ergebnisse der Simulationsstudie präsentiert und diskutiert. Die Arbeit schließt in **Kapitel fünf** mit einer Zusammenfassung und einem Ausblick ab.

2 Das Statistische Matching

Um adäquat auf die Bewertung der Qualität eines fusionierten Datensatzes und damit verbundener Probleme eingehen zu können, ist zunächst ein allgemeines Verständnis der Grundlagen des Statistischen Matchings von Nöten. Deshalb wird im folgenden Teil der Arbeit ein Einblick in die Theorie dieser Methode vermittelt.
Zu Beginn wird dabei der Begriff des Statistischen Matchings genau differenziert und eine allgemeingültige Definition festgelegt. Ausgehend davon werden in einem weiteren Abschnitt die Problemstellung beleuchtet und mögliche Lösungsansätze aufgezeigt.
Im Anschluss daran werden verschiedene Methoden des Statistischen Matchings vorgestellt. Den Abschluss des Kapitels bildet ein Überblick über das Statistische Matching in der Literatur.

2.1 Definition und begriffliche Abgrenzung

Im Bereich der Verknüpfung von Datensätzen aus verschiedenen Quellen muss zunächst zwischen zwei verschiedenen Ansätzen, die mit den Begriffen „Record Linkage“ und „Statistisches Matching“ bezeichnet werden, unterschieden werden.
Beim Record Linkage werden Beobachtungen identischer Individuen aus mehreren Datensätzen zusammengefügt. Die Verknüpfung erfolgt meist anhand von Linkvariablen, die ein Objekt eindeutig identifizierbar machen, wie beispielsweise der Name oder die Steuernummer einer Person (vgl. z. B. Rässler, 2002, Kap 1.2). Allerdings besteht eine wichtige Voraussetzung für die Durchführung dieser Methode darin, dass ein ausreichend großer Teil der Individuen in beiden Stichproben enthalten sein muss (vgl. D'Orazio et al., 2006, S. 2). Zudem sind für viele Datensätze die benötigten Informationen für eine Verknüpfung der Beobachtungen nicht vorhanden oder aufgrund von Datenschutzbestimmungen nicht verwertbar. Aus diesen Gründen stellt das Record Linkage häufig nicht das richtige Vorgehen dar.

In diesen Fällen kann alternativ die Methode des Statistischen Matchings verwendet werden. Bei diesem Vorgehen werden ebenfalls zwei oder mehrere Datensätze zu einem Datensatz zusammengefügt; allerdings werden dabei nicht die Beobachtungen identischer Individuen, sondern möglichst ähnlicher Individuen verbunden (vgl. z. B. Rässler, 2002, Kap. 1.2). Diese werden oft als „statistische Zwillinge“ bezeichnet (siehe z. B. Rässler, 2002, S. 17).
Unter Statistischem Matching werden daher Verfahren verstanden, in denen zwei oder mehrere Datensätze zusammengefügt werden, die sich zwar auf dieselbe Grundgesamtheit beziehen, jedoch keine (oder nur sehr wenige) identische Individuen enthalten (vgl. D'Orazio et al., 2006, Kap. 1.2).
Eine mathematische Formulierung dieser Methodik erfolgt im nächsten Abschnitt.

2.2 Ausgangssituation und Problemstellung

Um die mit dem Statistischen Matching verbundene Problemstellung erschließen zu können, wird in diesem Kapitel zugleich eine einheitliche Notation eingeführt, welche weitestgehend aus D'Orazio et al. (2006, Kap. 1.2) übernommen wird.
Aus Gründen der leichteren Verständlichkeit wird im weiteren Verlauf der Arbeit stets von einer Analyse zweier verschiedener Stichprobenerhebungen A und B ausgegangen. Eine zentrale Annahme beim Statistischen Matching besteht darin, dass diese unabhängig voneinander erzeugt wurden und aus n_A bzw. n_B unabhängig und identisch verteilten Beobachtungen bestehen. Diese wurden jeweils durch die wahre gemeinsame Verteilung $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ generiert.
Während jedoch in A nur die Zufallsvektoren $\mathbf{X} = (X_1, ..., X_P)$ und $\mathbf{Y} = (Y_1, ..., Y_Q)$ erhoben wurden, konnten in B nur $\mathbf{X}$ und $\mathbf{Z} = (Z_1, ..., Z_R)$ untersucht werden. Somit erhält man für jede Einheit a mit $a = 1, ..., n_A$ den Beobachtungsvektor

$$(\mathbf{x}_a, \mathbf{y}_a) = (x_{a1}, ..., x_{aP}, y_{a1}, ..., y_{aQ}).$$

Analog kann für jede Untersuchungseinheit b mit $b = 1, ..., n_B$ die Beobachtung

$$(\mathbf{x}_b, \mathbf{z}_b) = (x_{b1}, ..., x_{bP}, z_{b1}, ..., z_{bR})$$

untersucht werden. Diese Ausgangssituation des Statistischen Matchings ist in Abbildung 2.1 veranschaulicht. Grau hinterlegte Zellen markieren dabei fehlende Beobachtungen.

Stichprobe	X_1	...	X_P	Y_1	...	Y_Q	Z_1	...	Z_R
A	x_{11}	...	x_{1P}	y_{11}	...	y_{1Q}			
	x_{11}	...	x_{2P}	y_{21}	...	y_{2Q}			
	...	...	...	...	...	...			
	x_{n_A1}	...	x_{n_AP}	y_{n_A1}	...	y_{n_AQ}			
B	x_{11}	...	x_{1P}				z_{11}	...	z_{1R}
	x_{21}	...	x_{2P}				z_{21}	...	z_{2R}
	...	...	...				...	...	...
	x_{n_B1}	...	x_{n_BP}				z_{n_B1}	...	z_{n_BR}

Abbildung 2.1: Ausgangssituation des Statistischen Matchings für $\mathbf{X} = (X_1, ..., X_P)$, $\mathbf{Y} = (Y_1, ..., Y_Q)$ und $\mathbf{Z} = (Z_1, ..., Z_R)$ nach D'Orazio et al. (2006, S. 5)

Der Datensatz aus Abbildung 2.1 wird auch als $A \cup B$ bezeichnet, während der wahre vollständig beobachtete Datensatz, der in der praktischen Anwendung unbekannt ist, mit $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ angegeben wird. Um Verwechslungen mit dem gleichnamigen Zufallsvektor $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ zu vermeiden, wird es stets kenntlich gemacht, wenn der vollständige Datensatz gemeint ist.

Die Variablen in $\mathbf{X}$ wurden in beiden Stichproben A und B erhoben und werden daher als „gemeinsame Variablen‘ bezeichnet. Anhand dieser Variablen werden die Beobachtungen von A und B verbunden, um so die gemeinsame Verteilung $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ schätzen zu können.

Aufgrund der fehlenden Daten $\mathbf{Z}$ in A und $\mathbf{Y}$ in B liegt es nahe, das Statistische Matching zunächst einer Missing-Data-Problematik, bei der die Daten blockweise fehlen, zuzuordnen. Für diesen Fall werden für gewöhnlich verschiedene Imputationstechniken verwendet, um die fehlenden Daten zu ersetzen.

Allerdings unterscheidet sich $A \cup B$ von einem „gewöhnlichen“ Datensatz mit fehlenden

Werten dadurch, dass $\mathbf{Y}$ und $\mathbf{Z}$ in keiner Beobachtung gemeinsam erhoben wurden und $A \cup B$ somit keinerlei Informationen über die Beziehung dieser Variablen enthält (vgl. D'Orazio et al., 2006, Kap.1.2). Zusätzlich zur Missing-Data-Problematik muss also ein Identifikationsproblem gelöst werden: Mithilfe der zur Verfügung stehenden Daten kann keine Entscheidung darüber getroffen werden, welche gemeinsame Verteilung $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ den Daten zugrunde liegt, da $f_{\mathbf{YZ}}(\mathbf{y}, \mathbf{z})$ nicht spezifiziert werden kann (vgl. Rässler, 2002, Kap. 1.1).
Diese beiden Eigenschaften von $A \cup B$ charakterisieren das „Matching-Problem“ (vgl. D'Orazio et al., 2006, S. 4). Um also die Datensätze A und B erfolgreich fusionieren zu können, muss sowohl die Missing-Data-Problematik als auch das Identifikationsproblem gelöst werden.
Aus diesem Grund wird im nächsten Kapitel zunächst das Statistische Matching als Problem fehlender Daten beleuchtet, bevor anschließend Lösungsansätze für das Identifikationsproblem vorgestellt werden.

2.3 Statistisches Matching als Problem fehlender Daten

Um einen passenden Umgang mit dem unvollständigen Datensatz $A \cup B$ zu gewährleisten, werden zu Beginn die mathematischen Rahmenbedingungen für die fehlenden Daten $\mathbf{Y}$ in A und $\mathbf{Z}$ in B im Kontext des Statistischen Matchings erläutert.
Dazu muss zunächst der Mechanismus der fehlenden Daten in $A \cup B$ untersucht werden. Rubin (1976) definiert hierfür drei verschiedene Missing-Data-Mechanismen: „Missing completely at random“ (MCAR), „Missing at random“ (MAR) und „Missing not at random“ (MNAR) (siehe z. B. auch Enders, 2010, Kap. 1.4).
Sofern alle Annahmen aus Kapitel 2.2 zutreffen, wird das Fehlen der Daten beim Statistischen Matching durch das Stichprobendesign verursacht. Aus diesem Grund kann dem Datensatz $A \cup B$ der MCAR-Mechanismus zugeordnet werden. Bei diesem Mechanismus ist das Fehlen der Werte unabhängig von den Ausprägungen der beobachteten und unbeobachteten Variablen $\mathbf{X}, \mathbf{Y}$ und $\mathbf{Z}$. Detailliertere Informationen zum

Missing-Data-Mechanismus im Statistischen Matching finden sich beispielsweise in D'Orazio et al. (2006, Kap. 1.3).
Unter diesen Umständen ist es möglich, die beobachtete Stichprobenverteilung mit $n_A + n_B$ Einheiten anhand der Likelihood-Funktion

$$L_{A \cup B}(\boldsymbol{\theta}) = \prod_{a=1}^{n_A} f_{\mathbf{XY}}(\mathbf{x}_a, \mathbf{y}_a) \prod_{b=1}^{n_B} f_{\mathbf{XZ}}(\mathbf{x}_b, \mathbf{z}_b) \tag{2.1}$$

zu berechnen (vgl. Rässler, 2002, S. 78).
In einem gewöhnlichen Datensatz mit fehlenden Daten enthält die beobachtete Stichprobenverteilung in diesem Fall die gesamte relevante Information zur Schätzung des unbekannten Parameters $\boldsymbol{\theta}$. Für die Datensituation im Statistischen Matching kann jedoch die korrekte Beziehung von $\mathbf{Y}$ und $\mathbf{Z}$ aufgrund des zuvor beschriebenen Identifikationsproblems nicht eindeutig bestimmt werden. Im nächsten Abschnitt werden daher Lösungswege für diese Problematik aufgezeigt.

2.4 Lösung des Identifikationsproblems

Das Matching-Problem im Statistischen Matching wird durch die Schätzung der gemeinsamen Verteilung $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ aus einer unvollständigen Menge marginaler Verteilungen $f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$ und $f_{\mathbf{XZ}}(\mathbf{x}, \mathbf{z})$ charakterisiert (siehe z. B. Sims, 1972). Dies soll am Beispiel multivariat normalverteilter Daten veranschaulicht werden.
Für diesen Fall wird der interessierende Parameter $\boldsymbol{\theta}$ von

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{XYZ}}) = \left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\mathbf{Y}} \\ \boldsymbol{\mu}_{\mathbf{Z}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XY}} & \boldsymbol{\Sigma}_{\mathbf{XZ}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{Y}} & \boldsymbol{\Sigma}_{\mathbf{YZ}} \\ \boldsymbol{\Sigma}_{\mathbf{ZX}} & \boldsymbol{\Sigma}_{\mathbf{ZY}} & \boldsymbol{\Sigma}_{\mathbf{Z}} \end{bmatrix} \right) \tag{2.2}$$

festgelegt. Für die Datensituation des Statistischen Matchings können der Mittelwertvektor $\boldsymbol{\mu}$ und alle schwarz geschriebenen Elemente der Kovarianzmatrix $\boldsymbol{\Sigma}_{\boldsymbol{XYZ}}$ aus den vorhandenen Daten $A \cup B$ konsistent geschätzt werden. Aufgrund der fehlenden Daten kann jedoch kein Punktschätzer für $\boldsymbol{\Sigma}_{\mathbf{YZ}}$, welches in 2.2 blau markiert wurde, berechnet werden. Dieser kann nur bei perfekter Abhängigkeit der Variablen $\mathbf{Y}$ und $\mathbf{Z}$ von $\mathbf{X}$

bestimmt werden. Liegt diese nicht vor, so kann je nach vorliegender Datensituation bestenfalls der mögliche Wertebereich für $\mathbf{\Sigma_{YZ}}$ eingeschränkt werden. Je stärker dabei der Zusammenhang von $\mathbf{X}$ und $\mathbf{Y}$ bzw. $\mathbf{Z}$ ist, desto schmaler ist der Bereich, der den möglichen Wertebereich von $\mathbf{\Sigma_{YZ}}$ beinhaltet (vgl. Rässler, 2002, S. 10).
Zum besseren Verständnis wird diese Problematik beispielhaft anhand eines standardisierten dreidimensionalen Datensatzes (X, Y, Z) mit

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{\Sigma}_{XYZ}) = \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & 1 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & 1 \end{bmatrix} \right) \tag{2.3}$$

und vorliegender linearer Abhängigkeit veranschaulicht. Aufgrund der standardisierten Daten kann die Kovarianzmatrix $\mathbf{\Sigma}_{XYZ}$ in diesem Fall auch als Korrelationsmatrix $\mathbf{R}_{XYZ}$ interpretiert werden. In Abbildung 2.2 werden die möglichen Wertebereiche von σ_{YZ} für verschiedene festgelegte Werte von σ_{XZ} und σ_{XY} angegeben.

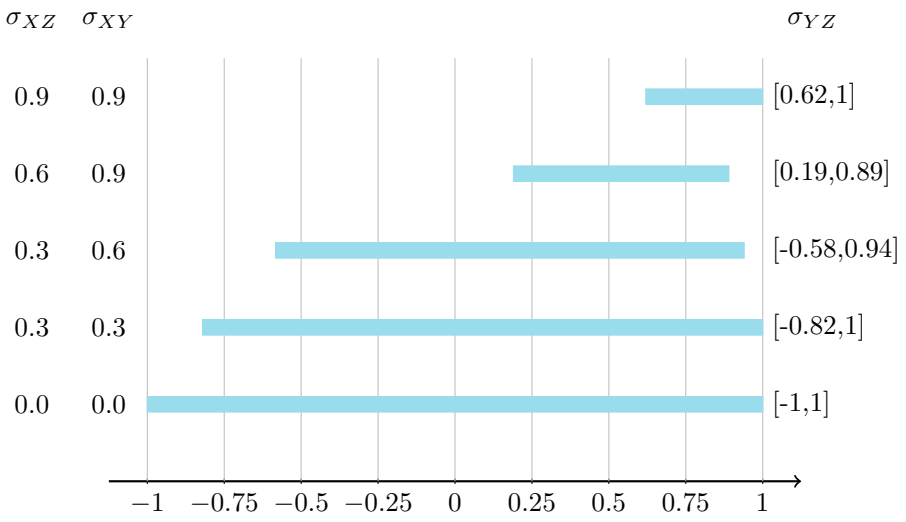

Abbildung 2.2: Möglicher Wertebereich von σ_{YZ} für verschiedene σ_{XZ} und σ_{XY} bei Vorliegen eines dreidimensionalen normalverteilten Datensatzes (X, Y, Z)

Je höher dabei die Werte der Kovarianzen σ_{XY} und σ_{XZ} gewählt werden, desto schmaler ist das Intervall des Wertebereichs für σ_{YZ}. Während für $\sigma_{XZ} = \sigma_{XY} = 0$ keinerlei Restriktionen für diesen Bereich getroffen werden können, kann er für $\sigma_{XZ} = \sigma_{XY} = 0.9$ zumindest auf $[0.62, 1]$ beschränkt werden.

Allerdings ist auch diese Beschränkung von σ_{YZ} für die meisten Analysen nicht genau genug. Möchte man also einen Punktschätzer $\hat{\sigma}_{YZ}$ (bzw. im mehrdimensionalen Fall $\hat{\mathbf{\Sigma}}_{\mathbf{YZ}}$) erhalten, muss eine Lösung für das Identifikationsproblem gefunden werden.
In der wissenschaftlichen Literatur werden dafür mehrere Ansätze vorgeschlagen, von denen drei im Folgenden näher beschrieben werden:
So kann die Unsicherheit als Folge des Identifikationsproblems einfach beibehalten werden (vgl. Kadane, 1978). Außerdem können zusätzliche Informationen hinzugezogen werden, anhand derer Schätzungen für $\mathbf{\Sigma}_{\mathbf{YZ}}$ getroffen werden können (siehe Paass, 1986).
Eine naheliegende Möglichkeit zur Lösung des Matching-Problems stellt das Treffen zusätzlicher Annahmen dar, so dass $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ vollständig spezifiziert werden kann. Die in diesem Kontext am häufigsten verwendete Annahme ist die bedingte Unabhängigkeitsannahme, welche im nächsten Abschnitt vorgestellt wird.

2.4.1 Die bedingte Unabhängigkeitsannahme (CIA)

Die bedingte Unabhängigkeitsannahme oder auch „Conditional Independence Assumption“ (CIA) im Kontext des Statistischen Matchings wurde zunächst von Sims (1972) definiert. Sie besagt, dass die gemeinsame Dichte $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ einer Wahrscheinlichkeitsverteilung eindeutig identifizierbar ist, wenn $\mathbf{Y}$ und $\mathbf{Z}$ gegeben $\mathbf{X}$ voneinander unabhängig sind. Trifft diese Annahme zu, so kann die Beziehung von $\mathbf{Y}$ und $\mathbf{Z}$ vollständig aus der Beziehung von $\mathbf{Y}$ bzw. $\mathbf{Z}$ zu $\mathbf{X}$ abgeleitet werden (vgl. Rodgers, 1984). $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ ist dann aus den vorhandenen Daten $A \cup B$ anhand folgender Formel schätzbar:

$$f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \overset{\text{i.i.d.}}{=} f_{\mathbf{YZ|X}}(\mathbf{y}, \mathbf{z}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \tag{2.4}$$

$$\overset{\text{CIA}}{=} f_{\mathbf{Y|X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{Z|X}}(\mathbf{z}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \tag{2.5}$$

$f_{\mathbf{YZ|X}}(\mathbf{y}, \mathbf{z}|\mathbf{x})$ bezeichnet hierbei die bedingte Dichte von $(\mathbf{Y}, \mathbf{Z})$ gegeben $\mathbf{X}$, während die bedingten Dichtefunktionen von $\mathbf{Y}$ bzw. $\mathbf{Z}$ auf $\mathbf{X}$ durch $f_{\mathbf{Y|X}}(\mathbf{y}|\mathbf{x})$ und $f_{\mathbf{Z|X}}(\mathbf{z}|\mathbf{x})$ dargestellt werden. $f_{\mathbf{X}}(\mathbf{x})$ bezeichnet weiterhin die marginale Dichtefunktion von $\mathbf{X}$.

Diese Form bietet den Vorteil, dass alle Dichtefunktionen in Gleichung 2.5 aus den vorhandenen Daten $A \cup B$ geschätzt werden können (vgl. D'Orazio et al., 2006, S. 13f.). Obwohl die CIA in frühen Anwendungen des Statischen Matchings stets explizit oder implizit angenommen wurde (Rodgers, 1984) und auch moderne Datenfusionen oft unter dieser Annahme durchgeführt werden (siehe z. B. Rasner et al., 2011), ist das Zutreffen der bedingten Unabhängigkeitsannahme für reale Datensituationen oft fragwürdig (vgl. Barr et al., 1982). Beispielsweise würden die gemeinsamen Variablen Alter und Geschlecht im Einführungsbeispiel die Beziehung zwischen Fernseh- und Konsumverhalten nicht vollständig erklären. Um die bedingte Unabhängigkeit gewährleisten zu können, wären demzufolge weitere gemeinsame Variablen mit hoher Explanatory Power auf $\mathbf{Y}$ und $\mathbf{Z}$ nötig (siehe auch Rässler, 2002, S. 57).
Da es keine Möglichkeit gibt, das Zutreffen der bedingten Unabhängigkeitsannahme zu testen, muss die Haltbarkeit der Annahme jeweils anhand von fachlichem Wissen und Erfahrungswerten diskutiert werden. Wird die Annahme nämlich fälschlicherweise getroffen, so können verzerrte Schätzer die Folge sein (siehe bspw. Rodgers, 1984).
Für multivariat normalverteilte Datensätze ist die Erfüllung der CIA gleichbedeutend mit $\mathbf{R_{YZ|X}} = 0$ (siehe z. B. Rässler, 2002, S. 116).

2.4.2 Heranziehen zusätzlicher Informationen

Bestehen Zweifel an der Korrektheit der bedingten Unabhängigkeitsannahme, so stellt das Heranziehen zusätzlicher Informationen einen alternativen Ansatz zur Berechnung eines Punktschätzers für $\boldsymbol{\theta_{YZ}}$ dar (vgl. Paass, 1986). Dazu gibt es verschiedene geeignete Informationsquellen:
Die benötigten Informationen können aus einem dritten Datensatz C stammen, der entweder eine gemeinsame Erhebung von ($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$) oder ($\mathbf{Y}, \mathbf{Z}$) enthält (vgl. Singh et al., 1993). Die so geschaffenen Datensituationen für das Statistische Matching werden in den Abbildungen 2.3 und 2.4 veranschaulicht. Anschließend kann $\boldsymbol{\theta_{YZ}}$ aus $A \cup B \cup C$ geschätzt werden, da die gemeinsamen Beobachtungen von $\mathbf{Y}$ und $\mathbf{Z}$ in C enthalten sind.

Stichprobe	X_1	...	X_P	Y_1	...	Y_Q	Z_1	...	Z_R
A	x_{11}	...	x_{1P}	y_{11}	...	y_{1Q}			
	x_{11}	...	x_{2P}	y_{21}	...	y_{2Q}			
	...	...	...	...	...	...			
	x_{n_A1}	...	x_{n_AP}	y_{n_A1}	...	y_{n_AQ}			
B	x_{11}	...	x_{1P}				z_{11}	...	z_{1R}
	x_{21}	...	x_{2P}				z_{21}	...	z_{2R}
	...	...	...				...	...	...
	x_{n_B1}	...	x_{n_BP}				z_{n_B1}	...	z_{n_BR}
C	x_{11}	...	x_{1P}	y_{11}	...	y_{1Q}	z_{11}	...	z_{1R}
	x_{21}	...	x_{2P}	y_{21}	...	y_{2Q}	z_{21}	...	z_{2R}
	...	...	...	...	...	...	...	...	...
	x_{n_C1}	...	x_{n_CP}	y_{n_C1}	...	y_{n_CQ}	z_{n_C1}	...	z_{n_CR}

Abbildung 2.3: Ausgangssituation des Statistischen Matchings beim Heranziehen eines zusätzlichen Datensatzes C, der alle Variablen $\mathbf{X}, \mathbf{Y}$ und $\mathbf{Z}$ enthält nach Singh et al. (1993)

Stichprobe	X_1	...	X_P	Y_1	...	Y_Q	Z_1	...	Z_R
A	x_{11}	...	x_{1P}	y_{11}	...	y_{1Q}			
	x_{11}	...	x_{2P}	y_{21}	...	y_{2Q}			
	...	...	...	...	...	...			
	x_{n_A1}	...	x_{n_AP}	y_{n_A1}	...	y_{n_AQ}			
B	x_{11}	...	x_{1P}				z_{11}	...	z_{1R}
	x_{21}	...	x_{2P}				z_{21}	...	z_{2R}
	...	...	...				...	...	...
	x_{n_B1}	...	x_{n_BP}				z_{n_B1}	...	z_{n_BR}
C				y_{11}	...	y_{1Q}	z_{11}	...	z_{1R}
				y_{21}	...	y_{2Q}	z_{21}	...	z_{2R}
				...	...	...	...	...	...
				y_{n_C1}	...	y_{n_CQ}	z_{n_C1}	...	z_{n_CR}

Abbildung 2.4: Ausgangssituation des Statistischen Matchings beim Heranziehen eines zusätzlichen Datensatzes C, in dem die Variablen $\mathbf{Y}$ und $\mathbf{Z}$ gemeinsam erhoben wurden nach Singh et al. (1993)

Der Datensatz C kann seinen Ursprung dabei beispielsweise in einer veralteten statistischen Untersuchung oder einer nicht-statistischen Quelle, wie z. B. einem administrativen Register, haben. Sind keine geeigneten Informationen verfügbar, so stellt auch die Durchführung einer Ad-Hoc-Umfrage zur Erhebung der geforderten Daten eine gute Quelle dar (vgl. D'Orazio et al., 2006, S. 67). Dies ist allerdings mit zeitlichem und finanziellem Aufwand verbunden.

Eine weitere Herangehensweise ist die Nutzung plausibler Hilfswerte für die unschätzbaren Parameter von $(\mathbf{Y}, \mathbf{Z}|\mathbf{X})$ oder $(\mathbf{Y}, \mathbf{Z})$. Dafür können beispielsweise Näherungsvariablen $\mathbf{Y}^*$ und $\mathbf{Z}^*$ herangezogen werden, wenn erwartet wird, dass deren Beziehung ähnlich der zwischen $\mathbf{Y}$ und $\mathbf{Z}$ ist (vgl. Singh et al., 1993). Allerdings wird bei diesem Ansatz ebenfalls eine Annahme getroffen, deren Zutreffen nicht überprüft werden kann: Damit diese Methode verlässliche Ergebnisse liefert, müssen die verwendeten Informationen und die vorhandenen Stichproben A und B kompatibel sein und dieselbe zugrunde liegende Verteilung aufweisen (vgl. D'Orazio et al., 2006, S. 67).
Eine gute Einführung in diesen Ansatz ist ebenfalls in D'Orazio et al. (2006, Kap. 3) zu finden.

2.4.3 Beibehalten der Unsicherheit

Die beiden bisher vorgestellten Ansätze beinhalten jeweils Annahmen, deren Zutreffen nicht überprüft werden kann. Dies ist insofern problematisch, da sich diese stets in den Ergebnissen widerspiegeln. Eine Verletzung dieser Annahmen kann daher zu verzerrten Schätzern führen.
In Kadane (1978) wurde daher ein Ansatz vorgestellt, bei dem keine zusätzlichen Annahmen getroffen werden, sondern die Unsicherheit, die durch das Identifikationsproblem verursacht wird, beibehalten wird. Aus diesem Grund kann dieses Vorgehen auch angewandt werden, falls keine zusätzlichen Informationen vorhanden sind und Zweifel am Zutreffen der bedingten Unabhängigkeitsannahme bestehen.
Da die vorhandenen Informationen in $A \cup B$ nicht ausreichen, um sichere und eindeutige Schlüsse über $\boldsymbol{\theta_{YZ}}$ (und damit auch über $\boldsymbol{\theta}$) ziehen zu können, wird für diesen Ansatz ein Parameterraum $\boldsymbol{\Theta}$ mit allen möglichen Schätzern $\boldsymbol{\theta}$ für $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ angegeben. Dieser Parameterraum ist maximal groß, wenn keinerlei Kenntnisse über den wahren unbekannten Schätzer $\boldsymbol{\theta}^*$ vorhanden sind. Aus A und B sind jedoch $\boldsymbol{\theta_{XY}}$ bzw. $\boldsymbol{\theta_{XZ}}$ bekannt, weshalb $\boldsymbol{\Theta}$ durch $\boldsymbol{\theta_{XY}} = \boldsymbol{\theta^*_{XY}}$ und $\boldsymbol{\theta_{XZ}} = \boldsymbol{\theta^*_{XZ}}$ beschränkt werden kann. $\boldsymbol{\theta_{YZ}}$ ist allerdings unbekannt und wird daher durch einen unsicheren Parameter, dessen Wertebereich nicht auf einen Punkt zentriert ist, beschrieben (vgl. D'Orazio et al., 2006, S. 100f.).

Für das Beispiel eines dreidimensionalen multivariat normalverteilten Datensatzes entspricht dieser unsichere Parameter $\boldsymbol{\theta}_{YZ} \in \boldsymbol{\theta}$

$$\boldsymbol{\theta}_{YZ} = (\boldsymbol{\mu}_{YZ}, \boldsymbol{\Sigma}_{YZ}) = \left(\begin{bmatrix} \mu_Y \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \sigma_Y & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z \end{bmatrix} \right),$$

wobei alle Elemente bis auf σ_{YZ} (blau markiert) eindeutig bestimmt werden können. σ_{YZ} kann dabei auf folgendes Intervall eingegrenzt werden:

$$\sigma_{YZ} \in \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_X} \pm \sqrt{\sigma_{Y|X}\sigma_{Z|X}}$$

(vgl. Rässler, 2002, S. 115f.). Für verschiedene feste Werte von σ_{XY} und σ_{XZ} ist ein solches Intervall auch in Abbildung 2.2 veranschaulicht. Moriarity und Scheuren (2001a) zeigen, dass σ_{YZ} unter der bedingten Unabhängigkeitsannahme stets im Mittelpunkt dieses Intervalls liegt.

In Rässler (2002, S. 116) wird die Berechnung des Intervalls auch für ein mehrdimensionales $\mathbf{X}$ angegeben.Diese Intervalle können durch weitere Beschränkungen, die sich bspw. aufgrund der positiven Semidefinitheit der Matrix ergeben, eingeschränkt werden (vgl. Kadane, 1978).

Generell findet sich ein guter Überblick über diesen Ansatz in D'Orazio et al. (2006, Kap. 4) sowie in Rässler (2002), die in diesem Kontext einen multiplen Imputationsansatz vorstellt.

2.5 Methoden des Statistischen Matchings unter der CIA

Im letzten Kapitel wurden mehrere Ansätze vorgestellt, mit denen das Matching-Problem letztendlich gelöst werden kann. Je nach Analyseziel und vorhandener Datenstruktur stellen einer oder mehrere dieser Ansätze eine gute Wahl dar.

In der wissenschaftlichen Literatur werden für jeden Ansatz verschiedene Methoden für das Erstellen eines fusionierten Datensatzes beschrieben. Da in der Simulationsstudie unter anderem die Auswirkungen der bedingten Unabhängigkeitsannahme auf die Qua-

lität des fusionierten Datensatzes untersucht werden sollen, sind jedoch ausschließlich Methoden unter der bedingten Unabhängigkeitsannahme Gegenstand dieses Kapitels. Zudem konzentriert sich der restliche Teil der Arbeit nur auf die Fusionierung stetiger Daten.

Die Funktionsweise der vorgestellten Methoden kann dabei in den „Mikro-“ und den „Makroansatz“ unterteilt werden. Das Ziel des statistischen Matchings, nämlich Aussagen über die Beziehung nicht gemeinsam untersuchter Variablen zu treffen, wird von beiden Ansätzen gleichermaßen erfüllt.

Während jedoch beim Makroansatz die gemeinsame Verteilung $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ oder ihre Eigenschaften geschätzt werden, wird beim Mikroansatz ein vollständiger synthetischer Datensatz $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ erzeugt (vgl. D'Orazio et al., 2006, S. 2f.).

Im Rahmen dieses Kapitels werden die Vor- und Nachteile dieser beiden Ansätze erläutert und für jeden Ansatz verschiedene Methoden zur Durchführung des Statistischen Matchings vorgestellt. Als Primärquelle dieses Kapitels wird D'Orazio et al. (2006, Kap. 2) verwendet, welches einen guten Überblick über diese und weitere Methoden bietet.

2.5.1 Makroansatz

Das Ziel des Makroansatzes besteht, wie bereits erwähnt, in der Schätzung von $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Dieser Ansatz kann sehr effizient eingesetzt werden, allerdings ist die Handhabung bei der späteren Analyse der Daten etwas komplizierter als beim Mikroansatz. Aus diesem Grund wird diese Methode seltener verwendet (vgl. D'Orazio et al., 2006, S. 2f.) und daher auch in dieser Arbeit weniger ausführlich erläutert.

2.5.1.1 Parametrische Methoden

Kann die gemeinsame Wahrscheinlichkeitsverteilung $f_{\mathbf{XYZ}}$ einer parametrischen Verteilungsfamilie $\mathcal{F}$ zugeordnet werden, so kann für die Durchführung des Statistischen Matchings auf parametrische Methoden zurückgegriffen werden. Jede Verteilung $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) \in \mathcal{F}$ wird dabei von $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq T$ definiert. T stellt hierbei eine endliche Zahl dar.

Für den Fall, dass die bedingte Unabhängigkeitsannahme zutrifft, kann $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$ anhand der Gleichung

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}) = f_{\mathbf{X}}(\boldsymbol{x}, \boldsymbol{\theta_X}) f_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_{Y|X}}) f_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta_{Z|X}})$$

vollständig von $\boldsymbol{\theta_X}$, $\boldsymbol{\theta_{Y|X}}$ und $\boldsymbol{\theta_{Z|X}}$ identifiziert werden (siehe D'Orazio et al., 2006, S. 14). Diese Parameter können direkt aus den vorhandenen Daten $A \cup B$ geschätzt werden: Während $\hat{\boldsymbol{\theta}}_{\boldsymbol{X}}$ aus der Gesamtstichprobe $A \cup B$ berechnet werden kann, werden $\hat{\boldsymbol{\theta}}_{\boldsymbol{Y|X}}$ und $\hat{\boldsymbol{\theta}}_{\boldsymbol{Z|X}}$ aus den jeweiligen Teilstichproben A bzw. B geschätzt. Dafür kann zum Beispiel die Maximum-Likelihood-Schätzung verwendet werden. Die dazugehörige Likelihoodfunktion und das genaue Vorgehen sind in D'Orazio et al. (2006, Kap.2.1) dargestellt. Dort werden auch Alternativen zur Verwendung des ML-Schätzers aufgeführt.
Verschiedene Ansätze zur Durchführung dieser Methode für stetige Daten sind im R-Paket `StatMatch` (D'Orazio, 2016) implementiert.

2.5.1.2 Nonparametrische Methoden

Oft enthält die Gesamtstichprobe $A \cup B$ nicht genügend Informationen, um die gemeinsame Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ einer parametrischen Verteilungsfamilie $\mathcal{F}$ zuordnen zu können. In diesem Fall kann $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ anhand nonparametrischer Methoden geschätzt werden.
In D'Orazio et al. (2006, Kap. 2.3) werden dafür mehrere Verfahren vorgestellt: Für das Vorliegen stetiger Daten eignet sich vor allem die Verwendung von Kern-Dichte-Schätzern (siehe Wand und Jones, 1995). Ein weiterer Ansatz konzentriert sich statt auf das Schätzen der kompletten Dichtefunktion $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ lediglich auf das Schätzen wichtiger Eigenschaften dieser Verteilung, wie beispielsweise $\mathbb{E}(Y_q|\mathbf{X})$ oder $\mathbb{E}(Z_r|\mathbf{X})$. Diese Erwartungswerte können anhand der nonparametrischen Regressionsfunktion

$$Z_r = g(\mathbf{X}) + \boldsymbol{\epsilon}_r$$

geschätzt werden. $g(\cdot)$ entspricht dabei einer deterministischen Funktion der Variablen

in $\mathbf{X}$, während $\boldsymbol{\epsilon}_r$ einen additiven Messfehler darstellt (siehe bspw. Fahrmeir et al., 2009, Kap. 7).

2.5.2 Mikroansatz

Beim Mikroansatz wird im Rahmen des Matchingprozesses ein vollständiger Datensatz $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ erstellt. Für die häufigere Anwendung dieses Ansatzes in der Praxis sprechen im Wesentlichen zwei Gründe:
So wird ein vollständiger Datensatz als Datenquelle für statistische Analysen von vielen Personen als zugänglicher empfunden. Zudem können bestimmte statistische Prozeduren nur auf Basis eines vollständigen Datensatzes durchgeführt werden (vgl. D'Orazio et al., 2006, S. 2f.). In der noch folgenden Simulationsstudie wird daher ebenfalls dieser Ansatz verwendet.
Allerdings birgt dieser Ansatz auch Gefahren: So muss man sich bei der Analyse eines fusionierten Datensatzes stets vor Augen halten, dass dieser künstlich erzeugt wurde. Er wurde nicht durch direkte Beobachtung, sondern durch das Ausbeuten der Informationen in $A \cup B$ erzeugt. Die Menge der Informationen aus $A \cup B$ hat sich also nicht erhöht, weshalb der fusionierte Datensatz nicht wie ein Single-Source-Datensatz behandelt werden kann (vgl. D'Orazio et al., 2006, S. 2). Im folgenden Abschnitt werden verschiedene Methoden des Mikroansatzes vorgestellt, mit denen die fehlenden Werte imputiert werden können.
Wie in D'Orazio et al. (2006, Kap.2) werden die vorgestellten Methoden in dieser Arbeit in parametrische, nonparametrische und semiparametrische Vorgehensweisen untergliedert. Aus Gründen der leichteren Verständlichkeit wird in den vorgestellten Methoden stets $\mathbf{Z}$ in A ersetzt. Die Imputation von $\mathbf{Y}$ in B funktioniert jedoch analog.

2.5.2.1 Parametrische Methoden

Bei den parametrischen Methoden des Mikroansatzes (siehe z. B. D'Orazio et al., 2006, Kap. 2.2) wird zunächst eine Schätzung des Parameters $\boldsymbol{\theta}$ der gemeinsamen zugrundeliegenden Verteilung $f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$ getroffen. Dieser Schritt ist identisch zum parametrischen Makroansatz, siehe Kapitel 2.5.1.1.

Anschließend werden die fehlenden Beobachtungen im Datensatz mithilfe von $\hat{\boldsymbol{\theta}}$ geschätzt, so dass man einen vollständigen Datensatz $(\widetilde{\mathbf{X}, \mathbf{Y}, \mathbf{Z}})$ erhält.
In D'Orazio et al. (2006, Kap. 2.2) werden mit dem „Conditional Mean Matching" und „Random Draws" zwei parametrische Matching-Methoden aufgeführt.
Beim „Conditional Mean Matching" wird jeder fehlende Wert mit seinem Erwartungswert gegeben $\mathbf{X}$, $\mathbb{E}(\mathbf{z}_{ar}|\mathbf{x}_a)$,ersetzt. Sind die zu fusionierenden Daten multivariat normalverteilt, so ist dieser Ansatz identisch zur Methode der Regressionsimputation (siehe z. B. Enders, 2010, Kap. 2.7)), bei der der fehlende Wert z_{ar} mit $a = 1, ..., n_A$ und $r = 1, ..., R$ folgendermaßen berechnet wird:

$$\tilde{z}_{ar} = \hat{\alpha} + \mathbf{x}_a \hat{\boldsymbol{\beta}}$$

$\hat{\alpha}$ und $\hat{\boldsymbol{\beta}}$ werden dabei aus Datensatz B geschätzt. Allerdings besitzt die Methode den Nachteil, dass alle Werte $\tilde{z}_{ar}$ auf einer Regressionsgeraden liegen und so die Variabilität in den Daten verloren geht (vgl. D'Orazio et al., 2006, Kap. 2.2.1). Dieser Mangel kann behoben werden, indem die fehlenden Werte durch zufällig gezogene Beobachtungen aus der bedingten prädiktiven Verteilung $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\boldsymbol{Z}|\boldsymbol{X}})$ ersetzt werden, die durch den ML-Schätzer $\hat{\boldsymbol{\theta}}_{\boldsymbol{Z}|\boldsymbol{X}}$ festgelegt wird. Dieses Vorgehen entspricht der Methode der „Random Draws".
Für multivariat normalverteilte Daten ist diese Berechnungsart identisch zur Stochastischen Regressionsimputation (siehe z. B. Enders, 2010, Kap. 2.8). Diese unterscheidet sich vom Conditional Mean Matching durch das Addieren eines Störterms $\hat{\epsilon}_{ar} \sim N(0, \hat{\sigma}_{Z|X})$:

$$\tilde{z}_{ar} = \hat{\alpha} + \mathbf{x}_a \hat{\boldsymbol{\beta}} + \hat{\epsilon}_{ar}$$

$\hat{\sigma}_{Y|X}$ entspricht dabei der Residuenvarianz (vgl. D'Orazio et al., 2006, Kap. 2.2.2).
Für diese Methode kann $\hat{\boldsymbol{\theta}}$ bei Zutreffen der Annahmen zumindest für große Datensätze als approximativ gleich zum realen unbekannten Parameter $\boldsymbol{\theta}$ betrachtet werden. Für das Conditional Mean Matching trifft dies nicht zu, da aufgrund der fehlenden Variabilität stets $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ unterschätzt wird. Aus diesem Grund wird diese Methode allgemein nicht zur Verwendung empfohlen (siehe D'Orazio et al., 2006, Kap. 2.2.3).

Allerdings sollte bei der Anwendung der parametrischen Methoden berücksichtigt werden, dass die imputierten Werte für $\mathbf{Z}$ künstlich sind und zudem lediglich von $\mathbb{E}(\mathbf{Y}|\mathbf{X})$ bzw. $\mathbb{E}(\mathbf{Z}|\mathbf{X})$ abhängen (vgl. D'Orazio et al., 2006, S. 26).

2.5.2.2 Nonparametrische Methoden

Als nonparametrische Methoden (siehe z. B. D'Orazio et al., 2006, Kap. 2.4) werden im Kontext des Statistischen Matchings solche Methoden bezeichnet, die einen vollständigen synthetischen Datensatz $(\widetilde{\mathbf{X}, \mathbf{Y}, \mathbf{Z}})$ erzeugen, ohne die Daten einer bestimmten Verteilungsfamilie zuzuordnen. So kann das Risiko möglicher Fehlspezifikationen der zugrundeliegenden Verteilung umgangen werden.
Obwohl es eine breite Auswahl an nonparametrischen Verfahren im Mikroansatz gibt, beschränkt sich dieser Abschnitt auf die Vorstellung zweier Variationen von „Hot-Deck"-Verfahren (siehe z. B. Enders, 2010, Kap.2.9), wie sie auch bei herkömmlichen Datenimputationen verwendet werden. Bei diesem Ansatz werden die fehlenden Werte durch reale Beobachtungen ersetzt.
Im Statistischen Matching laufen Hot-Deck-Verfahren meist folgendermaßen ab: Einer der beiden Datensätze nimmt die Rolle des „Empfängerdatensatzes" ein, der andere die des „Spenderdatensatzes". Die fehlenden Werte im Empfängerdatensatz werden anschließend durch Beobachtungen aus dem Spenderdatensatz ersetzt (vgl. z. B. D'Orazio et al., 2006, S. 35). Im Einführungsbeispiel (siehe Abbildung 1.1) wird ebenfalls diese Methode mit dem Fernsehpanel als Spenderdatensatz angewandt.
Für den weiteren Verlauf der Arbeit wird A stets als Empfängerdatensatz definiert. In diesem Fall erhält man einen vollständigen Datensatz A, in dem die fehlenden Variablen $\mathbf{Z}$ durch real beobachtete Werte aus dem Spenderdatensatz B ersetzt werden.
Welcher Datensatz dabei welche Rolle einnimmt, muss generell auf Grundlage der jeweiligen Datensituation entschieden werden. Meist wird derjenige Datensatz mit der kleineren Fallzahl als Empfängerdatensatz definiert (vgl. bspw. D'Orazio et al., 2006, S. 35).
Prinzipiell besteht auch die Möglichkeit, dass beide Datensätze sowohl als Empfänger- als auch als Spenderdatensatz verwendet werden und man somit am Ende des Mat-

chingprozesses einen vervollständigten Datensatz $A \cup B$ erhält (vgl. D'Orazio et al., 2006, S. 36). Ob damit eine höhere Qualität des fusionierten Datensatzes erzielt werden kann, wurde in der wissenschaftlichen Literatur noch nicht ausführlich untersucht. Aus diesem Grund wird diese Möglichkeit in der Simulationsstudie in Kapitel 3.4 mit dem herkömmlichen Vorgehen verglichen.

Als weiterer Schritt muss vor der Durchführung einer nonparametrischen Hot-Deck-Methode stets entschieden werden, ob ein beschränktes oder ein unbeschränktes Matching durchgeführt werden soll:

Unbeschränktes Matching bedeutet, dass eine Beobachtung aus dem Spenderdatensatz B mehrmals zur Vervollständigung fehlender Beobachtungen in A herangezogen werden kann. So können stets die Beobachtungen der Individuen, die sich am ähnlichsten sind, fusioniert werden. Allerdings können $f_{\mathbf{Z}}(\mathbf{z})$ und $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})$ mit dieser Methode oft nicht korrekt nachgebildet werden (vgl. Rässler, 2002, Kap. 3.3.1).

Eine bessere Nachbildung dieser Verteilungen kann mit dem **beschränkten Matching** erreicht werden (vgl. z. B. Rodgers, 1984, Rässler, 2002, S. 57f. oder Paass, 1986). Bei diesem Ansatz kann jeder Eintrag in B nur ein einziges Mal als Spender ausgewählt werden, weshalb für dessen Anwendung $n_B \geq n_A$ gelten muss (vgl. D'Orazio et al., 2006, S. 42).

Obwohl für diese Methode die Abstände von $\mathbf{x_a}$ und $\mathbf{x_b}$ für die fusionierten Fälle in der Regel größer als beim unbeschränkten Matching sind und zudem eine höhere Rechenlaufzeit benötigt wird (vgl. D'Orazio et al., 2006, S. 43), wird im weiteren Verlauf der Arbeit daher stets ein beschränktes Matching verwendet.

Anschließend werden beispielhaft für die Hot-Deck-Methodik die Random-Hot-Deck- sowie die Distance-Hot-Deck-Methode vorgestellt:

Bei der **Random-Hot-Deck-Methode** (siehe bspw. D'Orazio et al., 2006, Kap. 2.4.1) wird für jede fehlende Beobachtung $\mathbf{z}_a$ im Empfängerdatensatz A zufällig eine Beobachtung $\mathbf{z}_b$ aus dem Spenderdatensatz B gezogen.

In manchen Fällen ist es jedoch wichtig, dass nur die Beobachtungen von Personen mit den gleichen Eigenschaften bezüglich eines oder mehrerer Merkmale fusioniert werden. Dazu werden sowohl der Empfänger- als auch der Spenderdatensatz anhand

dieser Merkmale in verschiedene homogene Subgruppen, die auch Spenderklassen genannt werden, eingeteilt. Um anschließend die fehlende Beobachtung eines Individuums im Empfängerdatensatz A zu ersetzen, kann nur eine Beobachtung aus derselben Spenderklasse verwendet werden. Solche Spenderklassen werden mithilfe bestimmter Variablen aus $\mathbf{X}$ gebildet, wobei dafür meist bestimmte demographische Eigenschaften wie Geschlecht, Bildungsstand oder Wohnort verwendet werden. Diese sind in der Regel kategorial (vgl. D'Orazio et al., 2006, S. 39).
(Quasi-)stetige Variablen, wie beispielsweise die Körpergröße, können durch das Bilden von Kategorien in Spenderklassen aufgeteilt werden. So kann zumindest gewährleistet werden, dass Beobachtungen von Individuen mit ähnlichen Ausprägungen bezüglich eines Merkmals fusioniert werden. Ein Verfahren dieser Art wird beispielsweise in Ruggles et al. (1977) eingesetzt.
Wird die Random-Hot-Deck-Methode ohne spezielle Spenderklassen angewandt, so impliziert dies die Annahme, dass $\mathbf{Z}$ und $\mathbf{X}$ unabhängig voneinander sind. Diese Annahme kann jedoch im Datensatz B relativ einfach durch das Schätzen von $\hat{F}_{\mathbf{Z}|\mathbf{X}}(\mathbf{x}|\mathbf{z})$ überprüft werden (vgl. D'Orazio et al., 2006, S. 39).

Die **Distance-Hot-Deck-Methode** wurde vor allem in den frühen Veröffentlichungen des Statistischen Matchings verwendet, weshalb mittlerweile viele verschiedene Variationen dieser Methode vorgestellt wurden (siehe z. B. Okner, 1972 oder Rodgers, 1984). Die Funktionsweise der Distance-Hot-Deck-Methode stellt eine sehr intuitive Herangehensweise dar: Jede Beobachtung im Empfängerdatensatz A wird mit derjenigen Beobachtung aus dem Spenderdatensatz fusioniert, welche ihr am ähnlichsten ist. Um dies zu gewährleisten, muss eine klare Definition der Ähnlichkeit zweier Beobachtungen getroffen werden. Bei der Distance-Hot-Deck-Methode wird diese auf Basis einer Abstandsmessung auf Grundlage der gemeinsamen Variablen $\mathbf{X}$ bestimmt.
Der fehlende Beobachtungsvektor $\mathbf{z}_a$ der Beobachtungseinheit a wird folglich durch den Vektor $\mathbf{z}_{b*}$ des Individuums b^* mit

$$d_{ab^*} = \min_{1 \leq b \leq n_b} \|\mathbf{x}_a - \mathbf{x}_b\|$$

ersetzt. Weisen für ein Individuum aus A zwei oder mehr Einträge aus dem Spenderdatensatz B denselben geringsten Abstand d_{ab^*} auf, so wird einer der Einträge zufällig ausgewählt.
Welches Distanzmaß dabei gewählt wird, muss anhand der jeweiligen Datensituation entschieden werden (vgl. D'Orazio et al., 2006, Kap 2.4.3).
Generell ist zu vermuten, dass diese Methode zu schlechteren Ergebnissen führt, wenn für viele Beobachtungen kein passender statistischer Zwilling gefunden werden kann und die Explanatory Power von $\mathbf{X}$ relativ gering ist. In diesem Fall erhöht sich der Matching Noise, auf den später noch genau eingegangen wird (vgl. D'Orazio et al., 2006, Kap. 2.4.4). Eine Eignung dieser Methode sollte daher anhand der vorliegenden Datensituation bewertet werden.
Im R-Paket `StatMatch` (D'Orazio, 2016) findet sich eine Implementierung dieser Methode für kategoriale, gemischte und stetige Daten

2.5.2.3 Gemischte Methoden

Der semiparametrische Ansatz versucht, die Vorteile parametrischer und nonparametrischer Methoden zu vereinen, indem er sich Elemente aus beiden Ansätzen zu Eigen macht: In der Regel verwenden diese Methoden ein parametrisches Modell, welches sparsamer als ein nonparametrisches Modell ist. Zur Vervollständigung des Empfängerdatensatzes A werden jedoch reale Werte verwendet, welche durch nonparametrische Hot-Deck-Methoden imputiert werden. Dadurch wird das Verfahren robuster gegenüber Modellfehlspezifikationen als eine rein parametrische Methode (vgl. D'Orazio et al., 2006, S. 47).
Solche gemischten Verfahren für stetige Variablen wurden zuerst von Kadane (1978) und Rubin (1986) veröffentlicht. Anschließend entwickelten vor allem Moriarity und Scheuren (2001a, 2003, 2004) weitere Variationen dieser Methodik.
Da all diese Methoden einen ähnlichen Ablauf besitzen, wird in diesem Abschnitt nur ein generelles Schema für die Fusionierung eindimensionaler Variablen Y und Z vorgestellt (siehe D'Orazio et al., 2006, Kap. 2.5.1). In diesem wird die Durchführung einer gemischten Methode in drei Schritte unterteilt:

Im ersten Schritt werden die Regressionsschätzer α und $\boldsymbol{\beta}$ auf Basis des Spenderdatensatzes B geschätzt, so dass mit

$$\hat{z}_{ar} = \hat{\alpha} + \mathbf{x}_a \hat{\boldsymbol{\beta}}$$

ein prädiktiver Zwischenwert $\hat{z}_{ar}$ für jede Variable $r = 1, ..., R$ und Beobachtung $a = 1, ..., n_A$ berechnet werden kann. Unter Umständen kann durch das Addieren eines Störterms $\hat{\epsilon}_{ar} \sim N(0, \hat{\sigma}_{Z|X})$ auf $\hat{z}_{ar}$ eine bessere Qualität der fusionierten Daten gewährleistet werden (vgl. Moriarity und Scheuren, 2001a).
Der letzte Schritt besteht darin, für jede Beobachtung $a = 1, ..., n_A$ einen real beobachteten Vektor $\mathbf{z}_{b^*}$ zu imputieren. b^* wird durch eine passende Distance-Hot-Deck-Methode gefunden, welche $d(\hat{\mathbf{z}}_a, \mathbf{z}_b)$ minimiert.
Bei alternativen Methoden wird auch $d(\hat{\mathbf{z}}_a, \hat{\mathbf{z}}_b)$ minimiert, wobei $\hat{z}_{br}$ die Vorhersage für z_{br} auf Grundlage des berechneten linearen Modells darstellt (siehe z. B. D'Orazio et al., 2006, Kap. 2.5.1). Dort wurden auch einige dieser gemischten Verfahren zusammengefasst.
Generell ist die gemischte Methode robuster bezüglich Missspezifikationen im Modell als eine parametrische Methode. Allerdings sollte stets beachtet werden, dass sie aufgrund der Verwendung nonparametrischer Bausteine stärker durch Matching Noise beeinträchtigt werden kann (vgl. D'Orazio et al., 2006, S. 49).
Eine Implementierung in R ist ebenfalls im Paket `StatMatch` (D'Orazio, 2016) gegeben.

2.6 Diskussion in der wissenschaftlichen Literatur

Das Statistische Matching stellt eine kontroverse Methode dar, über die in der wissenschaftlichen Literatur viel diskutiert wurde. In diesem Kapitel folgt daher eine kurze Zusammenfassung verschiedener Stimmen zum Statistischen Matching.
Aufgrund der in der Einführung genannten Vorteile wurden Datenfusionierungen durch das Statistische Matching bereits in vielen Bereichen durchgeführt. D'Orazio et al. (2006, Kap. 7) nennen mit der Marktforschung, der amtlichen Statistik und der Mikrosimulation drei Hauptanwendungsgebiete:

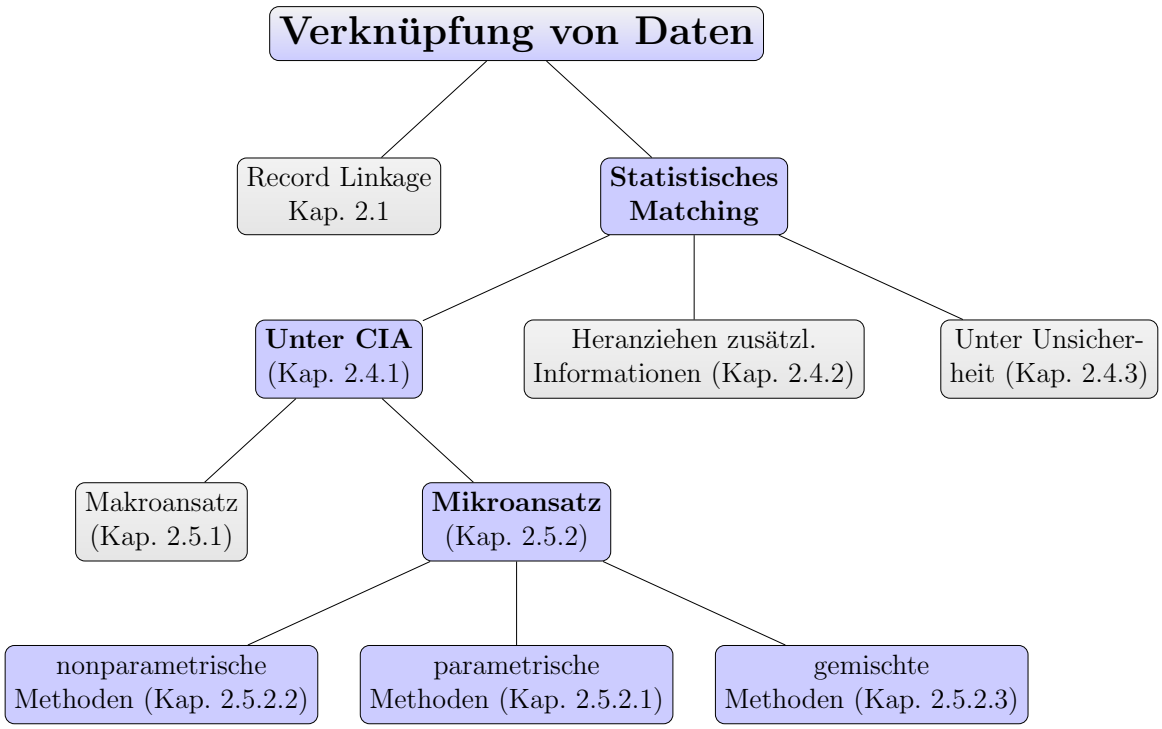

Abbildung 2.5: Überblick über die einzelnen Themen des Statistischen Matchings, welche bis zu diesem Punkt der Arbeit behandelt wurden

Vor allem in der Mikrosimulation werden Datensätze mit einer hohen Zahl an Beobachtungen und Variablen benötigt, welche für gewöhnlich nicht in einem Single-Source-Datensatz erhoben werden können. Um Änderungen in der Steuer- und Transferpolitik zu analysieren, wurden beispielsweise in Kanada mehrere Datenquellen zur „Social Policy Simulation Database“ (SPSD) zusammengefügt (vgl. D'Orazio et al., 2006, S. 173f.).

Ferner entwickelte das deutsche Marktforschungsunternehmen GfK in den 90er Jahren eine Methode zur Fusionierung der Daten aus einem Fernseh- und einem Verbraucherpanel (vgl. Rässler, 2002, S. 48). Auch Eurostat, das statistische Amt der Europäischen Union, forscht seit einigen Jahren ebenfalls auf dem Gebiet des Statistischen Matchings und kann beispielsweise positive Ergebnisse im Bereich der Fusionierung von Daten zur Armutsmessung vorweisen (siehe z.B. Webber und Tonkin, 2013). Auch in der nichtamtlichen Statistik wird das Statistische Matching angewandt: So berichten z. B. Rasner et al. (2011) von positiven Ergebnissen bei Verbindung von Daten aus dem sozio-ökonomischen Panel (SOEP) und der Versicherungskontenstichprobe (VSKT).

Andererseits gibt es viele negative Stimmen, die das Statistical Matching insbesondere unter der bedingten Unabhängigkeitsannahme stark kritisieren. So merkt Rodgers (1984) an, dass während des Matchingprozesses keine Informationen hinzugefügt werden, sondern nur die implizit und explizit getroffenen Annahmen reflektiert werden. Deshalb sollte der Anwender sich bei der Analyse statistisch fusionierter Datensätze darüber bewusst sein, dass diese nur bei korrekten Annahmen gute Schätzer liefern können.

Gerade das Zutreffen der zentralen Annahmen des Statistischen Matchings wird jedoch stark kritisiert. So halten Barr et al. (1982) die Erfüllung der bedingten Unabhängigkeitsannahme in der praktischen Anwendung für fragwürdig.

Auch das Zutreffen der Annahme, dass die beiden Datensätze A und B aus derselben gemeinsamen Verteilung $f_{\mathbf{XYZ}}$ stammen, wird beispielsweise in D'Orazio et al. (2006, S. 4) in Frage gestellt. So können sich die Verteilungen unter Umständen bereits unterscheiden, wenn die beiden Stichproben zu verschiedenen Zeitpunkten erhoben wurden. Im Zuge dessen bemängelt Gabler (1997) die vielen Manipulationsmöglichkeiten im komplexen Fusionierungsprozess.

Die Kritik am Statistischen Matching führt so weit, dass das Verfahren von mehreren kritischen Stimmen als unnötig bezeichnet wird (vgl. Rässler, 2002, S. 4). Judkins (1998) bezeichnet diese Methodik zudem als „imputation out of desperation and against our better judgement“ und auch Fellegi (1977) charakterisiert das Statistische Matching als gefährliches Verfahren und warnt vor einer allzu optimistischen und erwartungsvollen Anwendung statistisch fusionierter Datensätze bzw. Verteilungen.

Kadane (1978) gibt daher die Maxime aus, dass das Wichtigste bei der weiteren Verwendung fusionierter Datensätze nicht die Qualität der Fusionierung an sich, sondern die korrekte Verwendung und Interpretation der aus dem fusionierten Datensatz gewonnenen Analysen ist. So können diese Analysen niemals genauso gehandhabt und interpretiert werden wie Analysen, denen ein vollständig erhobener Datensatz aus einer einzigen Quelle zugrunde liegt.

Zudem ist es nötig, für jedes Matching-Problem die beste Strategie zu wählen, da eine Technik nur die beste für eine bestimmte Datensituation darstellt. Für andere Anwendungsfälle ist jedoch oft eine andere Vorgehensweise zu präferieren (vgl. z. B. Bennike, 1987 oder Moriarity und Scheuren, 2001b).

Es lässt sich also folgern, dass sich das Statistische Matching einerseits starker Kritik ausgesetzt sieht, andererseits trotz dieser Kritik in der Praxis verwendet wird (siehe dazu auch Rässler, 2002, S. 1). Dies wirft die Frage auf, wie gravierend sich die genannten Kritikpunkte auf die Qualität eines fusionierten Datensatzes auswirken. Um diese Frage zu beantworten, sind jedoch zunächst klare Anhaltspunkte zur Bewertung der Qualität eines fusionierten Datensatzes nötig, welche im nächsten Kapitel ausgearbeitet werden.

3 Kriterien zur Bewertung fusionierter Datensätze

Um Aussagen über die Qualität eines fusionierten Datensatzes treffen zu können, muss zunächst ein geeignetes Bewertungskriterium festgelegt werden. Dabei muss eine Entscheidung darüber getroffen werden, welche Eigenschaften eines fusionierten Datensatzes relevant sind und auf welcher Grundlage die Qualität des Datensatzes gemessen werden soll. Im Rahmen dieser Arbeit sollen zusätzlich Erkenntnisse über die Qualität der Analysen auf Basis eines fusionierten Datensatzes getroffen werden.
Generell kann die Qualität eines solchen Datensatzes sowohl von den Eigenschaften bzw. der Performance des Matchingprozesses als auch von der Qualität und der Kompatibilität der zu verbindenden Datensätze A und B abhängen (vgl. D'Orazio, 2009). Demzufolge hat die Wahl von A und B natürlich einen erheblichen Einfluss auf das Ergebnis des Matchingverfahrens. Je höher deren Qualität ist, desto höher ist auch die Qualität des fusionierten Datensatzes. Diese Qualität sinkt, je mehr Fehler, wie beispielsweise Messfehler, A und B beinhalten.
In der wissenschaftlichen Literatur existieren bereits verschiedene solcher Bewertungskriterien, von denen im folgenden Kapitel eine Auswahl vorgestellt wird. Zu Beginn werden frühe Bewertungskriterien für das Statistische Matching und der Matching Noise angesprochen. Den Schwerpunkt des Kapitels bildet allerdings das Bewertungskonzept von Rässler (2002, Kap. 2.5), da dieses die Bewertungsgrundlage für die Simulationsstudie darstellt. Aus diesem Grund werden gleichzeitig Kennzahlen zum Messen dieses Konzepts diskutiert. Dabei werden sowohl in der wissenschaftlichen Literatur verwendete Kennzahlen vorgestellt, als auch die zur Verwendung in der Simulationsstudie gewählten Kennzahlen näher beschrieben.
Generell müssen diese Kennzahlen speziell für die vorliegende Datensituation gewählt werden. Die in diesem Kapitel gewählten Kennzahlen sind daher auf die in der Simulationsstudie vorliegende Datensituation, bei der multivariat normalverteilte Datensätze fusioniert werden, abgestimmt.

3.1 Frühe Bewertungskriterien fusionierter Datensätze

Da es schon seit der ersten Veröffentlichung zum Statistischen Matching zu kontroversen Diskussionen über das Verfahren kam (siehe bspw. Sims, 1972), werden bereits seit den 1980er Jahren Arbeiten veröffentlicht, welche die Repräsentativität statistisch fusionierter Datensätze in Evaluationsstudien beurteilen. In diesen Studien wurden verschiedene Bewertungskriterien für die Qualität eines solchen künstlichen Datensatzes entwickelt und angewandt. Eine Auswahl wird in diesem Abschnitt vorgestellt.
Generell wurde die Qualität eines fusionierten Datensatzes daran gemessen, wie gut die Eigenschaften des Originaldatensatzes nachgebildet werden konnten.
So wurden in den meisten betrachteten Evaluationsstudien die univariaten Verteilungen von Z_r im Spenderdatensatz und im fusionierten Datensatz verglichen (siehe z. B. Rodgers et al., 1981 oder Barry, 1988). Weiterhin wurde der Zusammenhang zwischen $\mathbf{X}$ und $\mathbf{Z}$ sowie zwischen $\mathbf{Y}$ und $\mathbf{Z}$ untersucht (vgl. z. B. Barr et al., 1982).
Barry (1988) erweiterte diese Kriterien zusätzlich um einen Vergleich der Regressionsschätzer, die auf Basis des Original- und des fusionierten Datensatz berechnet wurden. Barr und Turner (1990) bewerten zudem die Unterschiede von $f_{Y_q Z_r}$ im Original- und fusionierten Datensatz.
Ein weiteres Bewertungskonzept bildet auch der Matching Noise, welcher im nächsten Abschnitt näher beschrieben wird.

3.2 Matching Noise

Speziell für endliche Stichprobengrößen muss die Frage aufgeworfen werden, wie gut die Datensätze A und B und somit auch der fusionierte Datensatz die wahre zugrunde liegende Verteilung $f_{\mathbf{XYZ}}$ überhaupt widerspiegeln. Ein Konzept zur Beantwortung dieser Frage bietet der „Matching Noise“(siehe z. B. D'Orazio et al., 2006, Kap. 2.4.4): Dieser misst für einen fusionierten Datensatz A, ob $(\mathbf{X}, \tilde{\mathbf{Z}})$ tatsächlich als Stichprobe aus $f_{\mathbf{XZ}}$ angesehen werden kann. Dies hängt generell von zwei Elementen ab: Zum einen von der Diskrepanz zwischen $\mathbf{x}_a^A$ und dem dazugehörigen Spender $\mathbf{x}_{b*}^B$, weshalb der Matching Noise bei Hot-Deck-Ansätzen generell eine stärkere Rolle spielt. Zum

anderen hat auch die Explanatory Power von $\mathbf{x}_b$ auf $\mathbf{z}_b$ einen Einfluss auf die Höhe des Matching Noise (vgl. Conti et al., 2008). Weisen die Datensätze A und B einen hohen Matching Noise auf, so ist die Annahme, dass beide Datensätze derselben wahren Verteilung zugrunde liegen, verletzt. Folglich kann $f_{\mathbf{XYZ}}$ im fusionierten Datensatz sehr wahrscheinlich nicht korrekt wiedergegeben werden. Die Reduzierung des Matching Noise stellt daher eine fundamentale Aufgabe im Vorfeld einer Fusionierung dar (vgl. D'Orazio et al., 2006, S. 11).
Weitere Untersuchungen zur Thematik des Matching Noise liefern Marella et al. (2008). Während die frühen Bewertungskriterien also den Erfolg eines Matchings anhand des Unterschieds zwischen dem Originaldatensatz und dem fusionierten Datensatz bewerten, berücksichtigt der Matching Noise den Unterschied zur wahren zugrunde liegenden Verteilung (vgl. D'Orazio et al., 2006, S. 10).
So stellt er ein hilfreiches Mittel zur Bewertung der Kompatibilität der Datensätze A und B dar. Dennoch werden wichtige Eigenschaften des fusionierten Datensatzes, wie die Beziehung von $\mathbf{Y}$ und $\mathbf{Z}$, nicht berücksichtigt, weshalb der Matching Noise sich nicht als umfassendes Bewertungskonzept eignet. Ein solches wird im folgenden Kapitel vorgestellt.

3.3 Validitätsstufen von Rässler

In Rässler (2002, Kap. 2.5) wird ein umfassenderer Qualitätsbegriff definiert, der die Qualität des fusionierten Datensatzes ebenfalls anhand seiner Repräsentativität bzgl. des Originaldatensatzes misst. Dieses Konzept zur Messung der Repräsentativität beinhaltet alle bereits genannten Bewertungskriterien und fügt darüber hinaus weitere hinzu.
Die Repräsentativität eines fusionierten Datensatzes wird dabei in vier verschiedenen Validitätsstufen gemessen, welche den jeweiligen Grad der Repräsentativität angeben. Die erste Validitätsstufe ist dabei am schwersten einzuhalten, während die vierte Stufe vermeintlich am leichtesten zu erreichen ist. Je höher allerdings die zugeordnete Validitätsstufe ist, desto weniger Analysemethoden können auf den fusionierten Datensatz angewandt werden.

Diese Definition des Qualitätsbegriffs wird auch in der noch folgenden Simulationsstudie verwendet. Aus diesem Grund werden in diesem Kapitel neben der Definition der einzelnen Stufen gleichzeitig die Kennzahlen, die in der Simulationsstudie zur Messung der einzelnen Validitätsstufen verwendet werden, ausgewählt.

3.3.1 Erste Validitätsstufe: Erhalt der einzelnen Werte

Die erste Validitätsstufe (siehe Rässler, 2002, Kap. 2.5.1) stellt die am schwierigsten zu erreichende Stufe dar. Der fusionierte Datensatz kann ihr zugeordnet werden, wenn die durch das Matching erzeugten Einträge für $\mathbf{Z}$ exakt mit den wahren Werten übereinstimmen. Da in realen Datensituationen die wahren Werte unbekannt sind, kann diese Validitätsstufe nur mithilfe von Simulationsstudien überprüft werden.
Eine totale Übereinstimmung der künstlichen mit den wahren Werten kann jedoch nur erreicht werden, wenn die gemeinsamen Variablen $\mathbf{X}$ bereits jegliche Information über $\mathbf{Z}$ enthalten.
Dieses Szenario gilt für den Datensatz A beispielsweise als erfüllt, wenn sich z_{ar} exakt durch eine Linearkombination $z_{ar} = \alpha + \mathbf{x}_a\boldsymbol{\beta}$ ausdrücken lässt. Im Statistischen Matching kann die Relevanz eines solchen Szenarios jedoch vernachlässigt werden.
Nichtsdestotrotz wird in Rässler (2004) eine Möglichkeit vorgeschlagen, mit der das Zutreffen dieser Validitätsstufe überprüft werden kann: Dabei wird eine Übereinstimmung von $\mathbf{z}_a$ und $\tilde{\mathbf{z}}_a$ als „Hit“ definiert, so dass daraus eine „Hit Rate“ berechnet werden kann.
Da die Wahrscheinlichkeit, einen Wert exakt vorherzusagen, für stetige Verteilungen jedoch gleich null ist, stellt diese Stufe für stetig verteilte Datensätze kein sinnvolles Bewertungskriterium dar.
Für kategoriale Daten kann die Hit Rate hingegen durchaus einen Hinweis auf die Güte des fusionierten Datensatzes geben. Eine hohe Hit Rate stellt jedoch nicht automatisch sicher, dass die wahre gemeinsame Verteilung im künstlichen Datensatz erhalten werden konnte.
Des Weiteren ist eine exakte Imputation der wahren Werte für eine anschließende statistische Analyse meist gar nicht notwendig, da die relevante Information eines Datensatzes

bereits in der gemeinsamen Verteilung $f_{\mathbf{XYZ}}$ enthalten ist. Für eine valide statistische Analyse basierend auf dem fusionierten Datensatz ist also die Übereinstimmung von $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ ausreichend (vgl. Rässler, 2002, Kap 2.5.1). Aus diesem Grund wurde für diese Validitätsstufe keine Messung in der Simulationsstudie vorgenommen.

3.3.2 Zweite Validitätsstufe: Erhalt der gemeinsamen Verteilung

Die zweite Validitätsstufe stellt die wichtigste Valididtitätsstufe dar und testet, ob die gemeinsame Verteilung im fusionierten Datensatz mit der wahren Verteilung $f_{\mathbf{XYZ}}$ übereinstimmt. Sie gilt als erfüllt, wenn

$$\tilde{f}_{\mathbf{XYZ}} = f_{\mathbf{XYZ}}$$

zutrifft. In diesem Fall kann der fusionierte Datensatz, genau wie die Datensätze A und B, als Single-Source-Stichprobe aus $f_{\mathbf{XYZ}}$ angesehen werden. Alle statistischen Analysen, welche mit dem Originaldatensatz durchgeführt werden können, können dann ebenfalls auf Basis des fusionierten Datensatzes berechnet werden.
Das Zutreffen dieser Validitätsstufe muss ebenfalls in einer Simulationsstudie überprüft werden. Generell kann sie nur erfüllt werden, wenn die im Vorfeld des Matchings getroffenen Annahmen, wie beispielsweise die Annahme der bedingten Unabhängigkeit, zutreffen (vgl. Rässler, 2002, Kap.2.5.2).

3.3.2.1 Verwendete Kennzahlen in der Literatur

Da das Testen auf Homogenität zweier multivariater Verteilungen ein relativ komplexes Problem darstellt, konnte in der verwendeten Literatur kein Bewertungskriterium gefunden werden, das den Erhalt von $f_{\mathbf{XYZ}}$ auch für hochdimensionale Datensätze zufriedenstellend überprüft.
Beispielsweise wird in Rässler (2002, S. 153f.) die Verwendung eines „Crosstabulation Fit“ vorgeschlagen. Dabei werden zunächst alle Variablen eines Datensatzes in l Kategorien aufgeteilt. Anschließend werden für den Original- und den fusionierten Datensatz je eine paarweise Kontingenztabelle mit den Variablen Y_q und Z_r bzw. $\tilde{Y}_q$

und $\tilde{Z}_r$ gebildet. So können die Zellhäufigkeiten der beiden Kontingenztabellen anhand eines χ^2-Tests (siehe bspw. Agresti, 2002, Kap. 3.2 und 3.3) auf Homogenität getestet werden. Dieses Verfahren wird ebenfalls in Barr und Turner (1990) verwendet.
Als weiteres Kriterium wird in Rässler (2002, S. 153f.) ein Korrelationskoeffizient zwischen den Zellhäufigkeiten im Original- und im fusionierten Datensatz berechnet. Für stetige Daten besitzen beide Kennzahlen allerdings den Nachteil, dass die Information in den Datensätzen durch das Bilden von Kategorien reduziert wird. Ferner sollte auch erwähnt werden, dass diese Kennzahlen nicht exakt den Erhalt von $f_{\mathbf{XYZ}}$, sondern nur von $f_{Y_q Z_r}$ untersuchen. Aus diesem Grund wurde in der Simulationsstudie ein alternatives Kriterium ausgewählt.

3.3.2.2 Verwendete Kennzahlen in der Simulationsstudie

Als Bewertungskriterium für die zweite Validitätsstufe wurde die Hellinger-Distanz h (siehe z. B. Rüschendorf, 2014, S. 62) gewählt. Diese stellt eine Metrik zwischen zwei Wahrscheinlichkeitsmaßen dar und wird beispielsweise auch in Webber und Tonkin (2013, Kap. 2.4) zum Vergleich zweier Wahrscheinlichkeitsverteilungen verwendet. Die Berechnung erfolgt folgendermaßen:

$$h(f_{\mathbf{XYZ}}, \tilde{f}_{\mathbf{XYZ}}) = \left(1 - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(f_{\mathbf{XYZ}}(\mathbf{x},\mathbf{y},\mathbf{z})\tilde{f}_{\mathbf{XYZ}}(\mathbf{x},\mathbf{y},\mathbf{z})\right)^{\frac{1}{2}} \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{z}\right)^{\frac{1}{2}}$$

Für $f_{\mathbf{XYZ}} = \tilde{f}_{\mathbf{XYZ}}$ gilt

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \underbrace{\left(f_{\mathbf{XYZ}}(\mathbf{x},\mathbf{y},\mathbf{z})\tilde{f}_{\mathbf{XYZ}}(\mathbf{x},\mathbf{y},\mathbf{z})\right)^{\frac{1}{2}}}_{f_{\mathbf{XYZ}}} \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{z} = 1,$$

was einer Hellinger-Distanz von 0 entspricht. Kann hingegen kein Überlappen von $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ beobachtet werden, so nimmt die Hellinger-Distanz einen Wert von 1 an.
Inhaltlich kann die Hellinger-Distanz also eine gute Bewertung der zweiten Validitätsstufe gewährleisten. Aufgrund ihrer Normiertheit auf $h \in [0, 1]$ können die Ergebnisse der Distanz auch gut zum Vergleich von Verteilungen mit unterschiedlichen

Eigenschaften herangezogen werden. Zudem findet kein Informationsverlust durch das Bilden von Kategorien statt.

Allerdings wird eine unterschiedliche Variabilität in den Daten, die beispielsweise in unterschiedlichen Stichprobengrößen begründet ist, nicht berücksichtigt. Zudem gibt es bisher keinen einheitlichen Grenzwert, bis zu dem $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ als gleich bzw. ähnlich definiert werden können. Leulescu und Agafitei (2013, S. 14) geben zwar einen Erfahrungswert von 0.05 an, dieser stellt jedoch kein objektives Kriterium, wie es statistische Tests bieten, dar.

Des Weiteren ist eine Implementierung der Hellinger-Distanz in R nur für (multivariate) Normalverteilungen vorhanden. Die Berechnung der Hellinger-Distanz hängt dann nur noch von $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma_{XYZ}}$ bzw. $\tilde{\boldsymbol{\mu}}$ und $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{XYZ}}$ ab, so dass Abweichungen von der Normalverteilung nicht berücksichtigt werden:

$$h(f_{\mathbf{XYZ}}, \tilde{f}_{\mathbf{XYZ}}) = \left(1 - \frac{\det(\boldsymbol{\Sigma_{XYZ}})^{1/4} \det(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{XYZ}})^{1/4}}{\det(\frac{\boldsymbol{\Sigma_{XYZ}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{XYZ}}}{2})^{1/2}} \cdot \exp\left(-\frac{1}{8}(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})^T \left(\frac{\boldsymbol{\Sigma_{XYZ}} + \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{XYZ}}}{2}\right)^{-1} (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})\right)\right)^{1/2}$$

Folglich muss hierbei die Annahme getroffen werden, dass $\tilde{f}_{\mathbf{XYZ}}$ einer multivariaten Normalverteilung folgt. Die Berechnung in R erfolgt dabei über die Bhattacharyya-Distanz b für multivariat normalverteilte Datensätze im Paket `fpc` (Hennig, 2015), welche zur Hellinger-Distanz in folgender Beziehung steht (siehe z. B. Askinadze, 2015):

$$h(f_{\mathbf{XYZ}}, \tilde{f}_{\mathbf{XYZ}}) = \sqrt{1 - \exp\left(-b(f_{\mathbf{XYZ}}, \tilde{f}_{\mathbf{XYZ}})\right)}$$

An dieser Stelle sei erwähnt, dass zur Beurteilung des Zutreffens der zweiten Validitätsstufe generell auch die Verwendung nonparametrischer Homogenitätstests sinnvoll erscheint. Anhand dieser Tests kann eine objektive Beurteilung über den Erhalt von $f_{\mathbf{XYZ}}$ im fusionierten Datensatz getroffen werden, ohne $f_{\mathbf{XYZ}}$ einer parametrischen Verteilungsfamilie zuzuordnen, worin die Schwachstellen der Hellinger-Distanz liegen. Zudem kann so die durch einen unterschiedlich hohen Stichprobenumfang verursachte Variabilität in den Daten berücksichtigt werden.

Tests dieser Art sind jedoch nur selten in R implementiert und erfordern eine hohe Rechenlaufzeit: So war die Verwendung des multivariaten Cramer-Tests (siehe Baringhaus und Franz, 2004), der im R-Paket `cramer` (Franz, 2014) implementiert ist, aufgrund seiner zu hohen Rechenlaufzeit nicht möglich.
Ebenfalls in der Literatur empfohlen wird der weniger rechenintensive Energy-Test (Székely und Rizzo, 2004). Da dessen Teststatistik jedoch auf der euklidischen Distanz zwischen den einzelnen Beobachtungen basiert, muss diskutiert werden, ob mit diesem Vorgehen nicht eher das Zutreffen der ersten Validitätsstufe bewertet wird.
Die Teststatistik und die bereits berechneten Ergebnisse des Energy-Tests sind im Anhang angegeben. Letztere zeigen zwei weitere Schwachstellen für die in der Simulationsstudie vorliegenden Datensituation auf: So ist zum einen die Testpower für das verwendete Simulationsdesign zu gering, zum anderen unterliegen Datensätze mit einer höheren Anzahl an Beobachtungen einer starken Bestrafung.

3.3.3 Dritte Validitätsstufe: Erhalt der Korrelationsstruktur

Können die ersten beiden Validitätsstufen nicht erreicht werden, so müssen bereits deutliche Abstriche in der Qualität eines fusionierten Datensatzes gemacht werden. Dennoch ist in vielen Situationen eine Analyse der Zusammenhänge zwischen den einzelnen Variablen (wie beispielsweise bei einem linearen Regressionsmodell) ausreichend. Für solche Analysen genügt meist das Zutreffen der dritten Validitätsstufe.
Diese wird erreicht, wenn die wahre Korrelationsstruktur im fusionierten Datensatz erhalten werden kann (vgl. Rässler, 2002, Kap. 2.5.3). Rässler definiert diese Validitätsstufe für

$$\tilde{\mathbf{\Sigma}}_{\mathbf{XYZ}} = \mathbf{\Sigma}_{\mathbf{XYZ}}$$

als erfüllt. Aufgrund der fehlenden gemeinsamen Beobachtungen von $\mathbf{Y}$ und $\mathbf{Z}$ ist vor allem die korrekte Spezifikation von $\tilde{\mathbf{\Sigma}}_{\mathbf{YZ}}$ problematisch. Rässler (2002, S. 23f.) konnte zeigen, dass für $\tilde{\mathbf{\Sigma}}_{\mathbf{YZ}}$ folgende Äquivalenz gilt:

$$\tilde{\mathbf{\Sigma}}_{\mathbf{YZ}} = \mathbf{\Sigma}_{E(\mathbf{Y}|\mathbf{X}),E(\mathbf{Z}|\mathbf{X})}.$$

Da $\mathbf{\Sigma_{YZ}}$ generell mit

$$\mathbf{\Sigma_{YZ}} = E(\mathbf{\Sigma_{YZ|X}}) + \underbrace{\mathbf{\Sigma}_{E(\mathbf{Y}|\mathbf{X}),E(\mathbf{Z}|\mathbf{X})}}_{\tilde{\mathbf{\Sigma}}_{\mathbf{YZ}}}$$

berechnet werden kann (siehe Whittaker, 1990, S. 125), kann die dritte Stufe nur für $E(\mathbf{\Sigma_{YZ|X}}) = 0$ korrekt spezifiziert werden. Dies entspricht dem Fall, dass $\mathbf{Y}$ und $\mathbf{Z}$ im Durchschnitt bedingt unkorreliert gegeben $\mathbf{X}$ sind. Für den Fall der bedingten Unabhängigkeit von $\mathbf{Y}$ und $\mathbf{Z}$ trifft diese Eigenschaft automatisch zu, umgekehrt gilt dies aber nicht (vgl. Rässler, 2002, Kap. 2.5.3 und 2.5.5).
Auch diese Stufe kann ebenfalls nur anhand von Simulationsstudien getestet werden. Zu beachten ist auch, dass bei Matching-Methoden, die auf der bedingten Unabhängigkeitsannahme basieren, stets $\mathbf{\Sigma_{YZ|X}} = 0$ angenommen wird. Dementsprechend spiegelt der fusionierte Datensatz automatisch diese Beziehung wider, auch wenn dies für die wahre Verteilung nicht zutreffend ist.

3.3.3.1 Verwendete Kennzahlen in der Literatur

Die dritte Validitätsstufe wurde bereits in mehreren Evaluationsstudien gemessen. Als Kennzahlen wurden dabei die Abweichungen von $\tilde{\mathbf{\Sigma}}_{\boldsymbol{XY}}$, $\tilde{\mathbf{\Sigma}}_{\boldsymbol{XZ}}$ oder $\tilde{\mathbf{\Sigma}}_{\boldsymbol{YZ}}$ bzw. der entsprechenden Korrelationen zu den Parametern aus dem Orginaldatensatz gemessen. Dazu wurde meist der Bias verwendet (siehe z. B. Barry, 1988 oder Rodgers, 1984). Rässler (2002, S. 151ff.) bewertet die Abweichungen von $\mathbf{\Sigma_{XYZ}}$ anhand einer Teststatistik.
In Rässler (2002, S. 155) und Barry (1988) wurden außerdem die linearen Regressionskoeffizienten aus dem Original- und dem fusionierten Datensatz verglichen. Zudem wird in Rässler (2002, S. 152f.) der korrekte Erhalt von $\boldsymbol{R_{YZ|X}}$ im fusionierten Datensatz untersucht.

3.3.3.2 Verwendete Kennzahlen in der Simulationsstudie

In der Simulationsstudie wurden aufgrund der besseren Vergleichbarkeit die Abweichungen der Korrelationsmatrix als Kennzahl verwendet. Dabei werden Bias und MSE der

einzelnen Elemente von $\mathbf{R_{YZ}}$ und $\boldsymbol{R_{YZ|X}}$ gemessen, da diese die einzigen Parameter der Korrelationsstruktur verkörpern, die nicht aus A oder B gewonnen werden können. Die Beziehung von $(\mathbf{X}, \mathbf{Z})$ wird innerhalb der vierten Validitätsstufe gemessen.
Eine Überprüfung verschiedener Regressionsschätzer wurde nicht durchgeführt, jedoch werden in Kapitel 4.5 die Auswirkungen einer Verletzung der bedingten Unabhängigkeitsannahme auf diese diskutiert.

3.3.4 Vierte Validitätsstufe: Erhalt der marginalen Verteilungen

Die schwächste Validitätstufe ist erreicht, wenn die marginalen und gemeinsamen Verteilungen der Spenderdaten auch im fusionierten Datensatz erhalten werden können. Dies ist erreicht, wenn

$$\tilde{f}_{\mathbf{Z}} = f_{\mathbf{Z}}$$
$$\text{bzw.} \quad \tilde{f}_{\mathbf{XZ}} = f_{\mathbf{XZ}}$$

zutreffen (vgl. Rässler, 2002, Kap. 2.5.4). Obwohl dies die am leichtesten zu erreichende Validitätsstufe darstellt, kann sie v. a. für hochdimensionale Daten und unterschiedliche Stichprobendesigns schwer zu erfüllen sein (vgl. Rässler, 2002, 2.5.5).
Erfüllt der fusionierte Datensatz diese Validitätsstufe, so führen die Analysen von $(\mathbf{X}, \mathbf{Z})$ und $(\mathbf{X}, \tilde{\mathbf{Z}})$ zu den selben Schlüssen (vgl. Rässler, 2002, Kap. 2.5.4.).
Dies ist die einzige Stufe, deren Einhaltung auch in der praktischen Anwendung kontrolliert werden kann. Daher wird das Matchingverfahren oftmals als erfolgreich bezeichnet, wenn die vierte Validitätsstufe approximativ zutrifft (siehe dazu bspw. Koschnick, 1995).
Generell misst der Unterschied zwischen $\tilde{f}_{\mathbf{XZ}}$ und $f_{\mathbf{XZ}}$ auch den Matching Noise.

3.3.4.1 Verwendete Kennzahlen in der Literatur

Obwohl die vierte Validitätsstufe erst nach dem Erscheinen vieler Evaluationsstudien definiert wurde, wurden in fast allen untersuchten Studien Kennzahlen verwendet, die den Erhalt der marginalen Verteilungen f_{Z_r} beurteilen. Meist wurden dazu einfach

Mittelwert und Standardabweichung der Original- und der fusionierten Datensätze verglichen (siehe dazu bspw. Barry, 1988).
In Leulescu und Agafitei (2013, S. 14) wird ebenfalls die Verwendung der Hellinger-Distanz von für die Messung der vierten Stufe diskutiert. Diese eignet sich besonders bei unbekanntem Stichprobendesign der Datensätze A und B.
Allerdings haben die genannten Kennzahlen den Nachteil, dass die Interpretation der Werte subjektiv und ohne theoretische Rechtfertigung erfolgt und die im Stichprobendesign begründete Variabilität nicht berücksichtigt wird (vgl. Leulescu und Agafitei, 2013, S. 14).
In Rässler (2002, S. 150f.) wurden verschiedene univariate Tests verwendet, um zu überprüfen, ob zwei Stichproben dieselbe zugrunde liegende Verteilung besitzen. Dabei wurde allerdings nur der Erhalt von f_{Z_r} überprüft.
$f_{\mathbf{XZ}}$ wurde durch den Vergleich von $\boldsymbol{\Sigma_{XZ}}$ im Original- und fusionierten Datensatz (vgl. z. B. Rodgers, 1984).

3.3.4.2 Verwendete Kennzahlen in der Simulationsstudie

Da das Stichprobendesign in der Simulationsstudie bekannt ist, stellen statistische Tests eine gute Wahl zur Bewertung von f_{Z_r} dar. Diese beziehen auch eine unterschiedlich hohe Variabilität in den Daten mit ein. Als Kennzahl für die vierte Validitätsstufe wird daher die Anzahl abgelehnter Kolmogorov-Smirnov-Tests für f_{Z_r} angegeben. Dabei wird für ein Signifikanzniveau von $\alpha = 0.05$ die Nullhypothese $f_{Z_r} = \tilde{f}_{Z_r}$ gegen $H_1 : f_{Z_r} \neq \tilde{f}_{Z_r}$ für $r = 1, 2, 3$ mithilfe der Teststatistik

$$d_{Z_r} = \sup_{z_r} \|\tilde{F}_{Z_r}(z_r) - F_{Z_r}(z_r)\|$$

bewertet (siehe z.B. Trenkler und Büning, 1994, S. 119-124). Zusätzlich wird jeweils der durchschnittliche Wert der Teststatistik d_{Z_r} angegeben, welche die maximale vertikale Distanz der beiden Verteilungsfunktionen beschreibt. Um auch die Beziehung der Variablen in $\mathbf{Z}$ untereinander miteinzubeziehen, wurden zusätzlich der Bias und der MSE von $\tilde{\mathbf{R}}_{\mathbf{Z}}$ betrachtet.

Da die Daten in der Simulationsstudie multivariat normalverteilt sind, wird der Erhalt von $f_{\mathbf{XZ}}$ bzw. $f_{\mathbf{XY}}$ als bivariate Struktur interpretiert und wie in Kapitel 3.3.3.1 anhand des Bias und des MSE von $\boldsymbol{R_{XZ}}$ bzw. $\boldsymbol{R_{XY}}$ im fusionierten und im Originaldatensatz gemessen. Thematisch wird diese Kennzahl allerdings der vierten Stufe zugeordnet, da hier alle Kennzahlen zusammengefasst sein sollen, welche auch ohne eine Fusion von A und B erhoben und bewertet werden können.

3.4 Bewertung der Kriterien

Bei erneuter Betrachtung der frühen Ansätze zur Bewertung eines fusionierten Datensatzes in Kapitel 2.6 fällt auf, dass die meisten von ihnen Eigenschaften messen, welche der dritten und vierten Validitätsstufe entsprechen. Dadurch wird nochmals verdeutlicht, dass das Konzept der vier Validitätsstufen auf diese früheren Arbeiten aufbaut und diese um wichtige Aspekte ergänzt. So kann eine umfassende Definition der Repräsentativität des fusionierten Datensatzes bereitgestellt werden. Abbildung 3.1 bietet eine Veranschaulichung der einzelnen Validitätsstufen und der gewählten Kennzahlen. An dieser Stelle soll nochmals betont werden, dass die Wahl der Kennzahlen stets speziell auf die vorliegenden Daten und ihre Eigenschaften abgestimmt werden sollte.

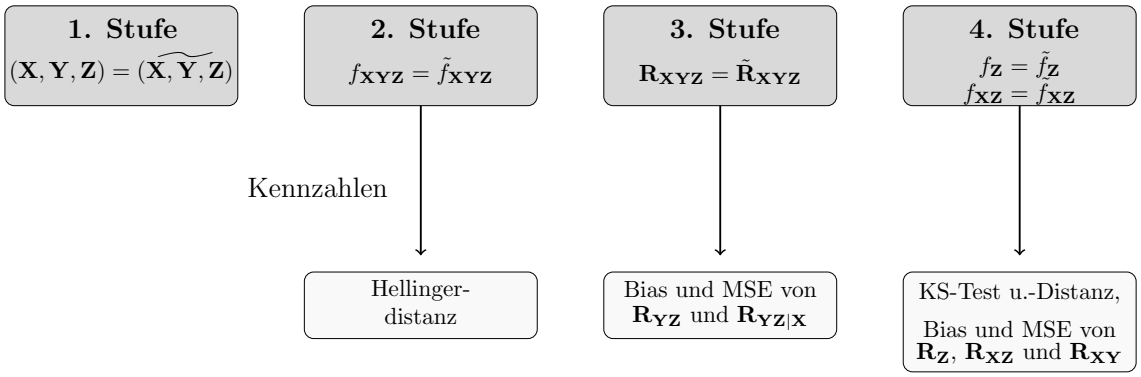

Abbildung 3.1: Veranschaulichung der in der Simulationsstudie verwendeten Kennzahlen

Die Zuordnung zu einer bestimmten Validitätsstufe bestimmt zugleich, welche Analyseverfahren mit dem fusionierten Datensatz korrekt durchgeführt werden können. Auf diesen Umstand wird in Kapitel 4.5 ausführlich eingegangen.
Allerdings weist Rässlers Konzept den Nachteil auf, dass in der praktischen Anwendung meist nur die vierte Stufe untersucht werden kann, da die notwendigen Daten zur Untersuchung der übrigen Stufen nicht vorhanden sind. Daher wird das Zutreffen der vierten Stufe oft bereits als Hinweis für eine erfolgreiche Fusionierung gesehen. Dies sollte kritisch betrachtet werden, da auf diese Weise nur die Kompatibilität der Datensätze A und B untersucht wird und das Erfüllen weiterer wichtiger Eigenschaften durch den fusionierten Datensatzes nicht berücksichtigt wird.
Umgekehrt gilt jedoch: Trifft die vierte Stufe nicht zu, so lässt sich durchaus folgern, dass die Durchführung eines Matchings keine guten Ergebnisse erzielen wird, da die Datensätze beispielsweise aufgrund von Messfehlern in der Erhebung nicht kompatibel sind. Rässler (2002, Kap.2.5.4) definiert dieses Kriterium daher auch als „Minimalanforderung“ an die erfolgreiche Durchführung eines Statistischen Matchings.
Da das Zutreffen der Annahmen des Statistischen Matchings in der praktischen Anwendung also nicht getestet werden kann, ist man in der Praxis auf verschiedene Heuristiken angewiesen, die eine Hilfestellung bei der Beurteilung dieser Annahmen gewährleisten. Anhand dieser kann die Qualität eines fusionierten Datensatzes eingeschätzt werden.
Abschließend muss bei der Bewertung durch die Validitätsstufen berücksichtigt werden, dass so nur bedingt Rückschlüsse über die Qualität der Datensätze A und B möglich sind. Diese sollte im Vorfeld durch verschiedene Diagnoseverfahren sichergestellt werden.

4 Simulationsstudie zur Bewertung der Qualität fusionierter Datensätze

In der praktischen Anwendung des Statistischen Matchings kann die Qualität eines fusionierten Datensatzes nicht ausreichend überprüft werden, da die „wahren“ Beobachtungen in $\mathbf{Z}$ nicht bekannt sind. Daher wird in dieser Arbeit eine Simulationsstudie zur Beurteilung der Qualität durchgeführt.
Wichtig für die praktische Anwendung des Statistischen Matchings sind dabei auch allgemeingültige Aussagen über die Auswirkungen von Annahmeverletzungen, welche sich auf die Qualität des fusionierten Datensatzes bzw. der damit durchgeführten Analysen auswirken. Dabei ist auch zu klären, inwieweit diese Auswirkungen tolerierbar sind.
Wichtig für die praktische Anwendung sind dabei auch allgemeingültige Aussagen über die Auswirkungen der Verletzungen der Annahmen des Statistischen Matchings auf die Qualität des fusionierten Datensatzes bzw. der damit durchgeführten Analysen und inwieweit diese Auswirkungen tolerierbar sind. In dieser Arbeit liegt der Schwerpunkt dabei auf Verletzungen der bedingten Unabhängigkeitsannahme.
Generell können für die Simulation künstlich erzeugte oder real beobachtete Daten verwendet werden. In dieser Arbeit fiel die Wahl auf künstlich erzeugte Daten, da diese den Vorteil besitzen, dass alle Parameter des generierten Datensatzes bestimmt werden können. So kann explizit die bedingte Abhängigkeit $\mathbf{\Sigma_{YZ|X}}$ in den Daten kontrolliert werden und der Einfluss der Verletzung der bedingten Unabhängigkeitsannahme auf die Qualität fusionierter Datensätze untersucht werden.
In einem weiteren Schwerpunkt der Studie soll untersucht werden, ob durch die Vervollständigung von $A \cup B$ ein Datensatz mit höherer Qualität erzeugt werden kann als bei einer ausschließlichen Imputation der fehlenden Werte in A. In diesem Zusammenhang wird auch der Einfluss verschiedener Größenverhältnisse der Stichproben A und B überprüft. Außerdem wird im Zuge der Simulationsstudie die Performance verschiedener in Kapitel 2.4.3 vorgestellter Matching-Methoden bewertet und vergli-

chen. Ein Ziel der Simulationsstudie besteht daher auch darin, herauszufinden, ob die Auswirkungen einer fälschlicherweise angenommenen CIA durch eine passende Wahl dieser Parameter abgemildert werden können.
Als Bewertungskriterien werden die in Kapitel 3.2 festgelegten Kennzahlen verwendet. Um die Auswirkungen der angesprochenen Parameter umfassend bewerten zu können, umfassen $\mathbf{Y}$ und $\mathbf{Z}$ je drei Variablen. Diese sind jeweils unterschiedlich hoch mit den gemeinsamen Variablen $\mathbf{X}$ korreliert, so dass auch die Auswirkungen einer unterschiedlich hohen Explanatory Power von $\mathbf{X}$ in die Bewertung einbezogen werden können. Zudem können so Schwierigkeiten bei der Fusion mehrerer Variablen aufgedeckt werden.
Dieses Kapitel ist folgendermaßen untergliedert: Zunächst wird das Simulationsdesign vorgestellt. Anschließend werden die Ergebnisse der Studie vorgestellt und mit bereits durchgeführten Evaluationsstudien verglichen. Im letzten Abschnitt wird die Analyse mit fusionierten Datensätzen diskutiert.
Alle Berechnungen wurden mit R (R Development Core Team, 2008, Version 3.3.1) durchgeführt. Eine Übersicht aller verwendeten R-Pakete ist in Anhang B zu finden.

4.1 Simulationsdesign

Zunächst wird in diesem Abschnitt der datengenerierende Prozess genau erläutert. Anschließend werden die festen und variierenden Parameter der Simulation vorgestellt. In einem weiteren Abschnitt wird schließlich die Implementierung der verwendeten Matching-Methoden in R behandelt.

4.1.1 Datengenerierender Prozess

Die Daten der Simulationsstudie werden nach dem Prinzip der „folded database“ (siehe Paass, 1986) erzeugt. Dabei werden zunächst die Datensätze A und B generiert, indem jeweils n_A bzw. n_B Beobachtungen unabhängig voneinander aus der gemeinsamen zugrundeliegenden Verteilung $\boldsymbol{f_{XYZ}}$ gezogen werden. Dafür wird das Paket `MASS` (Venables und Ripley, 2002) verwendet.
Um die Ausgangssituation des Statistischen Matchings zu erhalten, wird anschließend $\mathbf{Z}$ in Datensatz A und $\mathbf{Y}$ in Datensatz B gelöscht. Schließlich werden diese Datensätze zum Datensatz $A \cup B$ zusammengefügt (siehe dazu Abbildung 4.1).

	X_1	X_2	X_3	Y_1	Y_2	Y_3	Z_1	Z_2	Z_3
A	x_{11}	x_{12}	x_{13}	y_{11}	y_{12}	y_{13}	z_{11}	z_{12}	z_{13}
	x_{21}	x_{22}	x_{23}	y_{21}	y_{22}	y_{23}	z_{21}	z_{22}	z_{23}
	...	...	...	...	...	...	...	...	...
	x_{n_A1}	x_{n_A2}	x_{n_A3}	y_{n_A1}	y_{n_A2}	y_{n_A3}	z_{n_A1}	z_{n_A2}	z_{n_A3}
B	x_{11}	x_{12}	x_{13}	y_{11}	y_{12}	y_{13}	z_{11}	z_{12}	z_{13}
	x_{21}	x_{22}	x_{23}	y_{21}	y_{22}	y_{23}	z_{21}	z_{22}	z_{23}
	...	...	...	...	...	...	...	...	...
	x_{n_B1}	x_{n_B2}	x_{n_B3}	y_{n_B1}	y_{n_B2}	y_{n_B3}	z_{n_B1}	z_{n_B2}	z_{n_B3}

Abbildung 4.1: Datengenerierender Prozess der Simulationsstudie: A und B wird zunächst ohne fehlende Werte aus $f_{\mathbf{XYZ}}$ generiert. Die grau eingefärbten Bereiche werden anschließend gelöscht, um die Ausgangssituation des Statistischen Matchings zu erhalten.

Mit diesem Aufbau ist gewährleistet, dass den beiden Datensätzen A und B jeweils dieselbe Verteilung zugrunde liegt, was eine zentrale Annahme des Statistischen Matchings darstellt (vgl. D'Orazio et al., 2006, Kap. 1.2).

Man muss sich allerdings darüber bewusst sein, dass so auch der Matching Noise gering gehalten wird. Dies muss bei der Interpretation der Ergebnisse berücksichtigt werden. Aufgrund dieses Vorgehens sind die Werte des Originaldatensatzes $(\widetilde{\mathbf{X}, \mathbf{Y}}, \mathbf{Z})$ vollständig bekannt und können mit den durch das Statistische Matching erzeugten Werten verglichen werden. Die Qualität des fusionierten Datensatzes $(\widetilde{\mathbf{X}, \mathbf{Y}, \mathbf{Z}})$ wird daran gemessen, wie gut die gemeinsame Verteilung $\boldsymbol{f_{XYZ}}$ des simulierten Originaldatensatzes nachgebildet werden kann. Es ist jedoch zu beachten, dass diese für endliche Stichproben in der Regel nicht der wahren zugrundeliegenden Verteilung entspricht. Um diesen Umstand etwas abzumildern, wurden in diesem Simulationsdesign empirische Parameter bei der Generierung der Datensätze festgelegt, so dass sich die Parameter $\boldsymbol{\theta}$ der wahren zugrundeliegenden Verteilung und der Verteilung des Datensatzes nicht unterscheiden. Die genauen Einstellungen und Parameter dieser Simulationsstudie werden im nächsten Abschnitt genauer erläutert.

4.1.2 Feste Parameter

Zur Generierung des Datensatzes $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ wird eine multivariate Normalverteilung gewählt. $\mathbf{X}$,$\mathbf{Y}$ und $\mathbf{Z}$ bestehen dabei jeweils aus drei Zufallsvariablen. Eine Normalverteilung wurde unter anderem deshalb gewählt, da diese besonders günstige Eigenschaften bezüglich der bedingten Unabhängigkeit aufweist. So ist die bedingte Unabhängigkeitsannahme für normalverteilte Daten identisch mit $\mathbf{R_{YZ|X}} = 0$ (siehe z. B. Rässler, 2002, S. 116).

Die Beobachtungen werden aus einer multivariaten Normalverteilung $N(\boldsymbol{\mu}, \boldsymbol{\Sigma_{XYZ}})$ mit

$$\boldsymbol{\mu} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T,$$

$$\boldsymbol{\Sigma_{XYZ}} = \left(\begin{array}{ccc|ccc|ccc} 1.00 & 0.05 & 0.10 & 0.30 & 0.10 & 0.50 & 0.25 & 0.005 & 0.45 \\ 0.05 & 1.00 & 0.15 & 0.20 & 0.07 & 0.35 & 0.15 & 0.05 & 0.30 \\ 0.10 & 0.15 & 1.00 & 0.10 & 0.05 & 0.15 & 0.05 & 0.00 & 0.10 \\ \hline 0.30 & 0.20 & 0.10 & 1.00 & 0.10 & 0.20 & \sigma_{Y_1Z_1} & \sigma_{Y_1Z_2} & \sigma_{Y_1Z_3} \\ 0.10 & 0.07 & 0.05 & 0.10 & 1.00 & 0.05 & \sigma_{Y_1Z_1} & \sigma_{Y_2Z_2} & \sigma_{Y_2Z_3} \\ 0.50 & 0.35 & 0.15 & 0.20 & 0.05 & 1.00 & \sigma_{Y_1Z_1} & \sigma_{Y_3Z_2} & \sigma_{Y_2Z_3} \\ \hline 0.25 & 0.15 & 0.05 & \sigma_{Y_1Z_1} & \sigma_{Y_2Z_1} & \sigma_{Y_3Z_1} & 1.00 & 0.00 & 0.20 \\ 0.005 & 0.05 & 0.00 & \sigma_{Y_1Z_2} & \sigma_{Y_2Z_2} & \sigma_{Y_3Z_2} & 0.00 & 1.00 & 0.10 \\ 0.45 & 0.30 & 0.10 & \sigma_{Y_1Z_3} & \sigma_{Y_2Z_3} & \sigma_{Y_3Z_3} & 0.20 & 0.10 & 1.00 \end{array}\right)$$

gezogen. Da die Werte standardisiert sind, kann die Kovarianzmatrix auch als Korrelationsmatrix interpretiert werden.

Um die Aussage von Rässler (2002, S. 57) zu überprüfen, dass die Erfüllung der zweiten und dritten Validitätsstufe nur für eine hohe Explanatory Power von X_r realistisch ist, wurden für $\mathbf{\Sigma_{XY}}$ und $\mathbf{\Sigma_{XZ}}$ unterschiedlich hohe Korrelationen installiert. Dabei wurden die Korrelationen von Y_q und Z_r mit X_p so bestimmt, dass die Korrelation zu X_1 (bis auf eine Ausnahme) die höchsten und zu X_3 die niedrigsten Werte aufweist. Die Korrelation zu X_2 beträgt einen Wert zwischen diesen. Für $\mathbf{\Sigma_X}$, $\mathbf{\Sigma_Y}$ und $\mathbf{\Sigma_Z}$ wurden die Kovarianzen hingegen eher gering gewählt.

Abbildung 4.2 bietet eine zusätzliche Veranschaulichung der Höhe der Kovarianzen. Für eine bessere Übersichtlichkeit sind $\sigma_{X_pY_q}$ und $\sigma_{X_pZ_r}$ schwarz umrandet.

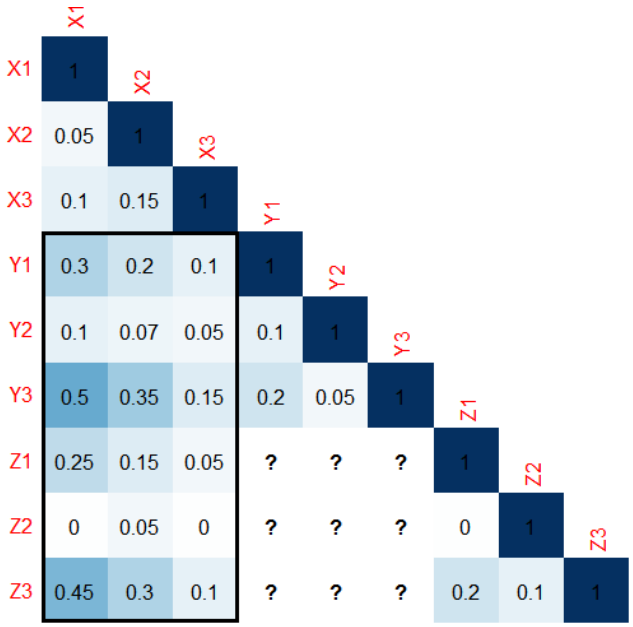

Abbildung 4.2: Veranschaulichung der einzelnen Elemente der in der Simulationsstudie verwendeten Kovarianz-/Korrelationsmatrix: Je dunkler die Farbe eines Feldes ist, desto höher ist die Kovarianz der beiden Variablen. Zur besseren Übersichtlichkeit sind die Kovarianzen von Y_q bzw. Z_r mit den einzelnen X-Variablen schwarz umrandet. Die Werte in den mit Fragezeichen besetzten Feldern werden später anhand der jeweiligen bedingten Korrelation festgelegt. Die Erstellung der Grafik erfolgte mit `corrplot` (Wei und Simko, 2016).

Zusätzlich wurde die Explanatory Power für Y_3 und Z_3 relativ hoch, für Y_2 und Z_2 eher niedrig und für Y_1 und Z_1 mittelhoch festgelegt. Sie wird anhand des Bestimmtheitsmaßes R^2 für das Regressionsmodell Y_q bzw. $Z_r = X_1 + X_2 + X_3$ gemessen (siehe Tabelle 4.1). Aufgrund des Simulationsdesigns weist R^2 für jeden Durchlauf die gleichen Werte auf.

Variable	R^2
Y_1	0.12623
Y_2	0.01520
Y_3	0.35876
Z_1	0.08148
Z_2	0.00257
Z_3	0.27992

Tabelle 4.1: Die Tabelle zeigt das Bestimmtheitsmaß R^2 für die Variablen Y_q und Z_r, welches anhand des Regressionsmodells Y_q bzw. $Z_r = X_1 + X_2 + X_3$ gemessen wurde.

Die Kovarianzen $\sigma_{Y_q Z_R}$ mit $q, r = 1, 2, 3$ (in 4.2 mit Fragezeichen gekennzeichnet) sind noch unbestimmt und werden später anhand der jeweiligen bedingten Korrelation $\mathbf{R_{YZ|X}}$ festgelegt.
Die folgenen Ausführungen beziehen sich auf die variierenden Parameter der Simulationsstudie.

4.1.3 Variierende Parameter

Das Ziel der Simulationsstudie besteht darin, den Einfluss verschiedener Eigenschaften auf die Qualität eines fusionierten Datensatzes zu untersuchen. In dieser Studie werden die Auswirkungen einer Verletzung der bedingten Unabhängigkeitsannahme analysiert. Zudem wird die Performance unterschiedlicher Matching-Methoden verglichen sowie die Ergebnisse verschiedener Matching-Szenarien, die im folgenden Abschnitt noch näher beschrieben werden, untersucht.

4.1.3.1 Bedingte Korrelation $\mathbf{R_{YZ|X}}$

Ein wichtiges Ziel dieser Studie besteht darin, die Auswirkungen einer Verletzung der CIA zu untersuchen. Diese stellt eine wesentliche Annahme beim Statistischen Matching dar und beeinflusst die Qualität eines fusionierten Datensatzes massiv.
Die Tatsache, dass das Zutreffen dieser Annahme in der praktischen Anwendung nicht überprüft werden kann, stellt ein zentrales Problem des Statistischen Matchings dar.
In der Simulationstudie werden daher drei verschiedene Werte (0, 0.2 und 0.35) für $\mathbf{R_{YZ|X}}$ verwendet. Für $\mathbf{R_{YZ|X}} = 0$ ist die bedingte Unabhängigkeitsannahme erfüllt. Die anderen beiden Werte ziehen verschieden starke Abweichungen von dieser Annahme nach sich.
Anhand der Werte für $\mathbf{R_{YZ|X}}$ können anschließend die einzelnen Elemente von $\boldsymbol{\Sigma_{YZ}}$ berechnet werden, wodurch sich drei verschiedene Kovarianzmatrizen ergeben. Dazu wurde folgende Formel (vgl. Rässler, 2002, Kap. 4.8.1) verwendet:

$$\sigma_{Y_q Z_r} = \sigma_{Y_q Z_r|\mathbf{X}} + \boldsymbol{\Sigma}_{Y_q \mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-\mathbf{1}} \boldsymbol{\Sigma}_{Z_r \mathbf{X}}^{T} \tag{4.1}$$

mit

$$\sigma_{Y_q Z_r|\mathbf{X}} = r_{Y_q Z_r|\mathbf{X}} \sqrt{\sigma^2_{Y_q|\mathbf{X}} \sigma^2_{Z_r|\mathbf{X}}} \tag{4.2}$$

$$\sigma^2_{Y_q|\mathbf{X}} = \sigma^2_{Y_q} - \boldsymbol{\Sigma}_{Y_q \mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-\mathbf{1}} \boldsymbol{\Sigma}_{Y_q \mathbf{X}}^{T} \tag{4.3}$$

$$\sigma^2_{Z_r|\mathbf{X}} = \sigma^2_{Z_r} - \boldsymbol{\Sigma}_{Z_r \mathbf{X}} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-\mathbf{1}} \boldsymbol{\Sigma}_{Z_r \mathbf{X}}^{T} \tag{4.4}$$

Während die Gleichungen 4.3 und 4.4 nur von den festen Parametern in Kapitel 4.1.2 abhängen, wird die Gleichung 4.2 auch vom variierenden Parameter $r_{Y_q Z_r|\mathbf{X}}$ bestimmt. Tabelle 4.2 zeigt die so berechneten Werte für die verwendeten bedingten Korrelationen:

	$r_{Y_q Z_r\|\mathbf{X}} = 0$	$r_{Y_q Z_r\|\mathbf{X}} = 0.2$	$r_{Y_q Z_r\|\mathbf{X}} = 0.35$
$\sigma_{Y_1 Z_1}$	0.10072	0.27990	0.41428
$\sigma_{Y_1 Z_2}$	0.01038	0.19709	0.33713
$\sigma_{Y_1 Z_3}$	0.18712	0.34576	0.46474
$\sigma_{Y_2 Z_1}$	0.03412	0.22433	0.36700
$\sigma_{Y_2 Z_2}$	0.00350	0.20172	0.35038
$\sigma_{Y_2 Z_3}$	0.06355	0.23197	0.35828
$\sigma_{Y_3 Z_1}$	0.17007	0.32356	0.43868
$\sigma_{Y_3 Z_2}$	0.01829	0.17824	0.29820
$\sigma_{Y_3 Z_3}$	0.31621	0.45212	0.55404

Tabelle 4.2: Werte von $\sigma_{Y_q Z_r}$ für die verschiedenen, in der Simulationsstudie verwendeten, bedingten Korrelationen $\mathbf{R_{YZ|X}}$

Diese Konstruktionsmethode für $\boldsymbol{\Sigma_{XYZ}}$ weist jedoch einen Mangel auf: Die anhand der Formel 4.1 berechneten Werte für $\sigma_{Y_q Z_r}$ bilden nicht automatisch eine positiv semidefinite Matrix $\boldsymbol{\Sigma_{XYZ}}$. Dies muss zusätzlich für jedes $\mathbf{R_{YZ|X}}$ überprüft werden. Abbildung 4.3 veranschaulicht die Werte aus Tabelle 4.2 grafisch. Für jedes $\sigma_{Y_q Z_r}$ ist dabei zusätzlich der von Gleichung 4.1. bestimmte Wertebereich in blau angegeben. Bei der Interpretation dieses Wertebereichs muss jedoch bedacht werden, dass in diesem weitere Eigenschaften von $\boldsymbol{\Sigma_{XYZ}}$ nicht berücksichtigt werden. So sind Beschränkungen, die sich aufgrund der positiven Semidefinitheit der Kovarianzmatrix $\boldsymbol{\Sigma_{XYZ}}$ ergeben, nicht eingezeichnet.

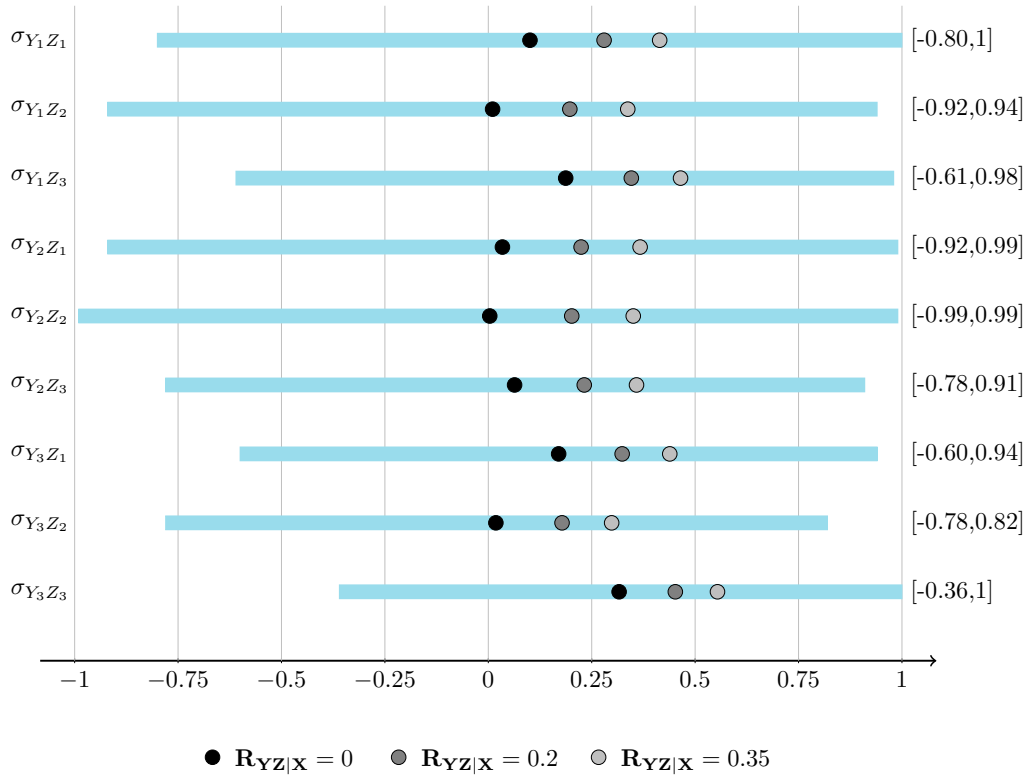

Abbildung 4.3: In dieser Grafik sind die einzelnen Werte für $\sigma_{Y_qZ_r}$ für die verschiedenen bedingten Korrelationen aufgezeigt. Der von $r_{Y_qZ_r|\mathbf{X}} \in [-1, 1]$ begrenzte Wertebereich ist ebenfalls in blau eingezeichnet.

4.1.3.2 Matching-Methoden

Die Qualität der fusionierten Datensätze wird auch im Hinblick auf die Verwendung verschiedener Matching-Methoden untersucht. So können eventuell Rückschlüsse darauf gezogen werden, welche Matching-Methode generell oder für bestimmte Datensituationen die besten Ergebnisse erzielt.

Aus jeder der in Kapitel 2.5.2 vorgestellten Methodengruppen des Mikroansatzes wird eine Methode ausgewählt, so dass in der Simulationsstudie je eine gemischte, eine nonparametrische und eine parametrische Methode Anwendung finden.

Als nonparametrische Matching-Methode wurde die Distance-Hot-Deck-Methode (S. 28) unter Verwendung der euklidischen Distanz ausgewählt, während die Stochastische Regressionsimputation (siehe S. 24) als parametrische Methode verwendet wird. Bei der gemischten Methode wird ein Störterm im Regressionsmodell hinzugefügt und der Abstand $d(\tilde{\mathbf{z}}_a, \mathbf{z}_b)$ ebenfalls anhand der euklidischen Distanz minimiert (siehe Methode MM4, D'Orazio et al., 2006, S. 49).

4.1.3.3 Matching-Szenarien

Zusätzlich werden in dieser Simulationsstudie verschiedene Matching-Szenarien verglichen. Dadurch soll herausgefunden werden, ob die Qualität des fusionierten Datensatzes erhöht werden kann, wenn der komplette Datensatz $A \cup B$ statt nur A oder B vervollständigt wird. Dies ist äquivalent zu der Frage, ob bessere Resultate erzielt werden können, wenn ein Datensatz gleichzeitig Spender- und Empfängerdatensatz ist oder nur eine dieser beiden Rollen einnimmt. Zum anderen sollen zwei verschiedene Größenverhältnisse von A und B überprüft werden.
Dazu werden drei verschiedene Szenarien getestet:

- $\underline{A \cup B_{500}}$: Imputation aller fehlenden Werte in $A \cup B$,
  $n_A = n_B = 500$
- $\underline{A_{500}}$: Imputation der fehlenden Werte in A,
  $n_A = n_B = 500$
- $\underline{A_{300}}$: Imputation der fehlenden Werte in A,
  $n_A = 300, n_B = 700$

Abbildung 4.4 veranschaulicht diese Szenarien grafisch. Dabei symbolisiert rot, dass dieser Datensatz vervollständigt wurde. Dieser Teil wird in der Simulationsstudie auf seine Repräsentativität geprüft. Grau eingefärbte Datensätze werden nur als Spenderdatensatz verwendet.

Während das Verhältnis von n_A und n_B folglich variiert, umfasst $n = n_A + n_B$ immer 1000 Beobachtungen. Diese Anzahl wurde gewählt, da sie einer realistischen Größenordnung entspricht und gleichzeitig eine gute Grundlage für verschiedene statistische Tests und Prozeduren bietet. Zudem ist die Rechenlaufzeit vergleichsweise gering. Prinzipiell sollte aber beachtet werden, dass $\tilde{f}_{\mathbf{XYZ}}$ für eine endliche Anzahl an Beobachtungen in A und B stets Unterschiede zum wahren zugrunde liegenden Modell aufweisen wird (vgl. D'Orazio et al., 2006, Kap. 2).
Die untersuchten Szenarien erheben dabei nicht den Anspruch umfassend zu sein, sondern sollen vielmehr als Orientierung für noch folgende Simulationsstudien dienen.

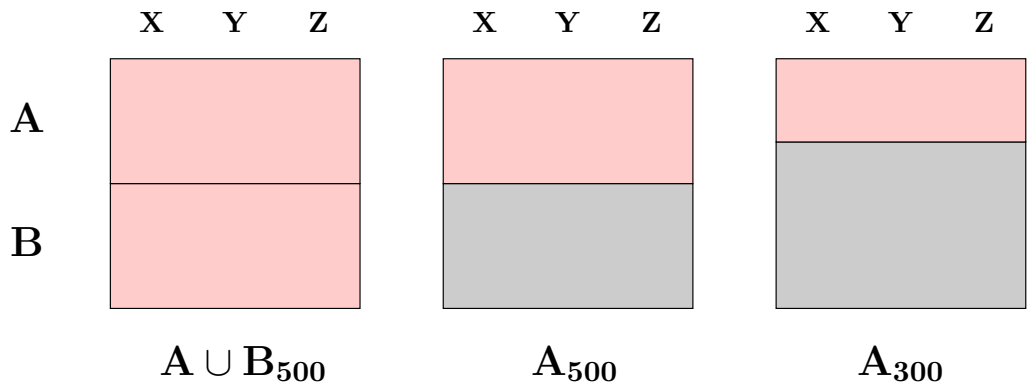

Abbildung 4.4: Darstellung der Matching-Szenarien. Rote Bereiche wurden vervollständigt und im Rahmen der Simulatsionsstudie analysiert. Graue Bereiche werden nur als Spenderdatensatz benutzt.

So besteht der Vorteil des Szenarios $A \cup B_{500}$ in der idealen Ausschöpfung aller vorhandenen Informationen (vgl. D'Orazio et al., 2006, S. 36). Da in diesem Fall auch $\mathbf{Y}$ in B imputiert wird, wird in der Simulationsstudie für die vierte Validitätsstufe auch der Erhalt von $\mathbf{R_{XY}}$ sowie von f_{Y_q} untersucht.
In D'Orazio et al. (2006, S. 36) wird jedoch angemerkt, dass sich bei einer vollständigen Imputation von $A \cup B$ auch der Matching Noise verstärken könnte, besonders wenn sich die Größen von n_A und n_B stark unterscheiden. Zum Vergleich wurden daher in Szenario A_{500} dieselben Datensätze verwendet, jedoch nur der Datensatz A vervollständigt. Aus Rässler und Fleischer (1998) geht hervor, dass der Spenderdatensatz für die Distance-Hot-Deck-Methode mindestens doppelt so viele Beobachtungen wie der Empfängerdatensatz aufweisen sollte. Im Szenario A_{300} wurde daher ein Verhältnis von 3:7 implementiert und dem Szenario A_{500} mit einem Verhältnis von 1:1 gegenüber gestellt.

Mit diesen vorgestellten variierenden Parametern stehen somit 27 Kombinationen der variierenden Parameter zur Verfügung. Die Zahl der Iterationen eines Simulationsdurchlaufs wurde mit k=500 festgesetzt. Einen grafischen Überblick liefert auch Abbildung 4.5. Weitere Parameter, die im Rahmen dieser Arbeit nicht behandelt werden konnten, werden im Ausblick aufgeführt.

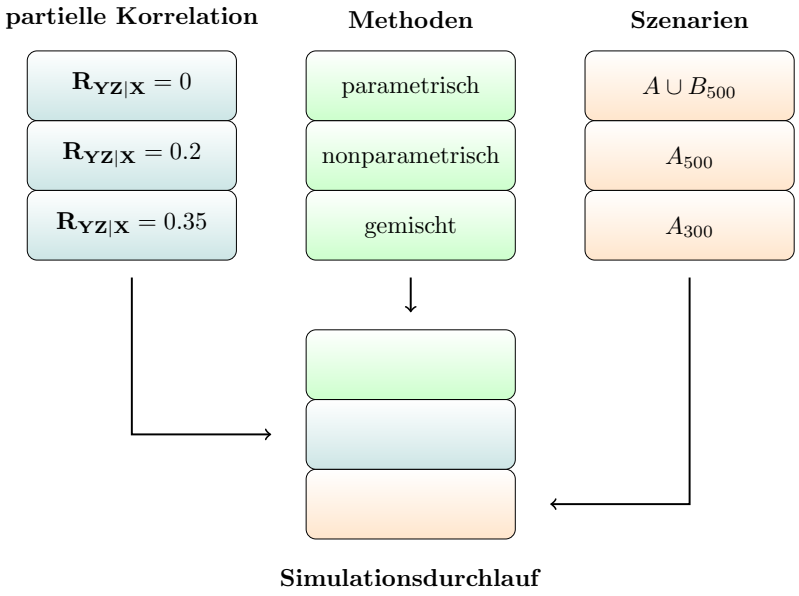

Abbildung 4.5: Überblick über die verschiedenen variierenden Parameter der Simulation

4.2 Implementierung der verwendeten Methoden in R

Wie bereits erwähnt, finden sich Implementierungen der meisten in D'Orazio et al. (2006) erwähnten Matching-Methoden für stetige Daten im R-Paket `StatMatch` (D'Orazio, 2016). Zur Berechnung der beschränkten Distance-Hot-Deck-Methode konnte daher die im Paket enthaltene Funktion NND.hotdeck(·) verwendet werden.

Für die Zuordnung der einzelnen Spender und Empfänger wird dabei die Ungarische Methode (Kuhn, 1955) verwendet. Mit dieser kann eine eineindeutige Zuordnung von Objekten zweier Gruppen mit verringerter Rechenlaufzeit durchgeführt werden, so dass $\sum_{a=1}^{n_A} d(\mathbf{x}_a, \mathbf{x}_{b*})$ minimal ist. Um eine Entscheidung über die Wahl der Distanzfunktion zu treffen, wurde eine kleine Simulationsstudie durchgeführt und anhand der Ergebnisse die euklidische Distanz gewählt (siehe Anhang 5.2).

Für die beiden anderen Methoden war die Verwendung einer Funktion aus dem `StatMatch`-Paket nicht möglich, da diese nur eine Fusionierung für eindimensionale Zufallsvariablen Y und Z durchführen konnten. Daher wurden für diese beiden Methoden eigene Funktionen geschrieben:

Die Stochastische Regressionsimputation wurde anhand folgender multivariater Regressionsgleichung durchgeführt:

$$\hat{z}_{ar} = \hat{\alpha}_r + \mathbf{x}_a\hat{\boldsymbol{\beta}}_r + \hat{\epsilon}_{ar},$$

Dabei gilt $\hat{\boldsymbol{\epsilon}}_a = (\hat{\epsilon}_{a1}, \hat{\epsilon}_{a2}, \hat{\epsilon}_{a3}) \sim N\left((0,0,0)^T, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z|X}}\right)$. $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z|X}}$ kann anhand folgender Formel (vgl. Rässler, 2002, S. 116) berechnet werden:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z|X}} = \boldsymbol{\Sigma}_{\boldsymbol{Z}} - \boldsymbol{\Sigma}_{\boldsymbol{XZ}}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{XZ}}$$

Von den untersuchten Methoden stellt die parametrische Methode die am wenigsten rechenintensive Methode dar.
Für die gemischte Methode wurde zur Berechnung der Hilfsvariablen ebenfalls das oben beschriebene Regressionsmodell verwendet. Zur Identifikation der statistischen Zwillinge wurde wieder die Ungarische Methode verwendet, welche im Paket `clue` (Hornik, 2016) implementiert ist. Dabei wurde $\sum d(\hat{\mathbf{z}}_a, \mathbf{z}_b)$ minimiert. Auch hierfür wurde eine kleine Simulationsstudie zur Wahl der Distanzfunktion durchgeführt und anschließend die euklidische Distanz gewählt, siehe Anhang A.2.

4.3 Ergebnisse der Simulationsstudie

In diesem Kapitel werden die Ergebnisse der in Kapitel 3.4 beschriebenen Simulationsstudie aufgeführt und diskutiert. Das Kapitel ist dabei nach den einzelnen Validitätsstufen gegliedert. Tabellen mit den ausführlichen Ergebnissen finden sich in Anhang C.

4.3.1 Kennzahlen der vierten Validitätsstufe

Die Güte der vierten Validitätsstufe wurde anhand der Anzahl abgelehnter Kolmogorov-Smirnov-Tests sowie der durchschnittlichen Kolmogorov-Smirnov-Distanz für f_{Z_r} bewertet. Um die Beziehung der Variablen in $\mathbf{Z}$ untereinander miteinzubeziehen wurde zusätzlich $\tilde{\mathbf{R}}_{\mathbf{Z}}$ erhoben. Zur Bewertung des Erhalts von $f_{\mathbf{XZ}}$ wurden die Abweichungen von $\mathbf{R}_{\mathbf{XZ}}$ im fusionierten Datensatz erhoben.

Für das Szenario $A \cup B_{500}$ wurden jeweils auch der KS-Test und die KS-Distanz für f_{Y_q} sowie die Abweichungen zu $\mathbf{R_{XY}}$ berechnet; die Ergebnisse dazu werden jedoch ausschließlich in Anhang C.1 aufgeführt.

Generell konnte für keine der Kennzahlen der vierten Validitätsstufe ein bedeutender Unterschied zwischen den Ergebnissen der verschiedenen bedingten Korrelationen festgestellt werden, weshalb die in diesem Abschnitt dargestellten Ergebnisse über alle $\mathbf{R_{YZ|X}}$ aggregiert wurden. Die ursprünglichen Ergebnisse können ebenfalls in Anhang C.1 eingesehen werden.

Anhand des Kolmogorov-Smirnov-Tests bzw. der KS-Distanz kann prinzipiell nur der Erhalt von f_{Z_r} beurteilt werden kann. Für die gemischte und die nonparametrische Methode, bei welchen der Datensatz B als Spenderdatensatz fungiert, ergibt sich jedoch ein Sonderfall für die Szenarien A_{500} und $A \cup B_{500}$: Da in dieser Situation der gesamte Vektor $\mathbf{z}_a$ „gespendet“ wird und die Methode des unbeschränkten Matchings verwendet wird, gilt $f_{Z_r} = \tilde{f}_{Z_r} \leftrightarrow f_{\mathbf{Z}} = \tilde{f}_{\mathbf{Z}}$. Diese Äquivalenz trifft jedoch für die restlichen Simulationsdurchläufe nicht zu, weshalb hierfür zusätzlich Bias und MSE von $\tilde{\mathbf{R}}_{\mathbf{Z}}$ untersucht wurden. Tabelle 4.3 zeigt die Ergebnisse in aggregierter Form:

Szenario	Methode	$r_{Z_1Z_2}$		$r_{Z_1Z_3}$		$r_{Z_2Z_3}$	
		Bias	MSE	Bias	MSE	Bias	MSE
A_{300}	gemischt	-0.00080	0.00291	0.01367	0.00260	0.00536	0.00315
	nonparam.	-0.00108	0.00201	-0.00870	0.00180	-0.00085	0.00199
	parametr.	-0.00010	0.00339	0.00290	0.00286	-0.00133	0.00368
A_{500}	parametr.	-0.00007	0.00198	-0.00007	0.00179	-0.00048	0.00196
$A \cup B_{500}$	parametr.	0.00011	0.00049	-0.00007	0.00044	-0.00025	0.00049

Tabelle 4.3: Bias und MSE von $\tilde{\mathbf{R}}_{\mathbf{Z}}$ aggregiert über alle partiellen Korrelationen $\mathbf{R_{YZ|X}}$

Die Resultate lassen keine nennenswerten Verzerrungen von $\tilde{\mathbf{R}}_{\mathbf{Z}}$ erkennen, weshalb davon ausgegangen werden kann, dass die Beziehung zwischen den einzelnen Variablen in $\mathbf{Z}$ für alle Simulationsdurchläufe erhalten werden konnte.

Im folgenden sind die Ergebnisse zur Bewertung von $\tilde{f}_{Z_r}$ in Tabelle 4.4 veranschaulicht:

		KS-Test				KS-Distanz			
Szenario	Methode	Z_1	Z_2	Z_3	$\sum$	Z_1	Z_2	Z_3	$\varnothing$
A_{300}	gemischt	16	8	6	**30**	0.06056	0.06093	0.05832	**0.05994**
	nonparam.	3	5	2	**10**	0.05774	0.05753	0.05698	**0.05742**
	parametr.	16	14	11	**41**	0.06146	0.06183	0.05943	**0.06090**
A_{500}	gemischt	1	0	0	**1**	0.03961	0.03880	0.03928	**0.03923**
	nonparam.	1	0	0	**1**	0.03961	0.03880	0.03928	**0.03923**
	parametr.	14	5	11	**30**	0.04707	0.04771	0.04681	**0.04720**
$A \cup B_{500}$	gemischt	0	0	0	**0**	0.01981	0.01940	0.01964	**0.01961**
	nonparam.	0	0	0	**0**	0.01981	0.01940	0.01964	**0.01961**
	parametr.	0	0	0	**0**	0.02354	0.02386	0.02340	**0.02360**

Tabelle 4.4: Anzahl abgelehnter KS-Tests und die durchschnittliche Teststatistik für $\tilde{f}_{Z_r}$ aggregiert über die verschiedenen bedingten Korrelationen: Jede weiß hinterlegte Zelle enthält so das Ergebnis von k=1500 Simulationsdurchläufen. In grau sind dabei die Summe aller abgelehnter KS-Tests und die durchschnittliche Teststatistik über alle Z_r angegeben.

Aus den aufgeführten Ergebnissen lässt sich schließen, dass auch f_{Z_r} im fusionierten Datensatz für alle Simulationsdurchläufe gut nachgebildet werden konnte. Dennoch weisen die einzelnen Szenarien und Methoden leichte Unterschiede auf, welche im Folgenden näher beschrieben werden. Dabei werden die aggregierten Werte über Z_r betrachtet, welche in der Tabelle grau hinterlegt sind:

- **Vergleich der Matching-Szenarien:**

  Die durchschnittliche Kolmogorov-Smirnov-Distanz ist hier für $A \cup B_{500}$ am geringsten und etwa halb so groß wie für A_{500}. Dies wird durch die Resultate für f_{Y_q} in Tabelle C.1.2 bestätigt und kann mit der zusätzlich verwendeten Information begründet werden.

  Die höchsten durchschnittlichen KS-Distanzen mit Werten zwischen 0.05698 und 0.06183 können für A_{300} beobachtet werden. Auch die Anzahl abgelehnter Tests fällt für dieses Szenario am höchsten aus, weist aber mit maximal einem Prozent abgelehnter Tests immer noch sehr gute Ergebnisse auf.

- **Vergleich der Matching-Methoden:**

  Der Vergleich der verschiedenen Methoden zeigt, dass die durchschnittliche Kolmogorov-Smirnov-Distanz und die Anzahl abgelehnter Tests bei der gemischten und der nonparametrischen Methode für $A \cup B_{500}$ und A_{500} stets dieselben Werte aufweisen.

Dies ist im Studiendesign begründet. Angesichts der Tatsache, dass A und B für diese Szenarien dieselbe Anzahl an Beobachtungen aufweisen und eine beschränkte Matching-Methode verwendet wird, wird jede Beobachtung $\mathbf{z}_b$ genau einmal an A gespendet. Somit gilt $f_{Z_r}(\mathbf{z}_b) = \tilde{f}_{Z_r}(\mathbf{z}_a)$. Da $f_{Z_r}(\mathbf{z}_a)$ und $f_{Z_r}(\mathbf{z}_b)$ die gleiche zugrundeliegende Verteilung besitzen, kann in diesen Fällen eine Ablehnung des Tests als stochastischer Fehler klassifiziert werden.

Demzufolge ist das Szenario A_{300} in dieser Studie das einzige Szenario, für das die Unterschiede der beiden Methoden sinnvoll bewertet werden können. Dass die gemischte Methode hier eine höhere Zahl abgelehnter Tests bzw. eine höhere durchschnittliche KS-Distanz aufweist, lässt darauf schließen, dass f_{Z_r} für diese Datensituation am besten mit der nonparametrischen Methode erhalten werden kann.

Die parametrische Methode scheint im Vergleich stets schlechtere Ergebnisse zu erzielen. Es ist jedoch zu vermuten, das die parametrische sowie auch die gemischte Methode bessere Resultate aufweisen würden, wenn die Explanatory Power der erklärenden Variablen höher wäre.

Abschließend wird auf die Ergebnisse für den Erhalt von $f_{\mathbf{XZ}}$ eingegangen, welcher anhand des Bias und des MSE von $\tilde{\boldsymbol{R}}_{\mathbf{XZ}}$ gemessen wurde. Aufgrund fehlender Unterschiede wurde diese Kennzahl zusätzlich über die einzelnen Elemente $r_{X_pZ_r}$ aggregiert:

Szenario	Methode	Bias	MSE
A_{300}	gemischt	-0.00133	0.00283
	nonparametrisch	-0.00700	0.00188
	parametrisch	-0.00500	0.00283
A_{500}	gemischt	-0.00526	0.00163
	nonparametrisch	-0.00475	0.00019
	parametrisch	-0.00086	0.00166
$A \cup B_{500}$	gemischt	-0.00263	0.00041
	nonparametrisch	-0.00238	0.00005
	parametrisch	-0.00046	0.00041

Tabelle 4.5: Die Tabelle zeigt Bias und MSE von $\frac{1}{9}\sum_{p=1}^{3}\sum_{r=1}^{3}\tilde{r}_{X_pZ_r}$ aggregiert für alle untersuchten bedingten Korrelationen.

Tabelle 4.5 dokumentiert, dass $\mathbf{R_{XZ}}$ aus dem fusionierten Datensatz für alle Szenarien und Methoden etwa gleich gut geschätzt werden kann. Dies gilt auch für $\mathbf{R_{XY}}$, welches nur für das Szenario $A \cup B_{500}$ erhoben wurde (siehe Tabelle C.1.2 im Anhang).
Insgesamt kann die vierte Validitätsstufe also für alle Simulationsdurchläufe als erfüllt angesehen werden.

4.3.2 Kennzahlen der dritten Validitätsstufe

Für die dritte Validitätsstufe wurde jeweils der Erhalt von $\mathbf{R_{YZ}}$ und $\mathbf{R_{YZ|X}}$ im fusionierten Datensatz gemessen: Dabei werden zunächst die Kennzahlen für die Güte von $\tilde{\mathbf{R}}_{\mathbf{YZ}}$ evaluiert. Exemplarisch werden dazu $r_{Y_3Z_3}$ und $r_{Y_2Z_2}$ dargestellt, da diese die höchste bzw. niedrigste Explanatory Power von $\mathbf{X}$ aufweisen. Die übrigen Elemente von $\mathbf{R_{YZ}}$ sind im Anhang C.2 aufgelistet.

Szenario	Methode	r_{Y2Z2}		r_{Y3Z3}	
		Bias	MSE	Bias	MSE
A_{300}	gemischt	0.00028	0.00322	-0.00405	0.00202
	nonparametrisch	0.00113	0.00301	-0.01615	0.00239
	parametrisch	0.00049	0.00331	0.00127	0.00201
A_{500}	gemischt	-0.00103	0.00189	-0.01145	0.00146
	nonparametrisch	0.00073	0.00224	-0.00975	0.00107
	parametrisch	-0.00102	0.00188	-0.00130	0.00128
$A \cup B_{500}$	gemischt	-0.00163	0.00095	-0.01111	0.00079
	nonparametrisch	0.00073	0.00224	-0.00975	0.00107
	parametrisch	-0.00108	0.00094	-0.00081	0.00060

Tabelle 4.6: Bias und MSE $\tilde{r}_{Y_2Z_2}$ und $\tilde{r}_{Y_3Z_3}$ für $\mathbf{R_{YZ|X}} = 0$

Trifft die bedingte Unabhängigkeitsannahme zu, so können $r_{Y_3Z_3}$ und $r_{Y_2Z_2}$ für alle Methoden und Szenarien korrekt geschätzt werden (siehe Tabelle 4.6). Dies gilt gleichermaßen für die übrigen im Anhang aufgeführten Elemente von $\mathbf{R_{YZ}}$ sowie alle Szenarien und Matching-Methoden. Für die betrachtete Datensituation sind sowohl die Höhe der Explanatory Power als auch das Verhältnis von Spenderdatensatz und Empfängerdatensatz unerheblich für die korrekte Schätzung von $\mathbf{R_{YZ}}$, sofern die CIA eingehalten wurde. Allerdings muss auch betont werden, dass das Zutreffen der CIA für reale Daten bei einer geringen Explanatory Power eher unwahrscheinlich ist.

Im Gegensatz dazu kann für eine bedingte Korrelation von 0.2 der Erhalt von $\mathbf{\Sigma_{YZ}}$ nicht mehr als gewährleistet angesehen werden (siehe Tabelle 4.7).

Szenario	Methode	r_{Y2Z2}		r_{Y3Z3}	
		Bias	MSE	Bias	MSE
A_{300}	gemischt	-0.19682	0.04251	-0.14070	0.02177
	nonparametrisch	-0.19779	0.04249	-0.15124	0.02482
	parametrisch	-0.19645	0.04234	-0.13603	0.02048
A_{500}	gemischt	-0.19908	0.04139	-0.15026	0.02396
	nonparametrisch	-0.20059	0.04225	-0.14680	0.02242
	parametrisch	-0.19846	0.04115	-0.13919	0.02066
$A \cup B_{500}$	gemischt	-0.19809	0.04017	-0.14760	0.02246
	nonparametrisch	-0.20059	0.04225	-0.14680	0.02242
	parametrisch	-0.19766	0.04000	-0.13712	0.01943

Tabelle 4.7: Bias und MSE $\tilde{r}_{Y_2Z_2}$ und $\tilde{r}_{Y_3Z_3}$ für $\mathbf{R_{YZ|X}} = 0.2$

Für alle Simulationsdurchläufe können Unterschätzungen der Korrelationen von $\mathbf{Y}$ und $\mathbf{Z}$ zwischen -0.13603 und -0.20059 beobachtet werden.
Für eine bedingte Korrelation von 0.35 (siehe Tabelle 4.8) fallen diese Verzerrungen mit Werten zwischen -0.23350 und -0.34651 noch drastischer aus:

Szenario	Methode	r_{Y2Z2}		r_{Y3Z3}	
		Bias	MSE	Bias	MSE
A_{300}	gemischt	-0.34651	0.12339	-0.23922	0.05942
	nonparametrisch	-0.34324	0.12079	-0.25345	0.06637
	parametrisch	-0.34621	0.12324	-0.23350	0.05664
A_{500}	gemischt	-0.34603	0.12156	-0.25110	0.06428
	nonparametrisch	-0.34524	0.12118	-0.25023	0.06365
	parametrisch	-0.34643	0.12183	-0.24025	0.05892
$A \cup B_{500}$	gemischt	-0.34635	0.12092	-0.24938	0.06278
	nonparametrisch	-0.34524	0.12118	-0.25023	0.06365
	parametrisch	-0.34639	0.12098	-0.23870	0.05756

Tabelle 4.8: Bias und MSE $\tilde{r}_{Y_2Z_2}$ und $\tilde{r}_{Y_3Z_3}$ für $\mathbf{R_{YZ|X}} = 0.35$

Die Unterschätzung in Folge einer Verletzung der CIA wird jedoch durch eine hohe Explanatory Power verringert (siehe auch die restlichen Elemente von $\mathbf{R_{YZ|X}}$ im Anhang C.2). Diese Erkenntnis bestätigt die Aussage von Rässler (vgl. 2002, S. 57), dass die dritte Validitätsstufe bei einer hohen Explanatory Power von $\mathbf{X}$ eher erfüllt werden kann, da so der mögliche Wertebereich für $\mathbf{R_{YZ}}$ verkleinert wird. Allerdings können der

Bias und der MSE von $\tilde{\boldsymbol{R}}_{\boldsymbol{YZ}}$ so nur abgemildert werden; die dritte Validitätsstufe kann in dieser Simulationsstudie für eine Verletzung der CIA dennoch in keiner Situation als erfüllt angesehen werden.

Der Vergleich der einzelnen Methoden zeigt, dass die parametrische Methode für eine hohe Explanatory Power im Vergleich etwas geringere Verzerrungen aufweist. Für diese Situation zeigen sich auch bessere Ergebnisse für das Szenario A_{300}, was darauf hindeuten könnte, dass die Unterschätzungen in Folge einer Verletzung der CIA durch einen größeren Spenderdatensatz abgemildert werden können. Dieser Effekt ist jedoch geringer als der Einfluss einer hohen Explanatory Power. Für eine geringe Explanatory Power zeigen sich hingegen keine Unterschiede zwischen den Methoden und Szenarien. Außerdem ist zu betonen, dass die nonparametrische Methode aufgrund der Symmetrie von $d(\mathbf{x}_a, \mathbf{x}_b)$ dieselben Ergebnisse für A_{500} und $A \cup B_{500}$ aufweist. Für den Erhalt von $\mathbf{R_{YZ}}$ macht es also keinen Unterschied, welches dieser beiden Szenarien gewählt wird. Abschließend veranschaulichen die Ergebnisse für $\tilde{\mathbf{R}}_{\mathbf{YZ|X}}$ erneut die Problematik der bedingten Unabhängigkeitsannahme. Tabelle 4.9 stellt den Bias und den MSE von $\tilde{\mathbf{R}}_{\mathbf{YZ|X}}$ aggregiert über alle $r_{Y_q Z_r|\mathbf{X}}$ dar.

Szenario	Methode	$\mathbf{R_{YZ\|X}} = 0$		$\mathbf{R_{YZ\|X}} = 0.2$		$\mathbf{R_{YZ\|X}} = 0.35$	
		Bias	MSE	Bias	MSE	Bias	MSE
A_{300}	gemischt	-0.00045	0.00328	-0.20074	0.04379	-0.35091	0.12661
	nonparam.	0.00048	0.00336	-0.20021	0.04351	-0.34993	0.12585
	parametr.	-0.00021	0.00332	-0.20070	0.04377	-0.35054	0.12637
A_{500}	gemischt	-0.00082	0.00202	-0.20108	0.04248	-0.35064	0.12498
	nonparam.	-0.00015	0.00202	-0.20056	0.04224	-0.35024	0.12475
	parametr.	-0.00085	0.00202	-0.20075	0.04232	-0.35077	0.12507
$A \cup B_{500}$	gemischt	-0.00059	0.00101	-0.20027	0.04110	-0.35006	0.12354
	nonparam.	-0.00006	0.00201	-0.20056	0.04223	-0.35034	0.12480
	parametr.	-0.00044	0.00100	-0.20012	0.04105	-0.35029	0.12371

Tabelle 4.9: Bias und MSE von $\frac{1}{9}\sum_{q=1}^{3}\sum_{r=1}^{3}\tilde{r}_{Y_q Z_r|\mathbf{X}}$

Aus ihr geht hervor, dass mit den verwendeten Matching-Methoden, die auf Basis der bedingten Unabhängigkeitsannahme durchgeführt werden, stets $\tilde{\mathbf{R}}_{\mathbf{YZ|X}} = 0$ im fusionierten Datensatz implementiert wird: Die bedingte Korrelation wird für $\mathbf{R_{YZ|X}} = 0.2$ um ca. 0.2 und für $\mathbf{R_{YZ|X}} = 0.35$ um ca. 0.35 unterschätzt. Mit den verwendeten

Methoden erhält man also ungeachtet der wahren Werte für $\boldsymbol{R_{YZ}}$ und $\boldsymbol{R_{YZ|X}}$ stets denselben fusionierten Datensatz.

Dass diese Unterschätzungen von $\tilde{\mathbf{R}}_{\mathbf{YZ|X}}$ für alle Methoden und Szenarien gleich sind, zeigt nochmals, wie hoch der Einfluss einer optimalen Wahl gemeinsamer Variablen $\mathbf{X}$ mit hoher Explanatory Power und die damit verbundene Verkleinerung des Wertebereichs von $\tilde{r}_{Y_q Z_r}$ für die Qualität des fusionierten Datensatzes ist.

4.3.3 Kennzahlen der zweiten Validitätsstufe

Die zweite und wichtigste Validitätsstufe gibt an, wie gut $f_{\mathbf{XYZ}}$ im fusionierten Datensatz nachgebildet werden kann und wird durch die Hellinger-Distanz gemessen. Die Resultate sind in Tabelle 4.10 angegeben:

Szenario	Methode	$\mathbf{R_{YZ\|X}} = 0$	$\mathbf{R_{YZ\|X}} = 0.2$	$\mathbf{R_{YZ\|X}} = 0.35$
A_{300}	gemischt	0.11277	0.25464	0.83214
	nonparametrisch	0.09170	0.24926	0.83239
	parametrisch	0.10687	0.25458	0.83274
A_{500}	gemischt	0.06730	0.24338	0.83200
	nonparametrisch	0.05024	0.23879	0.83159
	parametrisch	0.08147	0.24632	0.83215
$A \cup B_{500}$	gemischt	**0.04829***	0.23875	0.83167
	nonparametrisch	**0.04874***	0.23890	0.83171
	parametrisch	0.05784	0.23968	0.83156

Tabelle 4.10: Hellinger-Distanz zum Vergleich von $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$: Für Simulationsdurchläufe, deren Distanzen mit einem hochgestellten * gekennzeichnet sind, können $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ nach dem in Leulescu und Agafitei (2013) genannten Richtwert als gleich bzw. ähnlich angesehen werden.

Für das Zutreffen der CIA kann die zweite Validitätsstufe unter Anwendung der in Leulescu und Agafitei (2013) verwendeten Faustregel nur für die gemischte und die nonparametrische Methode in $A \cup B_{500}$ als erfüllt angesehen werden. Beide Methoden weisen im Szenario A_{500} jedoch nur geringfügig höhere Werte als 0.05 auf. Dasselbe gilt für die parametrische Methode in $A \cup B_{500}$. Die restlichen Simulationsdurchläufe erzielen jedoch schlechtere Ergebnisse.

Aus der separierten Betrachtung der einzelnen Szenarien für $\mathbf{R_{YZ|X}} = 0$ lässt sich schließen, dass für die vorliegende Datensituation mit dem Szenario $A \cup B_{500}$ die besten

Resultate erzielt werden konnten. A_{300} wies deutlich schlechtere Werte auf.
Bei den Methoden konnten die besten Werte für die nonparametrische Methode erzielt werden, während die parametrische Methode eine schlechtere Performance aufweisen. Auffällig ist das besonders schlechte Abschneiden der gemischten Methode im Szenario A_{300}. Dies könnte mit der schlechteren Auswahl der Spenderbeobachtungen im Vergleich zur nonparametrischen Methode erklärt werden (siehe Tabelle 4.4).
Für $\mathbf{R_{YZ|X}} = 0.2$ sind die Werte der Hellinger-Distanz so hoch, dass $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ in keinem Fall als gleich angesehen werden können und die zweite Validitätsstufe somit nicht als erfüllt betrachtet werden kann. Insgesamt zeigen auch hier das Szenario $A \cup B_{500}$ sowie die nonparametrische Methode etwas bessere Resultate.
Für $R_{\mathbf{YZ|X}} = 0.35$, welches die höchste gewählte bedingte Korrelation darstellt, zeigen sich große Unterschiede zwischen $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$: Die Hellinger-Distanz weist hier einen durchschnittlichen Wert von 0.83199 auf. Ein Unterschied zwischen den verschiedenen Szenarien und Methoden kann nicht mehr festgestellt werden.

4.3.4 Fazit der Simulationsstudie

Die vierte Validitätsstufe konnte aufgrund des Simulationsdesigns für alle Durchläufe gut erfüllt werden. Dies zeigt, dass das Erfüllen der vierten Validitätsstufe keinen Hinweis über das Zutreffen weiterer Validitätsstufen liefert. So wird nochmals verdeutlicht, wie bedenklich es ist, sein Urteil über das Gelingen eines Statistischen Matchings anhand der Ergebnisse der der vierten Validitätsstufe zu fällen.
Die Zuordnung des fusionierten Datensatzes zu den höheren Validitätsstufen erfolgt in dieser Simulationsstudie nur bei korrekt getroffener bedingter Unabhängigkeitsannahme: Die dritte Validitätsstufe konnte einzig für $\mathbf{R_{YZ|X}} = 0$ als erfüllt betrachtet werden. Verletzungen dieser Annahme führen wie erwartet zu einer Unterschätzung von $\tilde{\boldsymbol{R}}_{\boldsymbol{YZ}}$ im fusionierten Datensatz, die umso höher ausfällt, je stärker die bedingte Unabhängigkeitsannahme verletzt wurde. Jedoch fallen diese Unterschätzungen für Variablen mit einer hohen Explanatory Power von $\mathbf{X}$ geringer aus. Dies gilt in der Simulationsstudie insbesondere für die parametrische Methode.
Die zweite Validitätsstufe konnte wiederum nicht für alle Durchläufe mit $\mathbf{R_{YZ|X}} = 0$ erreicht werden. Je nachdem, welcher Grenzwert für die Hellinger-Distanz angewandt

wird, können $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ nur für $A \cup B_{500}$ bei gemischter und nonparametrischer Methode als erfüllt angesehen werden.

Insgesamt verdeutlichen die Ergebnisse der Simulationsstudie die Schwierigkeit, einen fusionierten Datensatz der zweiten Validitätsstufe zu erhalten, der ohne Einschränkungen analysiert werden kann. Der Vergleich der einzelnen Szenarien zeigt, dass unter dem Szenario $A \cup B_{500}$ insgesamt bessere Ergebnisse für die vierte und die zweite Validitätsstufe erzielt werden konnten. Sofern die gemeinsamen Variablen eine hohe Explanatory Power aufweisen, konnten für die dritte Validitätsstufe bessere Ergebnisse für A_{300} beobachtet werden.

Beim Vergleich der verschiedenen Methoden zeigt sich, dass die zweite und vierte Stufe am besten mit der nonparametrischen Stufe erfüllt werden konnten. Für $\mathbf{R_{YZ}}$ konnte die parametrische Methode bei einer hohen Explanatory Power ebenfalls eine etwas geringere Verzerrung aufweisen.

Insgesamt bestätigen die Ergebnisse der Simulationsstudie die Aussage von Rässler (vgl. 2002, S. 43), dass die Wahl der gemeinsamen Variablen für die Qualität des fusionierten Datensatzes eine viel höhere Relevanz besitzt als die Wahl der Matching-Methode. Im Vergleich dazu scheint in dieser Simulationsstudie auch die Wahl des Matching-Szenarios eher eine untergeordnete Rolle zu spielen.

Aus diesen Gründen ist eine sorgfältige Auswahl der gemeinsamen Variablen in der praktischen Anwendung sehr wichtig. D'Orazio et al. (2006, Kap. 6.2) beschreiben verschiedene Auswahlverfahren dieser Variablen für verschiedene Datentypen. Auch wird in Rässler (vgl. 2002, S.117) die Höhe der Explanatory Power der gemeinsamen Variablen und die damit einhergehende Beschränkung des möglichen Wertebereichs zur Bewertung der Qualität eines fusionierten Datensatzes vorgeschlagen.

Im Idealfall sollten bereits bei der Erhebung der Datensätze A und B inhaltliche Überlegungen über möglichst optimale gemeinsame Variablen $\mathbf{X}$ erfolgen. Die Messung dieser Variablen sollte anschließend normiert erfolgen (vgl. z. B. D'Orazio et al., 2006, S. 163). Dieses Vorgehen klingt relativ aufwendig, stellt aber vor allem für große Statistische Ämter wie beispielsweise Eurostat eine durchaus praktikable Methodik dar.

Als weitere Heuristik zur Bewertung der Validität der aus einem fusionierten Datensatz gewonnenen Analyseergebnisse, können Beschränkungen verwendet werden, die mit der positiven Semidefinitheit von $\boldsymbol{\Sigma_{XYZ}}$ einhergehen (siehe dazu auch verschiedene in Rässler (2002) vorgestellte multiple Imputationsmethoden).

Prinzipiell muss bei der Interpretation der Ergebnisse stets bedacht werden, dass diese einzig Rückschlüsse auf die hier behandelte Datensituation zulassen. Daher muss auch der Einfluss des Simulationsdesigns auf die Ergebnisse diskutiert werden:
Das Statistische Matching wurde in der Simulationsstudie nur durch einen geringen Matching Noise beeinträchtigt. Dies erklärt, warum die vierte Validitätsstufe für alle Simulationsdurchläufe als erfüllt betrachtet werden konnte. Für reale Daten ist dies jedoch äußerst unrealistisch, so dass selbst die vierte Stufe unter Umständen nur sehr schwierig und nur mit teils aufwendigen Harmonisierungsverfahren (vgl. D'Orazio et al., 2006, Kap. 6.1) erfüllt werden kann: Oft ist das Erreichen einer zufriedenstellenden Kompatibilität schlicht nicht möglich, weil die zu fusionierenden Datensätze eine zu schlechte Qualität aufweisen oder Variablen unterschiedlich gemessen wurden.
Ebenfalls durch den geringen Matching Noise kann das gute Abschneiden der nonparametrischen Methode erklärt werden. Auch beim Szenario $A \cup B_{500}$ werden bei höherem Matching Noise erhebliche Verschlechterungen erwartet (vgl. D'Orazio et al., 2006, S. 36). Daher wäre eine Folgestudie unter hohem Matching Noise interessant.
Auch wurde $A \cup B_{500}$ so konstruiert, dass die Durchführung eines beschränkten Matchings möglich ist. Weisen A und B eine ungleiche Anzahl an Beobachtungen auf, so können für die gemischte und die nonparametrische Methode nur unbeschränkte Fusionierungen durchgeführt werden. Diese gehen im Allgemeinen mit einer schlechteren Qualität des fusionierten Datensatzes einher. In einem solchen Fall gilt es zu prüfen, ob mit der parametrischen Methode bessere Ergebnisse erzielt werden können.
Des Weiteren werden die besseren Ergebnisse der parametrischen Methode für die dritte Validitätsstufe vermutlich durch die Tatsache begünstigt, dass der verwendete Datensatz einer multivariaten Normalverteilung folgt. Daher ist es fraglich, ob diese Methode bei Daten, die leichte oder starke Abweichungen von der angenommenen Verteilung aufweisen, ebenfalls gute Ergebnisse erzielt oder ob andere Methoden in so einem Fall besser abschneiden würden.

Zudem könnte das schlechte Abschneiden des Szenarios A_{300} für die zweite und vierte Stufe auch im geringeren Stichprobenumfang der fusionierten Beobachtungen begründet sein. Dadurch steigt die Variabilität in den Daten, welche beispielsweise von der Hellinger-Distanz oder dem Bias nicht berücksichtigt wird. Es stellt sich daher die Frage, ob das in A_{300} implementierte Verhältnis von Spender- und Empfängerdatensatz bessere Ergebnisse für einen höheren Stichprobenumfang erzielen könnte. Eventuell könnte auch der verwendete Grenzwert der Hellinger-Distanz an die Anzahl fusionierter Beobachtungen angepasst werden.

4.4 Vergleich mit anderen Evaluationsstudien

Abschließend werden die Ergebnisse der Simulationsstudie mit den Resultaten bereits durchgeführter Evaluationsstudien verglichen. Dabei wurden Studien herangezogen, in welchen ebenfalls die Qualität fusionierter Datensätze unter der bedingten Unabhängigkeitsannahme bewertet wurde. In den meisten dieser Studien wurden – anders als in dieser Arbeit – reale Datensätze verwendet. Nur Barr et al. (1982) evaluieren zusätzlich einen künstlichen Datensatz.
Eine Zusammenfassung ausgewählter Studien liefern Rodgers (1984) und Paass (1986). Meist wurden in den Studien verschiedene nonparametrische Methoden, wie Random Draws oder das Distance-Hot-Deck-Matching, verglichen (siehe bspw. Barry, 1988 Rodgers et al., 1981 Paass und Wauschkuhn, 1980 oder Ruggles et al., 1977). Auch Vergleiche von beschränktem und unbeschränktem Matching wurden durchgeführt (siehe z. B. Barr et al., 1982).
In allen Fällen kamen die Evaluationsstudien ebenfalls zu dem Ergebnis, dass die Beziehung von $\mathbf{X}$ und $\mathbf{Z}$ sowie die Verteilung von $\mathbf{Z}$ im fusionierten Datensatz relativ gut erhalten werden konnten. Allerdings konnte die Beziehung von $\mathbf{Y}$ und $\mathbf{Z}$ bei Verletzungen der CIA nicht korrekt spezifiziert werden. Die zweite Validitätsstufe bzw. deren inhaltliche Entsprechung wurde in diesen Studien nicht gemessen.
Dementsprechend wird das Statistische Matching unter der bedingten Unabhängigkeitsannahme eher kritisch betrachtet, da die Beziehung von $\mathbf{Y}$ und $\mathbf{Z}$ nur korrekt

nachgebildet werden kann, wenn die unsichere Annahme der bedingten Unabhängigkeit erfüllt ist.

Einzig Ruggles et al. (1977) ziehen in einer der ersten Studien zur qualitativen Bewertung fusionierter Datensätze ein positives Fazit. Diese Studie wurde jedoch u. a. von Rodgers (1984) kritisiert, da in die Bewertung der Qualität des fusionierten Datensatzes nur in wenigen Fällen die Beziehung von $\mathbf{Y}$ und $\mathbf{Z}$ berücksichtigt wurde.

Insgesamt betrachtet decken sich die Ergebnisse der untersuchten Evaluationsstudien auf Basis realer Daten mit den Ergebnissen der in dieser Arbeit durchgeführten Simulationsstudie.

Auch die Studie mit simulierten Daten in Barr et al. (1982) kommen zu dem Ergebnis, dass $f_{\mathbf{XYZ}}$ bei Verletzung der Unabhängigkeitsannahme nicht erhalten werden kann. Zudem merken sie an, dass $\boldsymbol{\Sigma}_{\boldsymbol{XZ}}$ im fusionierten Datensatz unterschätzt werden kann, wenn $\mathbf{X}$ und $\mathbf{Z}$ zu hoch korreliert sind, was zu einer zusätzlichen Verzerrung von $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{YZ}}$ führen kann.

4.5 Analysen mit fusionierten Datensätzen

Die Ergebnisse der Simulationsstudie zeigen, dass bei Verletzungen der bedingten Unabhängigkeitsannahme nur die vierte Validitätsstufe erfüllt werden kann. Für bestimmte Analysen mag diese Validitätsstufe zwar ausreichend sein, jedoch können viele statistische Fragestellungen mit einem fusionierten Datensatz der vierten Validitätsstufe nicht mehr beantwortet werden.

So wurde in Kapitel 3.2 bereits erwähnt, dass das Erfüllen der ersten oder zweiten Validitätsstufe die korrekte Durchführung aller Analysen gewährleistet, die auch mit dem Originaldatensatz möglich wären. Für Datensätze, die der dritten oder vierten Validitätsstufe zugeordnet werden können, sind die Analysemöglichkeiten eingeschränkt: Wird die vierte Validitätsstufe als erfüllt angesehen, so können alle univariaten Analysen durchgeführt werden (vgl. Rodgers, 1984). Analyseverfahren wie univariate Häufigkeitstabellen, Mittelwerte oder Varianzen können also korrekt spezifiziert werden. Auch univariate Tests können korrekte Ergebnisse erzielen.

Des Weiteren können für diese Validitätsstufe Analysen, welche $(\mathbf{X}, \mathbf{Y})$ bzw. $(\mathbf{X}, \mathbf{Z})$ betreffen, durchgeführt werden. So können $\boldsymbol{\Sigma_{XY}}$, $\boldsymbol{\Sigma_{XZ}}$,$\boldsymbol{\Sigma_X}$, $\boldsymbol{\Sigma_Y}$ und $\boldsymbol{\Sigma_Z}$ korrekt aus dem fusionierten Datensatz geschätzt werden und auch Regressionsmodelle, welche sich auf $(\mathbf{X}, \mathbf{Y})$ bzw. $(\mathbf{X}, \mathbf{Z})$ beziehen, liefern korrekte Schätzer.
Dies lässt sich auch anhand der Daten aus der Simulation zeigen: Angewandt auf einen Datensatz, welcher mit der gemischten Methode im Szenario $A \cup B_{500}$ fusioniert wurde, können für die Regressionsgleichung

$$z_{i1} = \alpha + x_{i1}\beta_{X_1} + x_{i2}\beta_{X_2} + x_{i3}\beta_{X_3} + \epsilon_{i3}$$

mit $i = (1, ..., n_A + n_B)$ die in Tabelle 4.11 aufgelisteten Ergebnisse berechnet werden. Je Simulationsdurchlauf wurden 500 Wiederholungen durchgeführt und jeweils Bias und MSE zu den Regressionsschätzern aus dem Originaldatensatz gemessen.

	β_0	β_{X_1}	β_{X_2}	β_{X_3}
$\mathbf{R_{YZ\|X}} = 0$	0	0.00769	0.00646	0.00155
	(0)	(0.00031)	(0.00038)	(0.00039)
$\mathbf{R_{YZ\|X}} = 0.2$	0	0.01019	0.00515	0.00105
	(0)	(0.00038)	(0.00033)	(0.00039)
$\mathbf{R_{YZ\|X}} = 0.35$	0	0.00985	0.00443	0.00052
	(0)	(0.00037)	(0.00030)	(0.00038)
wahre Werte	0	0.43469	0.27600	0.15132

Tabelle 4.11: Mittlere Verzerrung der Regressionskoeffizienten für das Szenario $A \cup B_{500}$ unter Verwendung der gemischten Methode für die verschiedenen Werte von $\mathbf{R_{YZ|X}}$. In Klammern wird jeweils der MSE angegeben.

Die Regressionsschätzer konnten in allen Fällen korrekt spezifiziert werden und auch der MSE weist sehr geringe Werte auf.
Für die korrekte Spezifikation der Analyseverfahren für $(\mathbf{X}, \mathbf{Y})$ und $(\mathbf{X}, \mathbf{Z})$ ist das Zutreffen der CIA also nicht nötig. Es muss allerdings auch beachtet werden, dass für die Analyse dieser Beziehungen auch die Durchführung eines Statistisches Matchings nicht nötig ist, da die dafür benötigten Informationen bereits vollständig in den Datensätzen A und B vorhanden sind (vgl. dazu Rodgers, 1984).

Gilt die dritte Stufe als erfüllt, so kann der fusionierte Datensatz zumindest für Analysen, welche auf Kovarianzen oder Korrelationen basieren, korrekte Ergebnisse erzielen (vgl. Rässler, 2002, Kap. 2.5.3). So können in diesem Fall lineare Regressionsmodelle, die eine wichtige Rolle in vielen Analysen spielen, korrekt spezifiziert werden.
Jedoch zeigt Tabelle 4.12 exemplarisch für die Verwendung der gemischten Matching-Methode und Szenario $A \cup B$ die Schwere der Auswirkungen auf die Regressionsschätzung bei fälschlicherweise angenommener CIA. Getestet wurde dabei das Regressionsmodell

$$y_{i3} = \alpha + x_{i1}\beta_{X_1} + x_{i2}\beta_{X_2} + x_{i3}\beta_{X_3} + z_{i1}\beta_{Z_1} + z_{i2}\beta_{Z_2} + z_{i3}\beta_{Z_3} + \epsilon_{i3},$$

welches ebenfalls auf Basis eines anhand der gemischten Methode und des Szenarios $A \cup B_{500}$ fusionierten Datensatzes berechnet wurde.

	β_0	β_{X_1}	β_{X_2}	β_{X_3}	β_{Z_1}	β_{Z_2}	β_{Z_3}
$\mathbf{R_{YZ\|X}} = 0$	0	0.00777	0.00698	0.00055	0.00300	0.00209	-0.00038
	(0)	(0.00054	(0.00036)	(0.00036)	(0.00070)	(0.00065)	(0.00095)
$\mathbf{R_{YZ\|X}} = 0.2$	0	-0.10046	-0.06924	0.00081	0.15775	0.14724	0.16193
	(0)	(0.01051)	(0.00515)	(0.00035)	(0.02558)	(0.02238)	(0.02713)
$\mathbf{R_{YZ\|X}} = 0.35$	0	0.28784	0.18879	0.05082	0.27967	0.25923	0.28094
	(0)	(0.03392)	(0.01588)	(0.00033)	(0.07911)	(0.06921)	(0.07996)
wahre Werte							
$\mathbf{R_{YZ\|X}} = 0$	0	0.47866	0.31790	0.05444	0	0	0
$\mathbf{R_{YZ\|X}} = 0.2$	0	0.36962	0.24412	0.05238	0.15981	0.14813	0.16053
$\mathbf{R_{YZ\|X}} = 0.35$	0	-0.18297	-0.12463	-0.00195	0.27989	0.2619	0.28116

Tabelle 4.12: Mittlere Verzerrung der Regressionskoeffizienten für das Szenario $A \cup B_{500}$ unter Verwendung der gemischten Methode für die verschiedenen Werte von $\mathbf{R_{YZ|X}}$. In Klammern wird jeweils der MSE angegeben.

Während für $\mathbf{R_{YZ|X}} = 0$ nur geringe Verzerrungen im Vergleich zu den Originalwerten zu beobachten sind, führt eine Verletzung der CIA zu einer Unterschätzung der Regressionskoeffizienten für X_1, X_2 und X_3 sowie zu einer Überschätzung für Z_1, Z_2 und Z_3. Die Abhängigkeitsstruktur wird anhand des fusionierten Datensatzes selbst für leichtere Verletzungen der CIA völlig falsch spezifiziert.
Für die korrekte Spezifikation vieler weiterer statistischer Analyseverfahren muss allerdings die zweite Validitätsstufe erfüllt sein. Dazu zählen vor allem nonparametrische Verfahren, die auf der Nächsten-Nachbar-Klassifikation beruhen. Auch ein Vergleich von $f_{\mathbf{XYZ}}$ oder $f_{\mathbf{YZ}}$ mit anderen Verteilungen ist nur für diese Validitätsstufe möglich.

5 Zusammenfassung und Ausblick

5.1 Zusammenfassung

In den letzten 15 Jahren wurde wieder vermehrt Forschung auf dem Gebiet des Statistischen Matchings betrieben. Besonders für die steigende Zahl verfügbarer Datenquellen im Zuge der Digitalisierung bietet diese Methode viele Chancen. Aufgrund von Datenschutzregelungen oder fehlender Überschneidungen der Individuen in den zu fusionierenden Datensätzen ist das Statistische Matching insbesondere für die Fusionierung hochdimensionaler Datensätze häufig die einzige praktikable Möglichkeit.
Speziell das Statistische Matching unter der bedingten Unabhängigkeitsannahme stellt dabei ein sehr attraktives Verfahren dar, da mit diesem Verfahren Punktschätzer für die fehlenden Werte berechnet werden können, ohne dass zusätzliche Daten erforderlich sind.
Zur korrekten Verwendung eines fusionierten Datensatzes sind jedoch Aussagen über dessen Qualität unabdingbar. Nur so können mögliche Analysetechniken für die fusionierten Daten festgelegt werden. Dies war auch das Ziel dieser Masterarbeit:
Anhand einer Simulationsstudie wurden die Auswirkungen verschiedener Parameter auf einen fusionierten Datensatz getestet. Dazu war im Vorfeld die Entwicklung einer passenden Bewertungsgrundlage nötig. Hierbei wurde sich auf die vier Validitätsstufen von Rässler (2002) bezogen, die den fusionierten Datensatz einer von vier Validitätsstufen zuordnen. Erfüllt ein Datensatz die erste oder zweite Validitätsstufe, so wird die Qualität des Datensatzes nicht durch den Fusionierungsprozess beeinträchtigt. Bei Erfüllung der dritten und vierten Validitätsstufe müssen hingegen Einschränkungen bei den möglichen Analysemethoden in Kauf genommen werden. Die vierte Validitätsstufe ist hierbei die einzige Stufe, die auch in der Realität getestet werden kann und gilt daher auch als Minimalanforderung an das Statistische Matching.

Die Simulationsstudie konnte zeigen, dass für eine fälschlicherweise getroffene bedingte Unabhängigkeitsannahme die Beziehung von **Y** und **Z** stets unterschätzt wird und deshalb höchstens die vierte Validitätsstufe erreicht werden kann. Die dritte Validitätsstufe, die u. a. eine korrekte Anwendung von linearen Regressionsmodellen erlaubt, konnte in der Simulationsstudie nur bei korrekt spezifizierter CIA erfüllt werden. Dies ist insofern problematisch, da die bedingte Unabhängigkeit von **Y** und **Z** gegeben **X** im Vorfeld nicht getestet werden kann.
Zudem bestätigten die Ergebnisse der Simulationsstudie, dass die von einer Verletzung der CIA hervorgerufenen Verzerrungen durch eine hohe Explanatory Power abgeschwächt werden. Daher gilt in der praktischen Anwendung, dass gemeinsame Variablen mit einer möglichst hohen Explanatory Power verwendet werden sollten (vgl. dazu auch die Erkenntnisse in Rässler, 2002).
Bemerkenswert war, dass die zweite Validitätsstufe selbst bei Einhaltung aller zentralen Annahmen nur für wenige Einstellungen der Simulation als erfüllt angesehen werden konnte. Möglicherweise wurden die Explanatory Power der gemeinsamen Variablen und die Stichprobengrößen zu gering festgelegt. Zudem wurde auf Basis eines relativ hochdimensionalen Datensatzes getestet, was ebenfalls zu einer verringerten Qualität des fusionierten Datensatzes geführt haben könnte.
Des Weiteren konnten für diese Datensituation mit einer vollständigen Imputation von $A \cup B$ insgesamt etwas bessere Ergebnisse erzielt werden als bei der Vervollständigung nur einer der beiden Datensätze. Unter den Matching-Methoden schnitt die nonparametrische Methode insgesamt am besten ab. Diese Methode wird in der praktischen Anwendung allerdings stärker durch den Matching Noise beeinflusst, der in dieser Simulationsstudie sehr gering war. Die dritte Stufe konnte mit der parametrischen Methode besser erfüllt werden.
Prinzipiell haben sich die Validitätsstufen in Simulationsstudien als gute Methode zur Bewertung der Qualität eines fusionierten Datensatzes erwiesen. In der realen Anwendung kann jedoch aufgrund der fehlenden Informationen nur die vierte Validitätsstufe untersucht werden. Auf deren Basis kann jedoch keine Aussage über das Gelingen eines Fusionierungsprozesses getroffen werden. In der Praxis ist man daher auf die Anwendung verschiedener Heuristiken angewiesen, die zumindest Hinweise auf

die Qualität des fusionierten Datensatzes geben können. Dazu zählen die Höhe der Explanatory Power von $\mathbf{X}$, der Matching Noise oder die Beschränkungen von $\boldsymbol{\Sigma_{YZ}}$ durch die positive Definitheit der Kovarianzmatrix $\boldsymbol{\Sigma_{XYZ}}$. Es wäre daher wünschenswert und sinnvoll, diese bereits in der Literatur verwendeten Heuristiken weiter zu entwickeln und eventuell neue Eigenschaften zu ermitteln, die eine Hilfestellung bei der Einschätzung der Qualität des fusionierten Datensatzes in der praktischen Anwendung bieten.
Generell bleibt das Statistische Matching unter der bedingten Unabhängigkeitsannahme eine unsichere Methode, die nur angewandt werden sollte, wenn aufgrund von Erfahrungswerten oder Fachwissen keine Zweifel am Zutreffen der nötigen Annahmen bestehen und die gemeinsamen Variablen sorgfältig ausgewählt wurden.
Sollte dies nicht der Fall sein, so ist die Wahl eines anderen Ansatzes, wie das Hinzuziehen zusätzlicher Informationen oder die Beibehaltung der Unsicherheit, vorzuziehen. Prinzipiell sollte bei der weiteren Bearbeitung des fusionierten Datensatzes jedoch stets beachtet werden, dass dieser künstlich erzeugt wurde.

5.2 Ausblick

An dieser Stelle soll nochmals betont werden, dass die Wahl der Kennzahlen an die vorliegende Datensituation und die Analyseziele angepasst werden sollte. Die Wahl der hier verwendeten Kennzahlen kann daher durchaus optimiert und erweitert werden. Für die Messung der vierten Stufe könnte insbesondere für größere Stichprobenerhebungen auch die Hellinger-Distanz verwendet werden. Damit könnte potentiell eine bessere inhaltliche Bewertung dieser Stufe gewährleistet werden und die Beziehung der Variablen in $\mathbf{Z}$ untereinander besser miteinbezogen werden. Des Weiteren könnte der Erhalt der verschiedenen Korrelationen im fusionierten Datensatz beispielsweise ebenfalls durch statistische Tests validiert werden.
Des Weiteren berechnet die in der Arbeit verwendete Variante der Hellinger-Distanz für die zweite Validitätsstufe die Distanz zwischen zwei multivariaten Normalverteilungen. Eine Implementierung in R für weitere Verteilungsfamilien ist jedoch nicht vorhanden. Auch die Möglichkeit einer Berechnung für den verteilungsfreien Fall wäre wünschenswert, so dass Ausreißer oder Verunreinigungen einer Verteilung adäquat in die Berechnung miteinbezogen werden können.

Generell bietet das Simulationsdesign viel Potential für Folgestudien. Eine Frage, die noch weiterer empirischer Untersuchungen bedarf, stellt beispielsweise der Einfluss des Matching Noise auf eine Imputation von $A \cup B$ dar. Dieses Forschungsziel kann mit der Untersuchung des Einflusses von Messfehlern oder Fehlern im Studiendesign von A und B auf den fusionierten Datensatz kombiniert werden. Zudem wäre interessant, wie gut das Statistische Matching für Datensätze mit deutlich unterschiedlicher Qualität und unterschiedlichen Fragestellungen funktioniert. Auch die Konsequenzen unterschiedlicher zugrunde liegender Verteilungen von A und B auf die Qualität des fusionierten Datensatzes stellen eine interessante Forschungsfrage dar.
Zudem sollte eine Bewertung der untersuchten Szenarien und Methoden auch für Daten getroffen werden, die nicht exakt einer Verteilungsfamilie folgen, sondern Ausreißer oder Verschmutzungen aufweisen. Insbesondere für die parametrische Methode sind hier schlechtere Ergebnisse zu erwarten.
Auch könnten die in dieser Arbeit untersuchten Methoden und Szenarien für die Fusionierung größerer Stichproben evaluiert werden.
Prinzipiell ist auch das Untersuchen von diskreten oder gemischten Datensätzen eine lohnenswerte Aufgabe für zukünftige Studien über die Qualität des Statistischen Matchings. Da diese Datensätze in der Realität häufiger auftreten als rein stetige Datensätze, ist die Qualität des Statistischen Matchings hier besonders interessant. Allerdings findet man in der wissenschaftlichen Literatur relativ wenig Methoden, welche die besonderen Eigenschaften dieser Datenstrukturen für die Datenfusion berücksichtigen.
Erst im letzten Jahrzehnt wurden Methoden für rein kategoriale Datensätze vorgestellt, welche auch die logischen Beschränkungen eines Datensatzes in die Anwendung einer Matching-Methode einbeziehen (siehe D'Orazio et al., 2006b). Verwendet man für diese Art von Daten Matching-Methoden, welche auf Ähnlichkeitsmaßen beruhen, so scheinen die herkömmlichen Distanzmaße nicht ideal zu sein. McCane und Albert (2008) und Boriah et al. (2008) stellen alternative Distanzfunktionen und Ähnlichkeitsmaße vor, deren Performance bei den einzelnen Matching-Methoden getestet werden könnte. Des Weiteren fehlt eine weitreichende Implementierung statistischer Matching-Methoden für kategoriale und stetige Datensätze in R.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

A. Agresti. *Categorical data analysis*. Wiley-Interscience, Hoboken, NJ, 2. Edition, 2002.

A. Askinadze. Vergleich von Distanzen und Kernel für Klassifikatoren zur Optimierung der Annotation von Bildern. In *Datenbanksysteme für Business, Technologie und Web (BTW 2015)*, Jahrgang P-242, S. 193–202, 2015. URL `http://www.btw-2015.de/res/proceedings/Workshops/Stud/Askinadze-Vergleich_von_Distanzen_u`. aufgerufen am 23.01.17.

L. Baringhaus und C. Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.

R. S. Barr und J. S. Turner. Quality issues and evidence in statistical file merging. In G. E. Liepins und V. Uppuluri, Herausgeber, *Data quality control: Theory and pragmatics*, S. 245–315. CRC Press, 1990.

R. S. Barr, W. H. Stewart und J. S. Turner. An empirical evaluation of statistical matching methodologies. `http://scholar.smu.edu/cgi/viewcontent.cgi?article=1180&context=business_workingpapers`, 1982. aufgerufen am 21.01.17.

J. T. Barry. An investigation of statistical matching. *Journal of Applied Statistics*, 15 (3):275–283, 1988.

S. Bennike. Fusion - An overview by an outside observer. In H. Henry, Herausgeber, *Readership research: theory and practice*, S. 334–335. Elsevier Science & Technology, 1987.

S. Bihorel und M. Baudin. *Optimbase: R port of the Scilab optimbase module*, 2014. URL `https://CRAN.R-project.org/package=optimbase`. R-Paket Version 1.0-9.

S. Boriah, V. Chandola und V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, S. 243–254. SIAM, 2008.

P. L. Conti, D. Marella und M. Scanu. Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational Statistics & Data Analysis*, 53(2):354–365, 2008.

D. B. Dahl. *Xtable: Export tables to LaTeX or HTML*, 2016. URL `https://CRAN.R-project.org/package=xtable`. R-Paket Version 1.8-2.

M. D'Orazio. Evaluation of the accuracy of statistical matching. In *Essnet Statistical Methodology Project on Integration of Survey and Administrative Data*. Eurostat, 2009.

M. D'Orazio. StatMatch: Statistical Matching , 2016. URL `https://CRAN.R-project.org/package=StatMatch`. R-Paket Version 1.2.4.

M. D'Orazio, M. Di Zio und M. Scanu. *Statistical matching: Theory and practice*. John Wiley, Chichester, 2006.

M. D'Orazio, M. Di Zio und M. Scanu. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics*, 22(1):137, 2006b.

C. K. Enders. *Applied missing data analysis*. Guilford Press, New York, 2010.

L. Fahrmeir, T. Kneib und S. Lang. *Regression: Modelle, Methoden und Anwendungen*. Springer, Berlin, Heidelberg, 2. Edition, 2009.

I. P. Fellegi. Discussion. In *Proceedings of the American Statistical Association, Social Statistics Section*, S. 762–764, 1977.

C. Franz. Cramer: Multivariate nonparametric Cramer-Test for the two-sample-problem, 2014. URL `https://CRAN.R-project.org/package=cramer`.

S. Gabler. Datenfusion. *ZUMA-Nachrichten*, 40:81–92, 1997.

C. Hennig. Fpc: Flexible procedures for clustering, 2015. URL `https://CRAN.R-project.org/package=fpc`. R-Paket Version 2.1-10.

K. Hornik. *Clue: Cluster ensembles*, 2016. URL `https://CRAN.R-project.org/package=clue`. R-Paket Version 0.3-51.

D. R. Judkins. Comment. *Journal of the American Statistical Association*, 93(443): 861–864, 1998.

J. B. Kadane. Some statistical problems in merging data files. *Compendium of Tax Research*, S. 159–179, 1978.

H. Kiesl und S. Rässler. Techniken und Einsatzgebiete von Datenintegration und Datenfusion. In C. König, M. Stahl und E. Wiegand, Herausgeber, *Datenfusion und Datenintegration: 6. wissenschaftliche Tagung*, Jahrgang 6, S. 17–32, 2005.

W. J. Koschnick. *Standard-Lexikon für Markt-und Konsumforschung*. Saur, 1995.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1–2):83–97, 1955.

S. Künn. The challenges of linking survey and administrative data. *IZA World of Labor*, 2015.

F. Leisch. A toolbox for K-centroids cluster analysis. *Computational Statistics and Data Analysis*, 51(2):526–544, 2006. R-Paket Version 1.3-4.

A. Leulescu und M. Agafitei. Statistical matching: a model based approach for data integration. *Eurostat - Methodologies and Working papers*, 2013.

D. Marella, M. Scanu und P. L. Conti. On the matching noise of some nonparametric imputation procedures. *Statistics & Probability Letters*, 78(12):1593–1600, 2008.

B. McCane und M. Albert. Distance functions for categorical and mixed variables. *Pattern Recognition Letters*, 29(7):986–993, 2008.

C. Moriarity und F. Scheuren. Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17(3):407–422, 2001a.

C. Moriarity und F. Scheuren. Statistical matching: Pitfalls of current procedures. In *ASA Proceedings of the Joint Statistical Meetings, American Statistical Association*, 2001b.

C. Moriarity und F. Scheuren. A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 21(1):65–73, 2003.

C. Moriarity und F. Scheuren. Regression-based statistical matching: Recent developments. In *ASA Proceedings of the Joint Statistical Meetings, American Statistical Association*, S. 4050–4057, 2004.

F. Novomestky. *Matrixcalc: Collection of functions for matrix calculations*, 2012. URL `https://CRAN.R-project.org/package=matrixcalc`. R-Paket Version 1.0-3.

D. Nychka, R. Furrer, J. Paige und S. Sain. Fields: Tools for spatial data, 2015. URL `www.image.ucar.edu/fields`. R-Paket Version 8.4-1.

B. Okner. Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1(3):325–362, 1972.

G. Paass. Statistical match: Evaluation of existing procedures and improvements by using additional information. *Microanalytic Simulation Models to Support Social and Financial Policy*, S. 401–422, 1986.

G. Paass und U. Wauschkuhn. Experimentelle Erprobung und Vergleichende Bewertung Statistischer Matchverfahren, Interner Bericht: IPES. 80.201, 1980.

R Development Core Team. R: A Language and Environment for Statistical Computing, 2008. URL `http://www.R-project.org`. Version 3.3.1.

A. Rasner, J. R. Frick und M. Grabka. Extending the empirical basis for wealth inequality research using statistical matching of administrative and survey data. *SOEP Papers on Multidisciplinary Panel Data Research*, 359:1–41, 2011.

S. Rässler. *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Springer Science & Business Media, New York, 2002.

S. Rässler. Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33(1-2):153–17, 2004.

S. Rässler und K. Fleischer. Aspects concerning data fusion techniques. In *International Workshop on Household Survey Nonresponse*, Jahrgang 4, S. 317–333. DEU, 1998.

M. L. Rizzo und G. J. Szekely. Energy: E-statistics: Multivariate inference via the energy of data, 2016. URL `https://CRAN.R-project.org/package=energy`. R-Paket Version 1.7-0.

W. L. Rodgers. An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2(1):91–102, 1984.

W. L. Rodgers, E. B. DeVol und G. Kalton. An evaluation of statistical matching. *Proceedings of the Survey Research Methods Section, American Statistical Association*, S. 128–132, 1981.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94, 1986.

N. Ruggles, R. Ruggles und E. N. Wolff. Merging microdata rationale practice and testing. *Annals of Economic and Social Measurement*, 6(4):407–428, 1977.

L. Rüschendorf. *Mathematische Statistik*. Springer, Berlin, Heidelberg, 2014.

C. A. Sims. Comments (on Okner 1972). *Annals of Economic and Social Measurement*, 1(3):343–345, 1972.

A. C. Singh, H. Mantel, M. Kinack und G. Rowe. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19(1):59–79, 1993.

P. Sólymos. Processing ecological data in R with the mefa package. *Journal of Statistical Software*, 29(8):1–28, 2009. R-Paket Version 3.2-7.

G. J. Székely. E-Statistics: The energy of statistical samples. Technical Report 02-16, Bowling Green State University, Ohio, October 2002. URL `http://personal.bgsu.edu/~mrizzo/energy/Szekely-E-statistics.pdf`. aufgerufen am 21.1.17.

G. J. Székely und M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5:1–6, 2004.

G. Trenkler und H. Büning. *Nichtparametrische statistische Methoden*. De Gruyter, Berlin, 1994.

W. N. Venables und B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4. Edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. R-Paket Version 7.3-45.

M. P. Wand und M. C. Jones. *Kernel Smoothing*, Jahrgang 60 of *Monographs on Statistics and Applied Probability*. Crc Press, Florida, 1995.

D. Webber und R. Tonkin. Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. *Eurostat - Working Papers and Methodologies*, 2013. URL `http://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF/37d4ffcc-e9fc-42bc-8d4f-fc89c65ff6b1`. aufgerufen am 04.02.17.

T. Wei und V. Simko. *Corrplot: Visualization of a correlation matrix*, 2016. URL `https://CRAN.R-project.org/package=corrplot`. R-Paket Version 0.77.

J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. John Wiley, Chichester, 1990.

M. Zambrano-Bigiarini. *HydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series*, 2014. URL `https://CRAN.R-project.org/package=hydroGOF`. R-Paket Version 0.3-8.

A Simulation zur Wahl der Distanzfunktion

Um die jeweils beste Distanzfunktion für die in der Simulationsstudie verwendete Datensituation zu finden, wurden für die beiden auf Abstandsmaßen beruhenden Matching-Methoden jeweils eine zusätzliche Simulationsstudie durchgeführt.
Dabei wurden die Einstellungen der Simulationsstudie in Kapitel 3.4 weitestgehend beibehalten. Allerdings wurden nur die ersten beiden Kovarianzmatrizen mit $\mathbf{R_{YZ|X}} = 0$ bzw. $\mathbf{R_{YZ|X}} = 0.2$ zur Erzeugung der Datensätze A und B verwendet. Ferner wurden nur $k = 100$ Iterationen pro Simulationsdurchlauf durchgeführt.
Untersucht wurde die Performance der folgenden drei Distanzmaße (siehe z. B. D'Orazio et al., 2006, Anhang A):

- Euklidische Distanz:
$$d_{ab} = \sqrt{\sum_{p=1}^{P} (\mathbf{x}_{ap} - \mathbf{x}_{bp})^2}$$
- Manhattan-Distanz:
$$d_{ab} = \sqrt{\sum_{p=1}^{P} |\mathbf{x}_{ap} - \mathbf{x}_{bp}|}$$
- Mahalanobis-Distanz:
$$d_{ab} = (\mathbf{x}_a - \mathbf{x}_b)^T \, \mathbf{\Sigma_X}^{-1} \, (\mathbf{x}_a - \mathbf{x}_b)$$

Für die Tests wurde exemplarisch das Szenario $A \cup B_{500}$ verwendet. In weiteren Arbeiten gilt es auszuloten, ob vor allem für das Szenario A_{300} bessere Ergebnisse mit einer anderen Distanzfunktion erzielt werden können.

A.1 Nonparametrische Methode

Die Ergebnisse für die nonparametrische Methode sind in den Tabellen A.1 und A.2 abgebildet. Als Kennzahlen wurden dabei jeweils die durchschnittlichen absoluten Abweichungen zum Originaldatensatz $\sum_{p=1}^{3}\sum_{r=1}^{3}|r_{X_pZ_r}|$, $\sum_{q=1}^{3}\sum_{r=1}^{3}|r_{Y_qZ_r}|$ und $\sum_{q=1}^{3}\sum_{r=1}^{3}|r_{Y_qZ_r|\mathbf{X}}|$ erhoben und aufsummiert.

$\mathbf{R_{YZ|X}} = 0$	eukl. Dist.	Manhattan-Dist.	Mahalanobis-Dist.
$\sum_{p=1}^{3}\sum_{r=1}^{3}\|r_{X_pZ_r}\|$	0.04424	0.05592	0.04960
$\sum_{q=1}^{3}\sum_{r=1}^{3}\|r_{Y_qZ_r}\|$	0.04473	0.05365	0.04923
$\sum_{q=1}^{3}\sum_{r=1}^{3}\|r_{Y_qZ_r\|\mathbf{X}}\|$	0.01981	0.03266	0.02945

Tabelle A.1: Die summierten absoluten Abweichungen zum Originaldatensatz unter Einhaltung der CIA für die nonparametrische Methode

$\mathbf{R_{YZ|X}} = 0.2$	eukl. Dist.	Manhattan-Dist.	Mahalanobis-Dist.
$\sum_{p=1}^{3}\sum_{r=1}^{3}\|r_{X_pZ_r}\|$	0.05540	0.05308	0.05515
$\sum_{q=1}^{3}\sum_{r=1}^{3}\|r_{Y_qZ_r}\|$	1.57580	1.56947	1.58098
$\sum_{q=1}^{3}\sum_{r=1}^{3}\|r_{Y_qZ_r\|\mathbf{X}}\|$	1.80866	1.80214	1.81494

Tabelle A.2: Die summierten absoluten Abweichungen zum Originaldatensatz unter einer bedingten Korrelation von 0.2 für die nonparametrische Methode

Während die euklidische Distanz für $\mathbf{R_{YZ|X}} = 0$ bessere Resultate aufweisen kann, scheint die Manhattan-Distanz für $\mathbf{R_{YZ|X}} = 0.2$ etwas besser abzuschneiden. Insgesamt bietet die euklidische Distanz für die hier verwendete Datensituation die beste Performance und wurde aus diesem Grund für die Verwendung in der Simulationsstudie ausgewählt.

Die Aussagen von Rodgers (1984) und Paass (1986), dass die Mahalanobis-Distanz verglichen mit anderen Distanzmaßen eine schlechtere Performance aufweist, lassen sich anhand der Ergebnisse dieser kleinen Simulationsstudie bestätigen. Allerdings ist der Unterschied für $\mathbf{R_{YZ|X}} = 0.2$ eher als gering anzusehen.

In einer weiteren Simulationsstudie könnten auch die Auswirkungen größerer Verletzungen der bedingten Unabhängigkeitsannahme auf die Performance der Distanzmaße untersucht werden. Zudem könnten weitere Distanzmaße zum Vergleich herangezogen werden.

A.2 Gemischte Methode

Zur Wahl der in der gemischten Methode zu verwendende Distanzfunktion wurden dieselben Simulationseinstellungen und Kennzahlen wie in Kapitel 5.2 verwendet. Hier zeigt die euklidische Distanz insgesamt etwas geringere Abweichungen (siehe Tabellen A.3 und A.4). Daher wird für die gemischte Methode ebenfalls diese Distanzfunktion herangezogen.

$\mathbf{R_{YZ\|X}} = 0$	eukl. Dist.	Manhattan-Dist.	Mahalanobis-Dist.
$\sum_{p=1}^{3}\sum_{r=1}^{3}\lvert r_{X_pZ_r}\rvert$	0.05812	0.06389	0.06504
$\sum_{q=1}^{3}\sum_{r=1}^{3}\lvert r_{Y_qZ_r}\rvert$	0.05249	0.05213	0.05851
$\sum_{q=1}^{3}\sum_{r=1}^{3}\lvert r_{Y_qZ_r\|\mathbf{X}}\rvert$	0.03128	0.02834	0.02941

Tabelle A.3: Die summierten absoluten Abweichungen zum Originaldatensatz unter Einhaltung der CIA für die gemischte Methode

$\mathbf{R_{YZ\|X}} = 0.2$	eukl. Dist.	Manhattan-Dist.	Mahalanobis-Dist.
$\sum_{p=1}^{3}\sum_{r=1}^{3}\lvert r_{X_pZ_r}\rvert$	0.06198	0.07624	0.07018
$\sum_{q=1}^{3}\sum_{r=1}^{3}\lvert r_{Y_qZ_r}\rvert$	1.58793	1.59407	1.59209
$\sum_{q=1}^{3}\sum_{r=1}^{3}\lvert r_{Y_qZ_r\|\mathbf{X}}\rvert$	1.81971	1.81719	1.81856

Tabelle A.4: Die summierten absoluten Abweichungen zum Originaldatensatz unter einer bedingten Korrelation von 0.2 für die gemischte Methode

B Übersicht verwendeter R-Pakete

Alle Berechnungen in dieser Arbeit wurden mit R (R Development Core Team, 2008, Version 3.3.1) durchgeführt. Die dazu verwendeten R-Pakete werden in diesem Abschnitt aufgelistet:

- `clue`, Version 0.3-51 (Hornik, 2016)
- `corrplot`, Version 0.77 (Wei und Simko, 2016)
- `energy`, Version 1.7-0(Rizzo und Szekely, 2016)
- `fields`, Version 8.4-1 (Nychka et al., 2015)
- `flexclust`, Version 1.3-4 (Leisch, 2006)
- `fpc`, Version 2.1-10 (Hennig, 2015)
- `hydroGOF`, Version 0.3-8 (Zambrano-Bigiarini, 2014)
- `MASS`, Version 7.3-45 (Venables und Ripley, 2002)
- `matrixcalc`, Version 1.0-3 (Novomestky, 2012)
- `mefa`, Version 3.2-7 (Sólymos, 2009)
- `optimbase`, Version 1.0-9 (Bihorel und Baudin, 2014)
- `StatMatch`, Version 1.2.4 (D'Orazio, 2016)
- `xtable`, Version 1.8-2 (Dahl, 2016)

C Ergebnisse der Simulationsstudie

C.1 Ergebnisse der vierten Validitätsstufe

C.1.1 Erhalt von $f_{\mathbf{Z}}$

$\mathbf{R_{YZ\|X}}$	Szenario	Methode	Z_1	Z_2	Z_3
$\mathbf{R_{YZ\|X}}$=0	A_{300}	gemischt	6	3	1
		nonparametrisch	0	3	0
		parametrisch	7	6	4
	A_{500}	gemischt	0	0	0
		nonparametrisch	0	0	0
		parametrisch	5	1	2
	$A \cup B_{500}$	gemischt	0	0	0
		nonparametrisch	0	0	0
		parametrisch	0	0	0
$\mathbf{R_{YZ\|X}}$=0.2	A_{300}	gemischt	4	2	3
		nonparametrisch	1	1	2
		parametrisch	3	6	2
	A_{500}	gemischt	0	0	0
		nonparametrisch	0	0	0
		parametrisch	2	2	3
	$A \cup B_{500}$	gemischt	0	0	0
		nonparametrisch	0	0	0
		parametrisch	0	0	0
$\mathbf{R_{YZ\|X}}$=0.35	A_{300}	gemischt	6	3	2
		nonparametrisch	2	1	0
		parametrisch	6	2	5
	A_{500}	gemischt	1	0	0
		nonparametrisch	1	0	0
		parametrisch	7	2	6
	$A \cup B_{500}$	gemischt	0	0	0
		nonparametrisch	0	0	0
		parametrisch	0	0	0

Tabelle C.1: Anzahl der abgelehnten Kolmogorov-Smirnov-Tests für $H_0 : f_{Z_r} = \tilde{f}_{Z_r}$

$\mathbf{R_{YZ\|X}}$	Szenario	Methode	Z_1	Z_2	Z_3
$\mathbf{R_{YZ\|X}}$=0	A_{300}	gemischt	0.06048	0.06055	0.05834
		nonparametrisch	0.05689	0.05727	0.05717
		parametrisch	0.06165	0.06151	0.05961
	A_{500}	gemischt	0.03941	0.03854	0.03936
		nonparametrisch	0.03941	0.03854	0.03936
		parametrisch	0.04718	0.04686	0.04662
	$A \cup B_{500}$	gemischt	0.01971	0.01927	0.01968
		nonparametrisch	0.01971	0.01927	0.01968
		parametrisch	0.02359	0.02343	0.02331
$\mathbf{R_{YZ\|X}}$=0.2	A_{300}	gemischt	0.05992	0.06148	0.05894
		nonparametrisch	0.05801	0.05744	0.05767
		parametrisch	0.06047	0.06260	0.05987
	A_{500}	gemischt	0.03951	0.03860	0.03923
		nonparametrisch	0.03951	0.03860	0.03923
		parametrisch	0.04734	0.04831	0.04736
	$A \cup B_{500}$	gemischt	0.01975	0.01930	0.01962
		nonparametrisch	0.01975	0.01930	0.01962
		parametrisch	0.02367	0.02415	0.02368
$\mathbf{R_{YZ\|X}}$=0.35	A_{300}	gemischt	0.06128	0.06075	0.05769
		nonparametrisch	0.05831	0.05788	0.05612
		parametrisch	0.06225	0.06137	0.05881
	A_{500}	gemischt	0.03992	0.03924	0.03924
		nonparametrisch	0.03992	0.03924	0.03924
		parametrisch	0.04670	0.04797	0.04645
	$A \cup B_{500}$	gemischt	0.01996	0.01962	0.01962
		nonparametrisch	0.01996	0.01962	0.01962
		parametrisch	0.02335	0.02399	0.02323

Tabelle C.2: Durchschnittliche Kolmogorov-Smirnov-Distanzen zwischen f_{Z_r} und $\tilde{f}_{Z_r}$

		KS-Test			KS-Distanz		
$\mathbf{R_{YZ\|X}}$	Methode	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
$\mathbf{R_{YZ\|X}} = 0$	gemischt	0	0	0	0.01965	0.01945	0.01962
	nonparametrisch	0	0	0	0.01965	0.01945	0.01962
	parametrisch	0	0	0	0.02416	0.02420	0.02366
$\mathbf{R_{YZ\|X}} = 0.2$	gemischt	0	0	0	0.01968	0.01974	0.01950
	nonparametrisch	0	0	0	0.01968	0.01974	0.01950
	parametrisch	0	0	0	0.02386	0.02368	0.02311
$\mathbf{R_{YZ\|X}} = 0.35$	gemischt	0	0	0	0.01952	0.01977	0.01956
	nonparametrisch	0	0	0	0.01952	0.01977	0.01956
	parametrisch	0	0	0	0.02387	0.02383	0.02309

Tabelle C.3: Anzahl abgelehnter Kolmogorov-Smirnov-Tests und die durchschnittlichen KS-Distanzen für $\tilde{f}_{Y_q}$. Diese wurden nur im Szenario $A \cup B_{500}$ beobachtet.

Szenario	Methode	$r_{Z_1Z_2}$	$r_{Z_1Z_3}$	$r_{Z_2Z_3}$
A_{300}	gemischt	0.00056	0.01411	0.0043
		0.00291	0.00257	0.00318
	nonparametrisch	-0.00278	-0.00943	0.00082
		0.00194	0.0018	0.00197
	parametrisch	-0.00003	0.00286	-0.00174
		0.00340	0.00278	0.00369
A_{500}	gemischt	0	0	0
		0	0	0
	nonparametrisch	0	0	0
		0	0	0
	parametrisch	0.00038	-0.00050	-0.00154
		0.00195	0.00186	0.00193
$A \cup B_{500}$	gemischt	0	0	0
		0	0	0
	nonparametrisch	0	0	0
		0	0	0
	parametrisch	0.00026	-0.00005	-0.00077
		0.00048	0.00046	0.00049

Tabelle C.4: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Z_rZ_s}$ mit $r \neq s$ und $r, s = 1, 2, 3$ für $\mathbf{R_{YZ|X}} = 0$

Szenario	Methode	$r_{Z_1Z_2}$	$r_{Z_1Z_3}$	$r_{Z_2Z_3}$
A_{300}	gemischt	0.00054	0.01295	0.00659
		0.00278	0.00252	0.0032
	nonparametrisch	-0.00033	-0.00887	-0.00396
		0.00218	0.00196	0.00199
	parametrisch	0.00022	0.00284	-0.00074
		0.00337	0.00294	0.00370
A_{300}	gemischt	0	0	0
		0	0	0
	nonparametrisch	0	0	0
		0	0	0
	parametrisch	-0.00006	0.00013	0.00030
		0.00197	0.00177	0.00197
$A \cup B_{500}$	gemischt	0	0	0
		0	0	0
	nonparametrisch	0	0	0
		0	0	0
	parametrisch	0.00050	0.00002	0.00014
		0.00049	0.00044	0.00049

Tabelle C.5: Bias (erste Zeile) und MSE (zweite Zeile) von $\tilde{r}_{Z_rZ_s}$ mit $r \neq s$ und $r, s = 1, 2, 3$ für $\mathbf{R_{YZ|X}} = 0.2$

Szenario	Methode	$r_{Z_1Z_2}$	$r_{Z_1Z_3}$	$r_{Z_2Z_3}$
A_{300}	gemischt	0.00129	0.01396	0.0052
		0.00304	0.00271	0.00306
	nonparametrisch	-0.00013	-0.00781	0.00060
		0.00191	0.00165	0.00200
	parametrisch	-0.00054	0.00299	-0.00152
		0.00339	0.00287	0.00365
A_{500}	gemischt	0	0	0
		0	0	0
	nonparametrisch	0	0	0
		0	0	0
	parametrisch	-0.00010	-0.00031	-0.00021
		0.00201	0.00174	0.00199
$A \cup B_{500}$	gemischt	0	0	0
		0	0	0
	nonparametrisch	0	0	0
		0	0	0
	parametrisch	0.00003	-0.00019	-0.00012
		0.00050	0.00043	0.00050

Tabelle C.6: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Z_rZ_s}$ mit $r \neq s$ und $r, s = 1, 2, 3$ für $\mathbf{R_{YZ|X}} = 0.35$

C.1.2 Erhalt von $f_{\mathbf{XZ}}$ bzw. $f_{\mathbf{XY}}$

Tabelle C.7: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{X_p Z_r}$ für $\mathbf{R_{YZ|X}} = 0$

Szenario	Methode	$r_{X_1Z_1}$	$r_{X_1Z_2}$	$r_{X_1Z_3}$	$r_{X_2Z_1}$	$r_{X_2Z_2}$	$r_{X_2Z_3}$	$r_{X_3Z_1}$	$r_{X_3Z_2}$	$r_{X_3Z_3}$
A_{300}	gemischt	-0.00297	0.00109	-0.00442	-0.00012	0.00166	-0.00539	-0.00173	-0.00022	-0.00328
		0.00286	0.00340	0.00179	0.00295	0.00323	0.00224	0.00342	0.00364	0.00275
	nonparametrisch	-0.01639	0.00228	-0.02048	-0.00557	-0.00330	-0.01305	0.00088	0.00023	-0.00243
		0.00209	0.00189	0.00166	0.00188	0.00201	0.00160	0.00196	0.00197	0.00148
	parametrisch	-0.00243	-0.00084	0.00245	0.00006	0.00209	-0.00060	-0.00134	-0.00030	-0.00050
		0.00299	0.00347	0.00177	0.00297	0.00330	0.00220	0.00340	0.00358	0.00272
A_{500}	gemischt	-0.00644	0.00103	-0.01471	-0.00480	-0.00186	-0.01279	0.00008	0.00217	-0.00603
		0.00163	0.00149	0.00126	0.00170	0.00171	0.00143	0.00175	0.00184	0.00151
	nonparametrisch	-0.00862	0.00040	-0.01597	-0.00479	-0.00231	-0.00993	-0.00040	-0.00014	0.00005
		0.00023	0.00014	0.00037	0.00017	0.00016	0.00022	0.00015	0.00014	0.00014
	parametrisch	-0.00070	-0.00103	0.00059	-0.00053	-0.00038	-0.00299	0.00091	0.00190	-0.00254
		0.00182	0.00194	0.00102	0.00181	0.00181	0.00128	0.00175	0.00182	0.00145
$A \cup B_{500}$	gemischt	-0.00322	0.00051	-0.00736	-0.00240	-0.00093	-0.00639	0.00004	0.00109	-0.00301
		0.00041	0.00037	0.00032	0.00043	0.00043	0.00036	0.00044	0.00046	0.00038
	nonparametrisch	-0.00431	0.00020	-0.00798	-0.00239	-0.00116	-0.00496	-0.00020	-0.00007	0.00003
		0.00006	0.00003	0.00009	0.00004	0.00004	0.00006	0.00004	0.00004	0.00003
	parametrisch	-0.00037	-0.00051	0.00022	-0.00028	-0.00021	-0.00155	0.00043	0.00093	-0.00128
		0.00045	0.00048	0.00025	0.00045	0.00045	0.00032	0.00044	0.00045	0.00036

Tabelle C.8: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{X_p Z_r}$ für $\mathbf{R_{YZ|X}} = 0.2$

Szenario	Methode	$r_{X_1Z_1}$	$r_{X_1Z_2}$	$r_{X_1Z_3}$	$r_{X_2Z_1}$	$r_{X_2Z_2}$	$r_{X_2Z_3}$	$r_{X_3Z_1}$	$r_{X_3Z_2}$	$r_{X_3Z_3}$
A_{300}	gemischt	-0.00126	0.00514	-0.00791	-0.00130	-0.00076	-0.00087	-0.00010	0.00012	-0.00153
		0.00271	0.00298	0.00200	0.00314	0.00342	0.00198	0.00316	0.00358	0.00233
	nonparametrisch	-0.01261	-0.00098	-0.02234	-0.00731	-0.00379	-0.01459	-0.00371	0.00130	-0.00064
		0.00205	0.00237	0.00173	0.00198	0.00193	0.00165	0.00191	0.00229	0.00153
	parametrisch	0.00027	0.00222	-0.00080	-0.00158	-0.00151	0.00341	-0.00048	-0.00065	-0.00037
		0.00279	0.00314	0.00185	0.00320	0.00341	0.00201	0.00307	0.00343	0.00231
A_{500}	gemischt	-0.00651	0.00084	-0.01921	-0.00465	-0.00235	-0.01074	-0.00289	0.00027	-0.00557
		0.00152	0.00153	0.00144	0.00188	0.00195	0.00131	0.00196	0.00199	0.00149
	nonparametrisch	-0.00946	0.00019	-0.01583	-0.00595	-0.00153	-0.01040	0.00053	0.00063	-0.00124
		0.00023	0.00014	0.00037	0.00016	0.00014	0.00025	0.00013	0.00014	0.00014
	parametrisch	0.00003	-0.00044	-0.00437	-0.00101	-0.00113	0.00002	-0.00163	0.00060	-0.00125
		0.00163	0.00187	0.00106	0.00201	0.00210	0.00107	0.00195	0.00203	0.00140
$A \cup B_{500}$	gemischt	-0.00326	0.00042	-0.00961	-0.00233	-0.00118	-0.00537	-0.00144	0.00014	-0.00278
		0.00038	0.00038	0.00036	0.00047	0.00049	0.00033	0.00049	0.00050	0.00037
	nonparametrisch	-0.00473	0.00010	-0.00792	-0.00297	-0.00076	-0.00520	0.00026	0.00031	-0.00062
		0.00006	0.00003	0.00009	0.00004	0.00004	0.00006	0.00003	0.00004	0.00004
	parametrisch	-0.00001	-0.00022	-0.00226	-0.00051	-0.00060	-0.00004	-0.00084	0.00029	-0.00065
		0.00041	0.00047	0.00026	0.00050	0.00052	0.00027	0.00049	0.00051	0.00035

Tabelle C.9: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{X_p Z_r}$ für $\mathbf{R_{YZ|X}} = 0.35$

Szenario	Methode	$r_{X_1Z_1}$	$r_{X_1Z_2}$	$r_{X_1Z_3}$	$r_{X_2Z_1}$	$r_{X_2Z_2}$	$r_{X_2Z_3}$	$r_{X_3Z_1}$	$r_{X_3Z_2}$	$r_{X_3Z_3}$
A_{300}	gemischt	-0.00088	0.00416	-0.00696	-0.00211	-0.00213	-0.00280	-0.00074	0.00014	-0.00066
		0.00244	0.00307	0.00187	0.00298	0.00330	0.00215	0.00319	0.00321	0.00259
	nonparametrisch	-0.01161	-0.00284	-0.02135	-0.00762	-0.00267	-0.01368	-0.00364	-0.00031	-0.00267
		0.00181	0.00208	0.00168	0.00191	0.00216	0.00163	0.00183	0.00202	0.00161
	parametrisch	0.00034	0.00168	-0.00025	-0.00114	-0.00241	0.00243	-0.00112	-0.00015	0.00070
		0.00247	0.00315	0.00176	0.00297	0.00343	0.00205	0.00313	0.00331	0.00254
A_{500}	nonparametrisch	-0.00632	0.00022	-0.01932	-0.00483	-0.00139	-0.00969	-0.00227	0.00030	-0.00446
		0.00157	0.00164	0.00141	0.00178	0.00177	0.00119	0.00192	0.00186	0.00147
	nonparametrisch	-0.00910	-0.00078	-0.01602	-0.00435	-0.00206	-0.01033	-0.00046	0.00055	-0.00094
		0.00021	0.00014	0.00039	0.00015	0.00013	0.00024	0.00015	0.00013	0.00014
	parametrisch	-0.00015	-0.00063	-0.00432	-0.00154	-0.00117	0.00060	-0.00134	0.00052	-0.00116
		0.00167	0.00193	0.00105	0.00194	0.00192	0.00105	0.00200	0.00197	0.00142
$A \cup B_{500}$	nonparametrisch	-0.00316	0.00011	-0.00966	-0.00241	-0.00069	-0.00484	-0.00114	0.00015	-0.00223
		0.00039	0.00041	0.00035	0.00045	0.00044	0.00030	0.00048	0.00047	0.00037
	nonparametrisch	-0.00455	-0.00039	-0.00801	-0.00218	-0.00103	-0.00517	-0.00023	0.00027	-0.00047
		0.00005	0.00004	0.00010	0.00004	0.00003	0.00006	0.00004	0.00003	0.00003
	parametrisch	-0.00010	-0.00031	-0.00224	-0.00079	-0.00061	0.00025	-0.00070	0.00025	-0.00061
		0.00042	0.00048	0.00026	0.00048	0.00048	0.00026	0.00050	0.00049	0.00035

Tabelle C.10: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{X_pY_q}$ für das Szenario $A \cup B_{500}$

bed. Korrelation	Methode	$r_{X_1Y_1}$	$r_{X_1Y_2}$	$r_{X_1Y_3}$	$r_{X_2Y_1}$	$r_{X_2Y_2}$	$r_{X_2Y_3}$	$r_{X_3Y_1}$	$r_{X_3Y_2}$	$r_{X_3Y_3}$
$\mathbf{R_{YZ\|X}} = 0$	gemischt	-0.00366	-0.00052	-0.00840	-0.00415	-0.00127	-0.00567	-0.00272	-0.00054	-0.00441
		0.00034	0.00039	0.00030	0.00039	0.00047	0.00033	0.00042	0.00046	0.00034
	nonparametrisch	-0.00539	-0.00140	-0.00890	-0.00382	-0.00124	-0.00622	-0.00080	-0.00046	-0.00137
		0.00006	0.00004	0.00011	0.00005	0.00004	0.00007	0.00003	0.00004	0.00003
	parametrisch	0.00011	0.00134	0.00014	-0.00125	0.00036	0.00031	-0.00145	0.00050	-0.00163
		0.00038	0.00049	0.00021	0.00040	0.00051	0.00029	0.00041	0.00045	0.00031
$\mathbf{R_{YZ\|X}} = 0.2$	gemischt	-0.00421	-0.00012	-0.00848	-0.00357	-0.00171	-0.00497	-0.00227	-0.00126	-0.00307
		0.00037	0.00039	0.00027	0.00043	0.00045	0.00028	0.00041	0.00049	0.00035
	nonparametrisch	-0.00552	-0.00176	-0.00902	-0.00324	-0.00092	-0.00610	-0.00119	-0.00058	-0.00111
		0.00006	0.00004	0.00011	0.00005	0.00004	0.00007	0.00004	0.00004	0.00004
	parametrisch	-0.00066	0.00178	-0.00004	-0.00102	-0.00035	0.00093	-0.00097	-0.00037	-0.00041
		0.00039	0.00046	0.00019	0.00044	0.00047	0.00025	0.00043	0.00049	0.00032
$\mathbf{R_{YZ\|X}} = 0.35$	gemischt	-0.00384	0.00049	-0.00879	-0.00379	-0.00163	-0.00539	-0.00073	-0.00051	-0.00309
		0.00036	0.00037	0.00026	0.00039	0.00047	0.00027	0.00043	0.00049	0.00033
	nonparametrisch	-0.00488	-0.00175	-0.00870	-0.00319	-0.00121	-0.00600	-0.00105	-0.00082	-0.00130
		0.00006	0.00004	0.00010	0.00004	0.00004	0.00007	0.00004	0.00003	0.00004
	parametrisch	-0.00022	0.00204	-0.00034	-0.00123	-0.00044	0.00068	0.00028	0.00051	-0.00079
		0.00039	0.00043	0.00018	0.00042	0.00053	0.00025	0.00042	0.00050	0.00030

C.2 Ergebnisse der dritten Validitätsstufe

C.2.1 Erhalt von $\mathbf{R_{YZ}}$

Tabelle C.11: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Y_qZ_r}$ für $\mathbf{R_{YZ|X}} = 0$

Szenario	Methode	$r_{Y_1Z_1}$	$r_{Y_1Z_2}$	$r_{Y_1Z_3}$	$r_{Y_2Z_1}$	$r_{Y_2Z_2}$	$r_{Y_2Z_3}$	$r_{Y_3Z_1}$	$r_{Y_3Z_2}$	$r_{Y_3Z_3}$
A_{300}	gemischt	-0.00379	0.00419	-0.00039	-0.00067	0.00028	-0.00204	-0.00413	-0.00067	-0.00405
		0.00269	0.00288	0.00231	0.00309	0.00322	0.00231	0.00301	0.00352	0.00202
	nonparametrisch	-0.00357	0.00442	-0.00801	-0.00120	0.00113	-0.00399	-0.01172	0.00181	-0.01615
		0.00314	0.00274	0.00237	0.00298	0.00301	0.00261	0.00285	0.00276	0.00239
	parametrisch	-0.00309	0.00379	0.00341	-0.00071	0.00049	-0.00130	-0.00418	-0.00111	0.00127
		0.00275	0.00298	0.00230	0.00317	0.00331	0.00223	0.00307	0.00367	0.00201
A_{500}	gemischt	-0.00251	-0.00175	-0.00823	-0.00132	-0.00103	-0.00386	-0.00569	0.00079	-0.01145
		0.00179	0.00194	0.00145	0.00186	0.00189	0.00152	0.00179	0.00172	0.00146
	nonparametrisch	-0.00527	-0.00030	-0.00747	-0.00148	0.00073	-0.00038	-0.00571	-0.00198	-0.00975
		0.00164	0.00191	0.00142	0.00177	0.00224	0.00140	0.00111	0.00138	0.00107
	parametrisch	0.00025	-0.00172	-0.00202	-0.00003	-0.00102	-0.00169	-0.00130	-0.00054	-0.00130
		0.00181	0.00204	0.00135	0.00188	0.00188	0.00147	0.00189	0.00185	0.00128
$A \cup B_{500}$	gemischt	-0.00185	-0.00147	-0.00848	-0.00004	-0.00163	-0.00281	-0.00677	0.00176	-0.01111
		0.00086	0.00101	0.00083	0.00095	0.00095	0.00083	0.00076	0.00077	0.00079
	nonparametrisch	-0.00527	-0.00030	-0.00747	-0.00148	0.00073	-0.00038	-0.00571	-0.00198	-0.00975
		0.00164	0.00191	0.00142	0.00177	0.00224	0.00140	0.00111	0.00138	0.00107
	parametrisch	0.00106	-0.00105	-0.00309	0.00130	-0.00108	-0.00020	-0.00159	0.00149	-0.00081
		0.00086	0.00099	0.00076	0.00097	0.00094	0.00089	0.00075	0.00080	0.00060

Tabelle C.12: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Y_q Z_r}$ für $\mathbf{R_{YZ|X}} = 0.2$

Szenario	Methode	$r_{Y_1Z_1}$	$r_{Y_1Z_2}$	$r_{Y_1Z_3}$	$r_{Y_2Z_1}$	$r_{Y_2Z_2}$	$r_{Y_2Z_3}$	$r_{Y_3Z_1}$	$r_{Y_3Z_2}$	$r_{Y_3Z_3}$
A_{300}	gemischt	-0.18290	-0.18606	-0.16057	-0.19005	-0.19682	-0.16853	-0.15662	-0.15896	-0.14070
		0.03676	0.03811	0.02857	0.03890	0.04251	0.03092	0.02749	0.02861	0.02177
	nonparametrisch	-0.18248	-0.18633	-0.16952	-0.19582	-0.19779	-0.17308	-0.15990	-0.16195	-0.15124
		0.03656	0.03768	0.03107	0.04148	0.04249	0.03254	0.02812	0.02929	0.02482
	parametrisch	-0.18259	-0.18646	-0.15758	-0.19054	-0.19645	-0.16737	-0.15573	-0.16123	-0.13603
		0.03668	0.03829	0.02758	0.03905	0.04234	0.03053	0.02715	0.02926	0.02048
A_{500}	gemischt	-0.18239	-0.18698	-0.16786	-0.19027	-0.19908	-0.17219	-0.16041	-0.16108	-0.15026
		0.03504	0.03686	0.02964	0.03824	0.04139	0.03127	0.02745	0.02767	0.02396
	nonparametrisch	-0.18532	-0.18826	-0.16696	-0.19226	-0.20059	-0.16844	-0.15782	-0.16085	-0.14680
		0.03597	0.03753	0.02918	0.03855	0.04225	0.02998	0.02608	0.02728	0.02242
	parametrisch	-0.17974	-0.18663	-0.16119	-0.18981	-0.19846	-0.16959	-0.15535	-0.16126	-0.13919
		0.03403	0.03681	0.02738	0.03803	0.04115	0.03035	0.02590	0.02798	0.02066
$A \cup B_{500}$	gemischt	-0.18394	-0.18587	-0.16495	-0.19036	-0.19809	-0.17012	-0.16113	-0.15982	-0.14760
		0.03467	0.03549	0.02796	0.03715	0.04017	0.02970	0.02665	0.02625	0.02246
	nonparametrisch	-0.18532	-0.18826	-0.16696	-0.19226	-0.20059	-0.16844	-0.15782	-0.16085	-0.14680
		0.03597	0.03753	0.02918	0.03855	0.04225	0.02998	0.02608	0.02728	0.02242
	parametrisch	-0.18077	-0.18559	-0.15909	-0.18932	-0.19766	-0.16777	-0.15619	-0.16001	-0.13712
		0.03355	0.03538	0.02610	0.03675	0.04000	0.02898	0.02510	0.02634	0.01943

Tabelle C.13: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Y_q Z_r}$ für $\mathbf{R_{YZ|X}} = 0.35$

Szenario	Methode	$r_{Y_1Z_1}$	$r_{Y_1Z_2}$	$r_{Y_1Z_3}$	$r_{Y_2Z_1}$	$r_{Y_2Z_2}$	$r_{Y_2Z_3}$	$r_{Y_3Z_1}$	$r_{Y_3Z_2}$	$r_{Y_3Z_3}$
A_{300}	gemischt	-0.31628	-0.32881	-0.28215	-0.33295	-0.34651	-0.29728	-0.27430	-0.27613	-0.23922
		0.10349	0.11191	0.08221	0.11405	0.12339	0.09078	0.07803	0.07930	0.05942
	nonparametrisch	-0.31717	-0.32880	-0.28394	-0.33529	-0.34324	-0.29706	-0.27795	-0.28536	-0.25345
		0.10354	0.11131	0.08297	0.11527	0.12079	0.09073	0.07991	0.08430	0.06637
	parametrisch	-0.31589	-0.32904	-0.27840	-0.33265	-0.34621	-0.29667	-0.27338	-0.27705	-0.23350
		0.10326	0.11209	0.08010	0.11379	0.12324	0.09047	0.07747	0.07985	0.05664
A_{500}	gemischt	-0.31726	-0.32508	-0.28609	-0.33405	-0.34603	-0.30087	-0.27324	-0.28101	-0.25110
		0.10242	0.10761	0.08351	0.11344	0.12156	0.09189	0.07632	0.08088	0.06428
	nonparametrisch	-0.31561	-0.32712	-0.28670	-0.33274	-0.34524	-0.29953	-0.27119	-0.28319	-0.25023
		0.10125	0.10889	0.08373	0.11250	0.12118	0.09141	0.07470	0.08144	0.06365
	parametrisch	-0.31452	-0.32614	-0.28026	-0.33338	-0.34643	-0.29835	-0.26906	-0.28163	-0.24025
		0.10065	0.10841	0.08016	0.11301	0.12183	0.09040	0.07416	0.08135	0.05892
$A \cup B_{500}$	gemischt	-0.31737	-0.32640	-0.28443	-0.33246	-0.34635	-0.29757	-0.27508	-0.27878	-0.24938
		0.10159	0.10744	0.08169	0.11149	0.12092	0.08927	0.07638	0.07852	0.06278
	nonparametrisch	-0.31561	-0.32712	-0.28670	-0.33274	-0.34524	-0.29953	-0.27119	-0.28319	-0.25023
		0.10125	0.10889	0.08373	0.11250	0.12118	0.09141	0.07470	0.08144	0.06365
	parametrisch	-0.31464	-0.32690	-0.27951	-0.33151	-0.34639	-0.29547	-0.27072	-0.27929	-0.23870
		0.09987	0.10783	0.07892	0.11087	0.12098	0.08806	0.07401	0.07881	0.05756

C.2.2 Erhalt von $\mathbf{R_{YZ|X}}$

Tabelle C.14: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Y_qZ_r|\mathbf{X}}$ im Original und im fusionierten Datensatz für $\mathbf{R_{YZ|X}} = 0$

Szenario	Methode	$r_{Y_1Z_1\|\mathbf{X}}$	$r_{Y_1Z_2\|\mathbf{X}}$	$r_{Y_1Z_3\|\mathbf{X}}$	$r_{Y_2Z_1\|\mathbf{X}}$	$r_{Y_2Z_2\|\mathbf{X}}$	$r_{Y_2Z_3\|\mathbf{X}}$	$r_{Y_3Z_1\|\mathbf{X}}$	$r_{Y_3Z_2\|\mathbf{X}}$	$r_{Y_3Z_3\|\mathbf{X}}$
A_{300}	gemischt	-0.00323	0.00386	0.00244	-0.00033	0.00007	-0.00130	-0.00340	-0.00212	-0.00007
		0.00303	0.00295	0.00344	0.00335	0.00330	0.00315	0.00361	0.00351	0.00316
	nonparametrisch	0.00233	0.00466	0.00040	0.00068	0.00112	-0.00137	-0.00275	0.00219	-0.00298
		0.00361	0.00302	0.00333	0.00328	0.00309	0.00352	0.00344	0.00329	0.00368
	parametrisch	-0.00268	0.00395	0.00358	-0.00046	0.00044	-0.00165	-0.00389	-0.00166	0.00044
		0.00310	0.00304	0.00342	0.00346	0.00339	0.00309	0.00356	0.00352	0.00329
A_{500}	gemischt	0.00019	-0.00195	-0.00188	-0.00049	-0.00110	-0.00173	-0.00140	0.00098	-0.00003
		0.00208	0.00200	0.00207	0.00207	0.00190	0.00203	0.00202	0.00189	0.00211
	nonparametrisch	-0.00212	0.00000	-0.00141	-0.00038	0.00085	0.00201	-0.00005	-0.00179	0.00150
		0.00194	0.00214	0.00213	0.00194	0.00229	0.00192	0.00175	0.00208	0.00203
	parametrisch	0.00053	-0.00155	-0.00201	-0.00002	-0.00097	-0.00173	-0.00110	-0.00003	-0.00074
		0.00205	0.00201	0.00206	0.00211	0.00186	0.00204	0.00208	0.00186	0.00209
$A \cup B_{500}$	gemischt	0.00104	-0.00139	-0.00293	0.00072	-0.00163	-0.00121	-0.00213	0.00255	-0.00029
		0.00100	0.00110	0.00105	0.00099	0.00096	0.00099	0.00089	0.00100	0.00107
	nonparametrisch	-0.00198	0.00006	-0.00116	-0.00044	0.00086	0.00191	0.00016	-0.00169	0.00171
		0.00196	0.00215	0.00210	0.00195	0.00229	0.00188	0.00172	0.00205	0.00198
	parametrisch	0.00150	-0.00093	-0.00313	0.00098	-0.00113	-0.00090	-0.00182	0.00218	-0.00072
		0.00098	0.00107	0.00105	0.00100	0.00095	0.00104	0.00094	0.00103	0.00098

Tabelle C.15: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Y_q Z_r|\mathbf{X}}$ im Original und im fusionierten Datensatz für $\mathbf{R_{YZ}}|\mathbf{X} = 0.2$

Szenario	Methode	$r_{Y_1Z_1\|\mathbf{X}}$	$r_{Y_1Z_2\|\mathbf{X}}$	$r_{Y_1Z_3\|\mathbf{X}}$	$r_{Y_2Z_1\|\mathbf{X}}$	$r_{Y_2Z_2\|\mathbf{X}}$	$r_{Y_2Z_3\|\mathbf{X}}$	$r_{Y_3Z_1\|\mathbf{X}}$	$r_{Y_3Z_2\|\mathbf{X}}$	$r_{Y_3Z_3\|\mathbf{X}}$
A_{300}	gemischt	-0.20344	-0.20074	-0.19919	-0.19972	-0.19903	-0.19922	-0.20279	-0.20154	-0.20101
		0.04494	0.04372	0.04365	0.04304	0.04346	0.04321	0.04461	0.04386	0.04359
	nonparametrisch	-0.19800	-0.19862	-0.20236	-0.20403	-0.19928	-0.20187	-0.19719	-0.20049	-0.20006
		0.04293	0.04259	0.04445	0.04504	0.04311	0.04422	0.04202	0.04365	0.04359
	parametrisch	-0.20351	-0.20010	-0.19898	-0.20036	-0.19832	-0.19900	-0.20248	-0.20230	-0.20121
		0.04499	0.04353	0.04354	0.04325	0.04316	0.04316	0.04448	0.04417	0.04368
A_{500}	gemischt	-0.20045	-0.20011	-0.20202	-0.19898	-0.20083	-0.20142	-0.20283	-0.20100	-0.20209
		0.04222	0.04195	0.04301	0.04180	0.04212	0.04280	0.04324	0.04218	0.04301
	nonparametrisch	-0.20266	-0.20146	-0.20230	-0.20085	-0.20234	-0.19752	-0.19732	-0.20068	-0.19992
		0.04307	0.04296	0.04291	0.04209	0.04299	0.04123	0.04083	0.04238	0.04172
	parametrisch	-0.20037	-0.19959	-0.20166	-0.19943	-0.20016	-0.20092	-0.20193	-0.20098	-0.20169
		0.04210	0.04172	0.04279	0.04195	0.04185	0.04260	0.04288	0.04226	0.04277
$A \cup B_{500}$	gemischt	-0.20204	-0.19883	-0.19961	-0.19931	-0.19983	-0.19986	-0.20327	-0.19933	-0.20033
		0.04180	0.04056	0.04089	0.04072	0.04088	0.04094	0.04228	0.04059	0.04123
	nonparametrisch	-0.20277	-0.20137	-0.20225	-0.20092	-0.20231	-0.19756	-0.19760	-0.20049	-0.19979
		0.04310	0.04292	0.04287	0.04211	0.04298	0.04124	0.04091	0.04229	0.04165
	parametrisch	-0.20127	-0.19857	-0.19898	-0.19934	-0.19943	-0.19980	-0.20339	-0.19981	-0.20047
		0.04151	0.04042	0.04069	0.04073	0.04072	0.04098	0.04236	0.04080	0.04124

Tabelle C.16: Bias (erste Zeile) und MSE (zweite Zeile) je Methode von $\tilde{r}_{Y_q Z_r|\mathbf{X}}$ im Original und im fusionierten Datensatz für $\mathbf{R_{YZ}}|\mathbf{X} = 0.35$

Szenario	Methode	$r_{Y_1Z_1\|\mathbf{X}}$	$r_{Y_1Z_2\|\mathbf{X}}$	$r_{Y_1Z_3\|\mathbf{X}}$	$r_{Y_2Z_1\|\mathbf{X}}$	$r_{Y_2Z_2\|\mathbf{X}}$	$r_{Y_2Z_3\|\mathbf{X}}$	$r_{Y_3Z_1\|\mathbf{X}}$	$r_{Y_3Z_2\|\mathbf{X}}$	$r_{Y_3Z_3\|\mathbf{X}}$
A_{300}	gemischt	-0.35221	-0.35309	-0.35237	-0.34989	-0.34990	-0.35209	-0.35591	-0.34693	-0.34581
		0.12771	0.12849	0.12782	0.12599	0.12579	0.12736	0.12983	0.12323	0.12327
	nonparametrisch	-0.34864	-0.35081	-0.34709	-0.35080	-0.34587	-0.34929	-0.35154	-0.35404	-0.35130
		0.12499	0.12647	0.12396	0.12618	0.12261	0.12537	0.12720	0.12883	0.12708
	parametrisch	-0.35236	-0.35251	-0.35135	-0.34975	-0.34933	-0.35254	-0.35586	-0.34646	-0.34467
		0.12789	0.12813	0.12714	0.12582	0.12543	0.12778	0.12975	0.12293	0.12244
A_{500}	gemischt	-0.35106	-0.34802	-0.35126	-0.35024	-0.34907	-0.35419	-0.34991	-0.35098	-0.35104
		0.12529	0.12307	0.12582	0.12468	0.12370	0.12731	0.12436	0.12533	0.12523
	nonparametrisch	-0.34849	-0.34979	-0.35319	-0.34866	-0.34816	-0.35308	-0.34586	-0.35285	-0.35205
		0.12346	0.12449	0.12706	0.12352	0.12323	0.12698	0.12152	0.12642	0.12609
	parametrisch	-0.35068	-0.34895	-0.35188	-0.35036	-0.34942	-0.35378	-0.34974	-0.35134	-0.35074
		0.12493	0.12372	0.12627	0.12478	0.12392	0.12709	0.12424	0.12560	0.12505
$A \cup B_{500}$	gemischt	-0.35105	-0.34933	-0.35051	-0.34892	-0.34939	-0.35156	-0.35166	-0.34807	-0.35001
		0.12423	0.12302	0.12400	0.12279	0.12305	0.12452	0.12462	0.12214	0.12351
	nonparametrisch	-0.34860	-0.34992	-0.35353	-0.34864	-0.34819	-0.35309	-0.34586	-0.35303	-0.35217
		0.12353	0.12458	0.12728	0.12350	0.12326	0.12697	0.12149	0.12651	0.12609
	parametrisch	-0.35072	-0.34989	-0.35107	-0.34891	-0.34945	-0.35152	-0.35234	-0.34891	-0.34983
		0.12398	0.12343	0.12439	0.12278	0.12312	0.12450	0.12508	0.12275	0.12339

D Energy-Test

Der Energy-Test wurde ursprünglich als Kennzahl für die zweite Validitätsstufe verwendet. Da dessen Verwendung aus den in Kapitel 3.3.2.2 genannten Gründen wieder verworfen wurde, werden die bereits berechneten Ergebnisse im Anhang aufgeführt und diskutiert.

D.1 Teststatistik und Funktionsweise

Anhand des Energy-Tests kann die Nullhypothese $H_0 : f_{\mathbf{XYZ}} = \tilde{f}_{\mathbf{XYZ}}$ überprüft werden. Dabei wird die Teststatistik

$$T = \frac{n}{2} E_n$$

mit

$$E_n = 2\mathbb{E}\|\mathbf{W} - \tilde{\mathbf{W}}\| - \mathbb{E}\|\tilde{\mathbf{W}} - \tilde{\mathbf{W}}\| - \mathbb{E}\|\mathbf{W} - \mathbf{W}\|$$

verwendet. Es gilt $\mathbf{W} = (\mathbf{X}, \mathbf{Y}, \mathbf{Z})$; zudem misst $\|\cdot\|$ die euklidische Distanz.
Die Nullhypothese wird abgelehnt, wenn $E_n > c_\alpha$, wobei $c_\alpha = \lim_{n\to\infty} P(E_n > c_\alpha) = \alpha$ (Székely und Rizzo, 2004). α wurde in der Simulation mit 0.05 festgelegt.
Eine Implementierung des Tests in R bietet das Paket `energy` (Rizzo und Szekely, 2016). In diesem wird anhand eines Permutationsverfahrens eine approximative Teststatistik mit B (hier mit B=1000) zufälligen Permutationen berechnet.
Zusätzlich wurde als weitere Kennzahl der Energy-Koeffizient der Inhomogenität (siehe Székely, 2002) verwendet, um die einzelnen Teststatistiken auch für eine unterschiedliche Anzahl an Beobachtungen vergleichbar zu machen. Dieser wird anhand folgender Formel berechnet:

$$e = \frac{E_n}{2\mathbb{E}\|\mathbf{W} - \tilde{\mathbf{W}}\|}$$

Für diesen gilt $e \in [0, 1]$, wobei ein Wert von 0 bedeutet, dass $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$ gleich sind.

D.2 Ergebnisse der Simulation

Als Kennzahlen für die zweite Validitätsstufe wurden ursprünglich die Anzahl abgelehnter Energy-Tests sowie der durchschnittliche Wert des Energy-Koeffizienten berechnet. Diese sind in den Tabellen D.1 und D.2 abgebildet.

Szenario	Methode	$\mathbf{R_{YZ\|X}} = 0$	$\mathbf{R_{YZ\|X}} = 0.2$	$\mathbf{R_{YZ\|X}} = 0.35$
A_{300}	gemischt	0	0	0
	nonparametrisch	0	0	0
	parametrisch	0	0	0
A_{500}	gemischt	0	0	0
	nonparametrisch	0	0	0
	parametrisch	0	0	0
$A \cup B_{500}$	gemischt	0	0	290
	nonparametrisch	0	0	420
	parametrisch	0	0	409

Tabelle D.1: Anzahl abgelehnter Energytests für $H_0 = \tilde{f}_{\mathbf{XYZ}} = f_{\mathbf{XYZ}}$

Abgelehnte Tests konnten nur im Szenario $A \cup B_{500}$ für $\mathbf{R_{YZ|X}} = 0.35$ beobachtet werden. Von je 500 Tests werden dabei bei der gemischten Methode 58% der Tests abgelehnt, während es bei der nonparametrischen Methode bereits 84% sowie bei der parametrischen Methode 81.8% der Tests sind. Die Ergebnisse decken sich nicht mit den im Hauptteil der Arbeit berechneten Resultaten und legen den Schluss nahe, dass die Testpower vor allem für die Szenarien A_{300} und A_{500} aufgrund der geringeren Zahl an Beobachtungen nicht ausreichend ist. Zudem könnte die Bestrafung von Datensätzen mit einer größeren Zahl an Beobachtungen durch die Teststatistik eine Rolle spielen.

Der durchschnittliche Energy-Koeffizient (siehe Tabelle D.2) weist für alle Simulationsdurchläufe sehr geringe Werte auf, was im Kontext der Ergebnisse aus dem Hauptteil der Arbeit nicht wirklich sinnvoll erscheint. Auffallend ist jedoch, dass der Energy-Koeffizient für eine bedingte Korrelation von 0.35 höhere Werte für A_{500} und A_{300} als für $A \cup B_{500}$ aufweist, obwohl die Energy-Tests für diese Simulationsdurchläufe nicht abgelehnt werden konnten. Dies kann ebenfalls mit der zusätzlichen Bestrafung von Datensätzen mit einer höheren Anzahl an Beobachtungen durch die Teststatistik

Szenario	Methode	$\mathbf{R_{YZ|X}} = 0$	$\mathbf{R_{YZ|X}} = 0.2$	$\mathbf{R_{YZ|X}} = 0.35$
A_{300}	gemischt	0.00273	0.00343	0.00511
	nonparametrisch	0.00248	0.00325	0.00492
	parametrisch	0.00275	0.00346	0.00515
A_{500}	gemischt	0.00130	0.00204	0.00375
	nonparametrisch	0.00127	0.00201	0.00371
	parametrisch	0.00163	0.00236	0.00406
$A \cup B_{500}$	gemischt	0.00064	0.00138	0.00308
	nonparametrisch	0.00083	0.00156	0.00327
	parametrisch	0.00081	0.00153	0.00323

Tabelle D.2: Durchschnittlicher Energy-Koeffizient zwischen $f_{\mathbf{XYZ}}$ und $\tilde{f}_{\mathbf{XYZ}}$

begründet werden. Insgesamt kann die Verwendung des Energy-Tests und des Energy-Koeffizienten für die in dieser Arbeit verwendete Datensituation daher nicht empfohlen werden.

E Elektronischer Anhang

Im elektronischen Anhang befinden sich folgende Ordner und Dateien:

- Ordner Masterarbeit:
  enthält die vorgelegte Masterarbeit als pdf-Dokument
- Ordner R-Code:
  enthält alle mit R durchgeführten Analysen sowie alle erstellten Grafiken. Diese sind in die folgenden Unterordner aufgeteilt:
  - Anhang
  - Ausgabe Ergebnistabellen
  - Grafiken
  - Simulation
- Die Datei „Beschreibung der R-Codes.pdf“ enthält zudem Beschreibungen der einzelnen R-Skripte in den genannten Unterordnern.