

Ludwig-Maximilians-Universität München
Institut für Statistik
Wintersemester 2014/2015

Statistical Matching

Seminararbeit

Seminar "Statistische Herausforderungen im Umgang mit
fehlenden bzw. fehlerbehafteten Daten"

Unter der Leitung von: Prof. Dr. Augustin und Eva Endres

Verfasst von: Katrin Hummrich

Inhaltsverzeichnis

| | | |
|----------|--|-----------|
| 1 | Einleitung | 1 |
| 2 | Theorie | 1 |
| 2.1 | Notation und theoretische Vorüberlegungen | 2 |
| 2.2 | Weitere wichtige Aspekte von Verfahren des statistischen Matchings . . | 4 |
| 2.3 | Die bedingte Unabhängigkeitsannahme (CIA) | 8 |
| 2.3.1 | Der Makroansatz | 8 |
| 2.3.2 | Der Mikroansatz | 10 |
| 2.3.3 | Gemischte Methoden | 15 |
| 3 | Anwendungsbeispiel Lebensqualität | 16 |
| 3.1 | Ausgangslage | 16 |
| 3.2 | Statistisches Matching | 18 |
| 3.3 | Fazit | 24 |
| 4 | Zusammenfassung | 24 |
| 5 | Literaturverzeichnis | 25 |

Abbildungsverzeichnis

| | | |
|---|--|----|
| 1 | Allgemeine Datensituation beim statistischen Matching | 2 |
| 2 | Empirische Datensituation beim statistischen Matching | 12 |
| 3 | Randverteilungen vor und nach dem statistischen Matching | 22 |
| 4 | Gemeinsame Verteilung der Zielvariablen mit einer Matchingvariable . | 23 |

Tabellenverzeichnis

| | | |
|---|---|----|
| 1 | Übersicht der beiden Datensätze | 18 |
| 2 | Übersicht der übrigen Matchingvariablen | 19 |

1 Einleitung

Wer sich ein bisschen mit Statistik auskennt, der ist sicher schonmal dem Begriff "statistisches Matching" begegnet. Doch was verbirgt sich wirklich hinter diesem Begriff? Was ist das Ziel dieser Methoden? Wie der Name schon sagt, soll etwas gematcht, also etwas zusammengefügt, werden. Die namentlich sehr ähnlichen Matchingverfahren wollen dabei sogenannte statistische Zwillinge finden, um beispielsweise die Treatment-Evaluationsproblematik zu lösen. Doch darum geht es hier nicht. Beim *statistischen* Matching geht es darum zwei Datensätze zusammen zu fügen. Wobei hier auch kein Record Linkage gemeint ist. Denn dort werden Datensätze aus verschiedenen Datenquellen zusammengefügt, die die gleichen Objekte enthalten und die anhand von Schlüsselvariablen eindeutig identifiziert werden können. Beim statistischen Matching sollen auch Datensätze aus unterschiedlichen Quellen zusammengefügt werden, jedoch handelt es sich hier nicht um dieselben Objekte, so dass die Zusammenführung nur mit Hilfe von gleichen oder ähnlichen Ausprägungen in den Matchingvariablen stattfinden kann. Gründe sich für die Anwendung solcher Methoden zu entscheiden, können sein: Zeit, jedes Mal neue Daten zu erheben benötigt viel Zeit; Geld, Datenerhebungen kosten meist viel Geld; zu lange Fragebögen, um wirklich alle benötigten Informationen zu erfassen müsste meist so viel gefragt werden, dass der Fragebogen viel zu lang werden würde; zu hohe Belastung der Befragten, heutzutage werden so viele Daten erhoben, dass die Bevölkerung bereits recht oft mit irgendwelchen Umfragen konfrontiert wird, dadurch senkt sich die Bereitschaft zu antworten und es kommt vermehrt zu Nonresponse (Vgl. D'Orazio, Di Zio and Scanu (2006), S. 1). Insgesamt kann man sagen, wenn nicht jedes Mal neue Daten erhoben werden müssen, spart man sich Zeit und Geld und verschlechtert nicht die Datenqualität durch zu viel Nonresponse oder anderer durch Unlust entstandener Antwortmuster.

Wie nun brauchbare Informationen aus der Zusammenfügung der bereits vorhandenen Datensätze aus verschiedenen Quellen mit unterschiedlichen Objekten gezogen werden kann, soll im Folgenden gezeigt werden. Dabei wird zuerst die Ausgangssituation genauer beschrieben und dann die Theorie einiger wichtiger Ansätze erklärt. Schließlich soll anhand von einem Beispiel auch die Anwendung kurz dargestellt werden.

2 Theorie

Ziel des statistisches Matchings ist es, möglichst viel Information aus den bereits vorhanden Datenquellen zu schöpfen. Dafür sollen zwei oder mehr Datensätze zusammen gelegt werden. Diese enthalten zum einen Informationen über Variablen, die in beiden (allen) Datensätzen enthalten sind und zum anderen über nicht gemeinsam erhobene Variablen. Außerdem stammen die Beobachtungen von unterschiedlichen statistischen Einheiten.

Das Zusammenlegen besteht hier also nicht daraus einfach neue Spalten (Variablen) oder neue Zeilen (Beobachtungen) anzufügen. Sondern es geht darum gemeinsame Information über nicht gemeinsam erhobene Variablen zu gewinnen. Dabei gibt es zwei

| | X | Y | Z |
|---|------------|------------|------------|
| A | beobachtet | beobachtet | fehlend |
| B | beobachtet | fehlend | beobachtet |

beobachtet
 fehlend

Abb. 1: Allgemeine Datensituation beim statistischen Matching (Meinfelder (2013), S. 85)

grundlegende Ansätze, wie dies ermöglicht werden kann (Vgl. D’Orazio et al. (2006), S. 2-3).

Es gibt zum einen den *Mikroansatz*. Hier ist es das Ziel einen ”kompletten” Datensatz zu erhalten, in dem Sinne, dass dieser alle interessierenden Variablen aus den verschiedenen Datensätzen enthält. Der resultierende Datensatz wird oft auch als ”künstlich” bezeichnet, da er nicht direkt aus der Erhebung von Daten entstanden ist. Der Datensatz enthält zwar beobachtete Werte, jedoch wurden diese nicht alle gemeinsam und in genau dieser Kombination beobachtet, sondern im Nachhinein erst zusammengefügt. Deshalb werden sie als ”künstlich”, man könnte vielleicht auch ”unnatürlich” oder ”konstruiert” sagen, angesehen.

Und dann gibt es noch den *Makroansatz*. Hier werden die Datenquellen nicht in dem Sinne zusammengefügt, dass wirklich ein zusammenhängender Datensatz entsteht, sondern sie werden dafür genutzt die gemeinsame Verteilung oder andere Eigenschaften der interessierenden Variablen zu schätzen.

2.1 Notation und theoretische Vorüberlegungen

Zur besseren Übersicht und leichterem Verständnis soll in diesem Abschnitt die, zum größten Teil von D’Orazio et al. (2006) übernommene, Notation, die im Folgenden gilt, kurz eingeführt und erläutert werden. Dabei sei vorab noch erwähnt, dass wie üblich fett gedruckte Buchstaben Vektoren oder Matrizen sind, kleine Buchstaben für Realisationen und große Buchstaben für Zufallsvariablen stehen.

Der Einfachheit halber sollen lediglich die drei Zufallsvariablen (\mathbf{X} , \mathbf{Y} , \mathbf{Z}) mit der Dichte $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ betrachtet werden, wobei die Beobachtungen aus den folgenden Beobachtungsräumen stammen $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, $\mathbf{z} \in \mathcal{Z}$. Die zugehörigen Vektoren der Zufallsvari-

ablen mit den Dimension P , Q und R haben dann diese Form $\mathbf{X} = (X_1, \dots, X_p)^T$, $\mathbf{Y} = (Y_1, \dots, Y_q)^T$ und $\mathbf{Z} = (Z_1, \dots, Z_r)^T$. Außerdem wird für die Dichte angenommen, dass sie aus der Verteilungsfamilie $\mathcal{F} = \{f\}$ kommt. Um es einfacher zu halten, sollen auch nur zwei Datensätze A und B mit n_a und n_b Beobachtungen betrachtet werden. Diese Beobachtungen unterliegen der Annahme, dass sie unabhängig und identisch (i.i.d.) verteilt sind und aus einer Verteilung mit der Dichte $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ stammen. Sei \mathbf{X} die Variable, die in beiden Datensätzen erhoben wurde und fehle die Variable \mathbf{Z} in A und \mathbf{Y} in B , so enthält A die beobachteten Werte $(\mathbf{x}_a, \mathbf{y}_a) = (x_{a1}, \dots, x_{ap}, y_{a1}, \dots, y_{aq})$ und B die Werte $(\mathbf{x}_b, \mathbf{z}_b) = (x_{b1}, \dots, x_{bp}, z_{b1}, \dots, z_{br})$. Der Datensatz der Vereinigung $A \cup B$ mit $n_a + n_b$ i.i.d. Beobachtungen aus $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ hat folgende charakteristische Eigenschaften (Vgl. D’Orazio et al. (2006), S. 4):

- das Auftreten von fehlenden Daten und daraus resultierenden Mechanismen
- der Mangel an gemeinsamen Informationen über \mathbf{X} , \mathbf{Y} und \mathbf{Z}

Um die erste Eigenschaft in den Griff zu bekommen, finden sich viele mögliche Lösungswege, wie beispielsweise Imputationsmethoden. Das statistische Matching befasst sich vor allem mit der zweiten Eigenschaft, dem Mangel an gemeinsamen Informationen über die interessierenden Variablen, die aber eben aus verschiedenen Datenquellen stammen. Um diese Besonderheit des nicht-vollständig beobachteten Datensatzes, der man beim statistischen Matching begegnet, darstellen zu können, werden noch weitere Notationen von D’Orazio et al. (2006) eingeführt.

Eine weitere Zufallsvariable \mathbf{R} soll Auskunft darüber geben welche Beobachtung vorhanden ist und welche fehlt. Dafür sei $\mathbf{R} = (\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z)$ bestehend aus den Zufallsvektoren $\mathbf{R}_x = (R_{X_1}, \dots, R_{X_p})^T$, $\mathbf{R}_y = (R_{Y_1}, \dots, R_{Y_q})^T$ und $\mathbf{R}_z = (R_{Z_1}, \dots, R_{Z_r})^T$ mit den Dimensionen P , Q und R . Hat zum Beispiel R_{X_j} den Wert 1, beutet das, dass X_j beobachtet wurde. Der Wert 0 steht demnach für eine fehlende Beobachtung X_j , mit $j = 1, \dots, P$. Für den Missing Mechanismus ist es nun interessant zu wissen wie die bedingte Verteilung von \mathbf{R} gegeben die interessierenden Variablen aussieht. Diese sei hier bezeichnet mit $h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z | \mathbf{x}, \mathbf{y}, \mathbf{z})$. Entscheidend ist nun welcher Missing Mechanismus in der Situation des statistischen Matchings vorliegt. ”Während das Datenausfallmuster designbasiert und somit nicht zufällig ist, ist der zugrundeliegende Datenausfallmechanismus im Idealfall absolut zufällig – nämlich dann, wenn die beteiligten Studien A und B Zufallsstichproben aus derselben Population sind. Der Datenausfallmechanismus wird in einem solchen Fall als Missing Completely at Random (MCAR) bezeichnet.” (Mainfelder (2013), S. 84). Man kann also sagen, dass das Fehlen der Werte von \mathbf{Y} bzw. von \mathbf{Z} weder von den Werten selbst noch von denen der anderen Variablen abhängt. Das heißt es gilt.

$$h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z | \mathbf{x}, \mathbf{y}, \mathbf{z}) = h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z). \quad (1)$$

D’Orazio et al. (2006) zeigt das analytisch indem er sich die bedingte Dichte von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ gegeben \mathbf{R} ansieht. Diese sei hier zur besseren Unterscheidung mit ϕ bezeichnet. Auf Grund der Symmetrie von Unabhängigkeit zwischen Zufallsvariablen, gilt auch umgekehrt:

$$\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z) = \phi(\mathbf{x}, \mathbf{y}, \mathbf{z}).$$

Wegen der bereits beschriebenen Struktur von A und B (siehe Abb. 1), kann es offensichtlich nur zwei mögliche Beobachtungsmuster für \mathbf{R} geben. $\mathbf{R} = (\mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R)$ für die Einheiten aus A und $\mathbf{R} = (\mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R)$ für die Einheiten aus B , wobei $\mathbf{1}_j$ und $\mathbf{0}_j$ zwei j -dimensionale Vektoren mit Einsen und Nullen sind. Auf Grund der i.i.d. Annahme für die $n_A + n_B$ Werte aus $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ gilt für alle $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$ und $\mathbf{z} \in \mathcal{Z}$:

$$\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R) = \phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R) = f(\mathbf{x}, \mathbf{y}, \mathbf{z}), \quad (2)$$

wobei $\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R)$ die Verteilung der Werte in A und $\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R)$ die Verteilung der Werte in B darstellt (Vgl. D’Orazio et al. (2013), S. 6-7). Das bedeutet, dass der Missing Mechanismus sowohl von den beobachteten Werten als auch von den fehlenden Werten unabhängig ist, was genau der Definition von MCAR entspricht. Nach D’Orazio et al. (2006) erlaubt dieser Sachverhalt es nicht nur Inferenz über die gemeinsame Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ zu betreiben ohne \mathbf{R} zu beachten, sondern auch diese auf die beobachtete Stichprobenverteilung zu stützen. Umgesetzt werden kann diese Inferenz mittels Marginalisierung der gemeinsamen Verteilung $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ unter Berücksichtigung der nicht erfassten Variablen. Folglich lässt sich die beobachtete Stichprobenverteilung der $n_A + n_B$ Beobachtungseinheiten berechnen mittels

$$\prod_{a=1}^{n_A} f_{\mathbf{X}\mathbf{Y}}(\mathbf{x}_a, \mathbf{y}_a) \prod_{b=1}^{n_B} f_{\mathbf{X}\mathbf{Z}}(\mathbf{x}_b, \mathbf{z}_b) \quad (3)$$

(Vgl. D’Orazio et al. (2006), S. 7).

2.2 Weitere wichtige Aspekte von Verfahren des statistischen Matchings

Nach theoretischen Überlegungen darüber was genau das Problem des statistischen Matchings ist und welche Annahmen in die Modelle gesteckt werden, sollte auch betrachtet werden wie überprüft bzw. beurteilt werden kann, ob das gewählte Matchingverfahren das gewünschte Ergebnis gebracht hat um die geplanten Analysen durchführen zu können. D’Orazio et al. (2006) stellen an dieser Stelle vier wichtige Fragen, die im Folgenden kurz betrachtet werden sollen:

- (a) Welche Annahmen für das gemeinsame Modell $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ können vernünftig in Betracht gezogen werden?

- (b) Welcher Schätzer unter allen, die den Modellannahmen aus (a) entsprechen, ist für die Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ zu bevorzugen?
- (c) Welche Methoden können verwendet werden um passende Werte für die fehlenden Variablen zu erzeugen, die zum gewählten Modell aus (a) und dem gewählten Schätzer aus (b) passen?
- (d) Welche Inferenzverfahren können auf den durch statistisches Matching erhaltenen Datensatz angewendet werden?

Die Fragen hängen so gesehen zusammen, dass die Bearbeitung einer Frage nur Sinn macht, wenn die vorhergehenden Fragen bereits beantwortet sind. Für den Makroansatz sind die Fragen (a) und (b) interessant, für Mikroansätze kommt Frage (c) hinzu und möchte man die Ergebnisse des Mikroansatzes analysieren muss man sich noch zusätzlich Frage (d) stellen (Vgl. D’Orazio et al. (2006), S. 8).

Zu Frage (a) - **die Modellannahmen**

Die Situation beim statistischen Matching ist etwas schwierig, da keine Beobachtung zur Verfügung steht, bei der alle interessierenden Variablen zusammen erfasst wurden. Anhand der Menge $A \cup B$, die vorliegt, ist es demnach nicht durchführbar alle möglichen Modelle für $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ zu identifizieren. Das liegt daran, dass die Menge $A \cup B$ keine gemeinsamen Informationen von \mathbf{Y} und \mathbf{Z} enthält um Parameter von (\mathbf{Y}, \mathbf{Z}) zu schätzen und es auch nicht machbar ist anhand von $A \cup B$ zu testen welches Modell passend für $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ ist. D’Orazio et al. (2006) schlagen drei verschiedene Möglichkeiten diese Problematik zu umgehen vor:

- Durch weitere Informationen (z.B. aus vorhergehenden Erfahrungen oder ad hoc Studien) die Verwendung eines identifizierbaren Modells von $A \cup B$ rechtfertigen.
- Verwende weitere Informationen (z.B. aus vorhergehenden Erfahrungen oder ad hoc Studien) zusammen mit $A \cup B$ um andere Modelle zu identifizieren.
- Es werden keine Modellannahmen über $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ gemacht und dieses Problem so behandelt, wie bei Unwissenheit über einige Modelleigenschaften.

Bei den ersten beiden Optionen ist es sehr wichtig, dass die dafür notwendigen Annahmen durch die erwähnten Informationsquellen gut begründet werden. Denn die Ergebnisse werden immer das Modell widerspiegeln, dem sie zu Grunde liegen und wenn die dazugehörigen Annahmen falsch sind und dem tatsächlich zu Grunde liegenden Modell nicht entsprechen, können somit auch keine richtigen Ergebnisse damit erzielt werden. Das gilt sowohl für die Makro- als auch für die Mikroansätze. Können die notwendigen Annahmen gut begründet werden, ist es möglich darauf basierend eine Punktschätzung für die interessierenden Parameter zu machen. Trotzdem kann es nicht schaden eine Abschätzung der Unsicherheit zu machen, um ein Gefühl für die Zuverlässigkeit der,

anhand der Analyse mit dem geschätzten Modell, gezogenen Schlüsse zu bekommen. Die dritte Option stellt eine konservativere Variante dar. Durch eine Intervallschätzung der Parameter an Stelle einer Punktschätzung legt man sich nicht auf einen Schätzwert fest, sondern gibt einen Bereich an, indem dieser liegt und beschreibt somit die Ungewissheit, die für den geschätzten Parameterwert, vorliegt. (Vgl. D’Orazio et al. (2006), S. 8-9)

Zu Frage (b) - **Genauigkeit des Schätzers**

Es wird nun davon ausgegangen, dass bereits ein Modell für $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ gefunden wurde. Um die Genauigkeit des Schätzers für die Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ beim Makroansatz zu beurteilen, können Messungen, wie der MSE (mean square error), herangezogen werden. Der MSE bezeichnet die erwartete mittlere quadratische Abweichung und wird bekannterweise aus der Summe der Varianz und des quadrierten Bias des Schätzers gebildet (Vgl. Fahrmeir et al. (2010), S. 371). Befindet man sich im parametrischen Fall, wird also eine Verteilungsannahme gemacht, kann der MSE häufig durch Verwendung des ML-Schätzers (Maximum Likelihood-Schätzers) minimiert werden, was an der Konsistenzeigenschaft des ML-Schätzers liegt (Vgl. D’Orazio et al. (2006), S. 9). Wenn der MSE für Stichprobengrößen n , die gegen Unendlich gehen, gleich Null ist, ist der Schätzer MSE-Konsistent (Vgl. Fahrmeir et al. (2010), S. 372). Auch im nichtparametrischen Fall spielt die Konsistenzeigenschaft des Schätzers eine große Rolle. Allgemein ist diese Eigenschaft für das statistische Matching sehr wichtig, da sie sicherstellt, dass für große Stichproben die Schätzer nah am wahren Wert dran sind. Für den Mikroansatz ist diese Tatsache ebenfalls relevant, wie im Folgenden noch klar wird.

Zu Frage (c) - **Repräsentativität des gematchten Datensatzes**

Hier können nach D’Orazio et al. (2006) vier Kategorien unterschieden werden, die vier Zielen entsprechen, was der erzeugte Datensatz leisten können soll (mit dem komplexesten und am schwierigsten erreichbaren Ziel beginnend):

- Die künstlichen Werte sollten mit den wahren, aber unbekanntenen Werten übereinstimmen.
- Die gemeinsame Verteilung aller Variablen wird durch den, anhand des statistischen Matchings erstellten, Datensatz wiedergegeben.
- Die Korrelationsstruktur der Variablen bleibt erhalten.
- Die marginalen und gemeinsamen Verteilungen der Variablen aus den Ursprungsdatensätzen bleiben in dem, durch statistisches Matching erstellten, Datensatz erhalten.

Das erste Ziel ist kaum erfüllbar und verlangt eigentlich sogar zu viel (Vgl. D’Orazio et al. (2006), S. 10). Denn es ist nicht notwendig, dass für jeden fehlenden Wert

der exakte wahre Wert gefunden wird. Schließlich interessieren nicht die Werte der einzelnen Objekte, sondern die gemeinsame Verteilung der beobachteten Variablen. Die Ziele drei und vier verlangen dagegen zu wenig. Hier wird nicht sichergestellt, dass bei Erfüllen dieser Eigenschaften auch angemessene Inferenz für $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ gemacht werden kann (Vgl. D’Orazio et al. (2006), S. 10). Denn die Korrelationsstruktur gibt zwar Auskunft über die Stärke des Zusammenhangs der Variablen wieder, jedoch nicht über dessen Verteilung. Außerdem bedeutet nicht, dass durch den Erhalt der Korrelationsstruktur oder auch der marginalen bzw. gemeinsamen Verteilungen der Ursprungsdatensätze automatisch die wahre gemeinsame Verteilung aller Variablen wiedergegeben wird. Das zweite Ziel stellt eine sinnvolle Zielsetzung für das statistische Matching dar. Wenn dieses erfüllt wird, kann der künstliche Datensatz als eine aus der gemeinsamen Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ generierte Stichprobe angesehen werden (Vgl. D’Orazio et al. (2006), S. 10). Das heißt der gematchte Datensatz ist repräsentativ für die Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ und stellt ein geeignetes Mittel dar um auf Eigenschaften der Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ zu schließen.

Eine hier noch wichtig zu erwähnende Größe ist das so genannte *matching noise*. D’Orazio et al. (2006) bezeichnen mit diesem Begriff den Unterschied zwischen dem wahren datengenerierenden Modell und dem Modell, das dem künstlichen Datensatz zu Grunde liegt. Die Frage, die an dieser Stelle aufkommt, ist, ob die Daten aus dem gematchten Datensatz von dem matching noise beeinträchtigt sind oder nicht. Allerdings ist diese Frage nur schwer zu beantworten. Wie bereits erwähnt wurde, ist die Konsistenzeigenschaft der verwendeten Schätzer sehr bedeutend. So kommt sie auch an dieser Stelle zum Tragen, da es möglich ist Mikroansätze mit reduziertem matching noise zu finden, wenn bei großen Datensätzen konsistente Schätzer für $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ verwendet werden (Vgl. D’Orazio et al. (2006), S. 11). Es gilt allerdings zu beachten, dass eine gute Schätzung von $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ zwar notwendig, aber nicht ausreichend ist um sicherzustellen, dass das matching noise so gering wie möglich ist (D’Orazio et al. (2006), S. 11).

Zu Frage (d) - **Genauigkeit der Schätzer basierend auf dem gematchten Datensatz**

Dieser Punkt stellt nach D’Orazio et al. (2006) einen kritischen Aspekt des Mikroansatzes dar. Ist das matching noise ausgeprägt, sind auch die Schätzungen basierend auf dem gematchten Datensatz nicht brauchbar. Denn in diesem Fall ist bereits der gematchte Datensatz nicht repräsentativ und folglich können auch darauf aufbauende Schätzungen nur zu falschen Ergebnissen und Schlussfolgerungen führen. Es ist somit elementar das matching noise so gut es geht zu reduzieren.

2.3 Die bedingte Unabhängigkeitsannahme (CIA)

Eine sehr wichtige, aber durchaus restriktive Annahme, die es ermöglicht $A \cup B$ zu identifizieren und direkt zu schätzen und vielen Verfahren des statistischen Matchings zu Grunde liegt, ist die Annahme der bedingten Unabhängigkeit von \mathbf{Y} und \mathbf{Z} gegeben \mathbf{X} . Diese Annahme wird auch mit der Abkürzung CIA, vom englischen conditional independence assumption, bezeichnet.

Ist diese Annahme zutreffend, lässt sich die Dichte der Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ wie folgt definieren:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})f_{\mathbf{X}}(\mathbf{x}), \quad (4)$$

für alle $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$ und $\mathbf{z} \in \mathcal{Z}$, wobei $f_{\mathbf{Y}|\mathbf{X}}$ die bedingte Dichte von \mathbf{Y} gegeben \mathbf{X} , $f_{\mathbf{Z}|\mathbf{X}}$ die bedingte Dichte von \mathbf{Z} gegeben \mathbf{X} und $f_{\mathbf{X}}$ die marginale Dichte von \mathbf{X} bezeichnet (Vgl. D’Orazio et al. (2006), S. 13). Die Informationen, die aus den einzelnen Datensätzen A und B gezogen werden können, reichen somit für die Schätzung der gemeinsamen Verteilung aus. Schließlich werden in (4) lediglich Informationen über die marginale Verteilung von \mathbf{X} und den paarweisen Beziehungen von \mathbf{X} und \mathbf{Y} sowie von \mathbf{X} und \mathbf{Z} gebraucht.

Es ist wichtig zu beachten, dass diese starke Annahme nur anhand von $A \cup B$ nicht getestet werden kann und somit auch falsch sein kann, was Fehlspezifikationen zur Folge hat (D’Orazio et al. (2006), S. 13). Trotzdem wird im Folgenden davon ausgegangen, dass CIA zutrifft.

2.3.1 Der Makroansatz

Beim Makroansatz geht es, wie bereits erwähnt, darum die gemeinsame Verteilung der interessierenden Variablen zu schätzen. Dabei kann zwischen parametrischen und nicht-parametrischen Methoden unterschieden werden. Beide werden im Folgenden kurz betrachtet.

Parametrische Makroansätze

Im parametrischen Fall handelt es sich folglich bei \mathcal{F} um eine parametrische Verteilungsfamilie. Somit wird jede Dichte $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) \in \mathcal{F}$ durch einen endlich dimensionalen Parametervektor $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^T$ mit T als ganze Zahl definiert. Nach D’Orazio et al. (2006) gilt analog zu Gleichung (4) unter der Annahme von CIA, dass \mathcal{F} als das Produkt von $\mathcal{F}_{\mathbf{X}}$, $\mathcal{F}_{\mathbf{Y}|\mathbf{X}}$ und $\mathcal{F}_{\mathbf{Z}|\mathbf{X}}$ dargestellt werden kann. Dabei gilt für die marginale Dichte von \mathbf{X} $f_{\mathbf{X}}(\cdot; \boldsymbol{\theta}_{\mathbf{X}}) \in \mathcal{F}_{\mathbf{X}}$, für die bedingte Dichte von \mathbf{Y} gegeben \mathbf{X} $f_{\mathbf{Y}|\mathbf{X}}(\cdot; \boldsymbol{\theta}_{\mathbf{Y}|\mathbf{X}}) \in \mathcal{F}_{\mathbf{Y}|\mathbf{X}}$ und für die bedingte Dichte von \mathbf{Z} gegeben \mathbf{X} analog $f_{\mathbf{Z}|\mathbf{X}}(\cdot; \boldsymbol{\theta}_{\mathbf{Z}|\mathbf{X}}) \in \mathcal{F}_{\mathbf{Z}|\mathbf{X}}$. Anhand der Faktorisierung aus Gleichung (4) lässt sich die Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ durch die

Parametervektoren $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_{Y|X}$ und $\boldsymbol{\theta}_{Z|X}$ darstellen als:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}_X) f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{Y|X}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{Z|X}), \quad (5)$$

mit $\boldsymbol{\theta}_X \in \Theta_X$, $\boldsymbol{\theta}_{Y|X} \in \Theta_{Y|X}$ und $\boldsymbol{\theta}_{Z|X} \in \Theta_{Z|X}$ (Vgl. D’Orazio et al. (2006), S. 14). Ziel des parametrischen Makroansatzes ist somit die Schätzung der Parameter $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}_{Z|X})$. Die beobachtete Likelihoodfunktion der gemeinsamen Menge $A \cup B$ beschreiben D’Orazio et al. (2006) dann folgendermaßen:

$$\begin{aligned} L(\boldsymbol{\theta}|A \cup B) &= \prod_{a=1}^{n_A} f_{\mathbf{X}\mathbf{Y}}(\mathbf{x}_a, \mathbf{y}_a; \boldsymbol{\theta}) \prod_{b=1}^{n_B} f_{\mathbf{X}\mathbf{Z}}(\mathbf{x}_b, \mathbf{z}_b; \boldsymbol{\theta}) \\ &= \prod_{a=1}^{n_A} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_a|\mathbf{x}_a; \boldsymbol{\theta}_{Y|X}) \prod_{b=1}^{n_B} f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}_b|\mathbf{x}_b; \boldsymbol{\theta}_{Z|X}) \cdot \prod_{a=1}^{n_A} f_{\mathbf{X}}(\mathbf{x}_a; \boldsymbol{\theta}_X) \prod_{b=1}^{n_B} f_{\mathbf{X}}(\mathbf{x}_b; \boldsymbol{\theta}_X). \end{aligned} \quad (6)$$

Hier kann folglich eine ML-Schätzung gemacht werden, obwohl die Daten aus $A \cup B$ durch fehlende Werte geprägt sind, indem jeweils der passende, komplette (Einzel-) Datensatz für die Schätzung verwendet wird. Das heißt der ML-Schätzer für $\boldsymbol{\theta}_X$ wird anhand von $A \cup B$ berechnet und die ML-Schätzer für $\boldsymbol{\theta}_{Y|X}$ und $\boldsymbol{\theta}_{Z|X}$ anhand der Einzeldatensätze A und B .

Nicht-parametrische Makroansätze

Es kann auch vorkommen, dass nicht ausreichend Informationen vorhanden sind um \mathcal{F} einer parametrischen Verteilungsfamilie zuzuordnen. In diesen Fällen ist es ratsam eine nicht-parametrische Methode vorzuziehen, da diese nicht durch möglicherweise falsche Annahmen bezüglich der parametrischen Form von \mathcal{F} beeinflusst wird.

Einige nicht-parametrische Makroansätze schätzten die Dichte $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ direkt und nutzten dafür die Faktorisierung aus Gleichung (4):

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}).$$

Für die nicht-parametrische Schätzung der einzelnen Dichten kann beispielsweise eine Kerndichteschätzung gemacht werden oder die Nächste-Nachbarn-Klassifikation (kNN von k nearest neighbour) verwendet werden (Vgl. D’Orazio et al. (2006), S. 32-33).

Ein weiterer möglicher nicht-parametrischer Makroansatz, den D’Orazio et al. (2006) vorstellen, schätzt die empirische kumulative Verteilungsfunktion mit Hilfe der gemeinsamen kumulativen Verteilungsfunktion von (\mathbf{Y}, \mathbf{Z}) gegeben \mathbf{X}

$$\mathbf{F}_{\mathbf{Y}\mathbf{Z}|\mathbf{X}}(\mathbf{y}, \mathbf{z}|\mathbf{x}) = \int_{\mathbf{t} \leq \mathbf{y}} \int_{\mathbf{v} \leq \mathbf{z}} f_{\mathbf{Y}\mathbf{Z}|\mathbf{X}}(\mathbf{t}, \mathbf{v}|\mathbf{x}) dt dv, \quad (7)$$

wobei \mathbf{X} hier kategorial ist und $\mathbf{t} \leq \mathbf{y}$ und $\mathbf{v} \leq \mathbf{z}$ komponentenweise Ungleichungen darstellen. Unter CIA gilt folgende, praktische Zergliederung

$$\mathbf{F}_{\mathbf{YZ}|\mathbf{X}}(\mathbf{y}, \mathbf{z}|\mathbf{x}) = \mathbf{F}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\mathbf{F}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}), \quad (8)$$

wobei die einzelnen Faktoren sich durch

$$\hat{\mathbf{F}}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{\sum_{a=1}^{n_A} \mathbf{I}(\mathbf{y}_a \leq \mathbf{y})\mathbf{I}(\mathbf{x}_a = \mathbf{x})}{\sum_{a=1}^{n_A} \mathbf{I}(\mathbf{x}_a = \mathbf{x})}, \quad (9)$$

$$\hat{\mathbf{F}}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}) = \frac{\sum_{b=1}^{n_B} \mathbf{I}(\mathbf{z}_b \leq \mathbf{z})\mathbf{I}(\mathbf{x}_b = \mathbf{x})}{\sum_{b=1}^{n_B} \mathbf{I}(\mathbf{x}_b = \mathbf{x})} \quad (10)$$

schätzen lassen (Vgl. D’Orazio et al. (2006), S. 31-32). Hier kann es vorkommen, dass manche Kategorien von \mathbf{X} in A und/oder B nicht beobachtet wurden und die zugehörige empirische kumulative Verteilungsfunktion für die Kategorie von \mathbf{X} dann nicht geschätzt werden kann.

2.3.2 Der Mikroansatz

Ziel der Mikroansätze ist es einen kompletten Datensatz für $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ zu erstellen, indem die fehlenden Werte in A und B "aufgefüllt" werden.

Parametrische Mikroansätze

Laut D’Orazio et al. (2006) wird bei den parametrischen Mikroansätzen ein explizites parametrisches Modell verwendet. Dieses wird ein Mal geschätzt und dann wird ein kompletter, künstlicher Datensatz erstellt, indem die fehlenden Werte in $A \cup B$ durch passende Werte aus der Verteilung der entsprechenden Variablen gegeben die beobachteten Variablen ersetzt werden. Daher kann dieser Ansatz auch als single Imputationsmethode bezeichnet werden. Die parametrischen Mikroansätze lassen sich in zwei große Kategorien einteilen, das sogenannte *conditional mean matching* und *draws based on conditional predictive distributions*.

Das conditional mean matching

Wenn die Variablen \mathbf{Y} und \mathbf{Z} stetig sind, kann jeder fehlende Wert durch den Erwartungswert der fehlenden Variable gegeben die beobachteten Variablen ersetzt werden, das heißt es gilt:

$$\tilde{\mathbf{z}}_a = E(\mathbf{Z}|\mathbf{X} = \mathbf{x}_a) = \int_{\mathbf{Z}} \mathbf{z} f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}_a; \boldsymbol{\theta}_{\mathbf{Z}|\mathbf{X}}) d\mathbf{z}, \quad a = 1, \dots, n_A \quad (11)$$

und

$$\tilde{y}_b = E(\mathbf{Y}|\mathbf{X} = \mathbf{x}_b) = \int_{\mathcal{Y}} \mathbf{y} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}_b; \boldsymbol{\theta}_{\mathbf{Y}|\mathbf{X}}) d\mathbf{y}, \quad b = 1, \dots, n_B \quad (12)$$

wobei die unbekannt Parameter $\boldsymbol{\theta}_{\mathbf{Z}|\mathbf{X}}$ und $\boldsymbol{\theta}_{\mathbf{Y}|\mathbf{X}}$ durch die zugehörigen ML-Schätzer ersetzt werden können (Vgl. D’Orazio et al. (2006), S. 26). Gemäß der Definition von der einfachen Regression, bedeutet das, dass die imputierten Werte den Werten aus der geschätzten Regressionsfunktion von \mathbf{Z} bzw. \mathbf{Y} auf \mathbf{X} entsprechen. D’Orazio et al. (2006) weisen darauf hin, dass diese Methode im Hinblick darauf, dass der ML-Schätzer bezüglich der quadratischen Verlustfunktion der beste Punktschätzer ist, sehr ansprechend wirkt. Sie erwähnen jedoch auch zwei Nachteile dieser Methode. Der erste Nachteil ist, dass die prognostizierten Werte keine tatsächlich beobachteten Werte sind. Der zweite Nachteil ist, dass die künstliche Verteilung der prognostizierten Werte von \mathbf{Y} bzw. \mathbf{Z} ausgerichtet ist auf den Erwartungswert von \mathbf{Y} bzw. \mathbf{Z} gegeben \mathbf{X} . Das heißt alle Punkte liegen auf der Regressionsgeraden und die Varianz wird somit unterschätzt. Wie später noch deutlich wird, können diese Werte trotzdem nützlich sein.

Draws based on conditinal predictive distributions

Eine Möglichkeit den Nachteil der fehlenden Variabilität der imputierten Werte des conditional mean matchings zu begegnen stellt die folgende Methode der Draws based on conditinal predictive distributions dar. Gemäß dem Namen werden hier also zufällige Ziehungen aus bedingten Vorhersageverteilungen gemacht. Unter der Annahme, dass der Mechanismus der fehlenden Werte einem MAR (missing at random) Mechanismus entspricht, schildern D’Orazio et al. (2006), dass die datengenerierende multivariate Verteilung besser bestimmt werden kann, wenn das Imputieren der fehlenden Werte durch eine zufällige Ziehung aus einer Vorhersageverteilung erfolgt. Hier wird nicht, wie vorher, MCAR angenommen, da in empirischen Datensituationen meist nicht davon ausgegangen werden kann, dass die Datensätze aus der gleichen Grundgesamtheit stammen, weil sie beispielsweise wegen Teilnahmeverweigerung nicht die gleiche Struktur haben (siehe Abb. 2). Durch CIA wird aber eben automatisch MAR angenommen (Vgl. Meinfelder (2013), S. 86). Das bedeutet in diesem Fall, dass für alle $a = 1, \dots, n_A$ ein zufälliger Wert aus $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}_a; \hat{\boldsymbol{\theta}}_{\mathbf{Z}|\mathbf{X}}^{(ML)})$ gezogen wird und für alle $b = 1, \dots, n_B$ aus $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}_b; \hat{\boldsymbol{\theta}}_{\mathbf{Y}|\mathbf{X}}^{(ML)})$ (Vgl. D’Orazio et al. (2006), S. 29). Dabei steht $\hat{\boldsymbol{\theta}}_{\mathbf{Z}|\mathbf{X}}^{(ML)}$ bzw. $\hat{\boldsymbol{\theta}}_{\mathbf{Y}|\mathbf{X}}^{(ML)}$ für den entsprechenden ML-Schätzer.

Bei Mikroansätzen des statistischen Matchings ist es wichtig, dass der resultierende Datensatz repräsentativ für die Verteilung mit der Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$ ist, um darauf basierende Inferenz betreiben zu können. In den eben beschriebenen Methoden wurde immer der ML-Schätzer verwendet um den unbekannt Parametervektor $\boldsymbol{\theta}$ zu schätzen. Dies bringt den Vorteil, dass auf Grund der Konsistenzeigenschaft des ML-Schätzers, vor allem bei großen Datensätzen, $\hat{\boldsymbol{\theta}}$ approximativ dem wahren unbekannt

| | X | Y | Z |
|---|------------|------------|------------|
| A | beobachtet | beobachtet | fehlend |
| B | beobachtet | fehlend | beobachtet |
| | fehlend | fehlend | fehlend |

beobachtet
 fehlend

Abb. 2: *Datensituation beim statistischen Matching mit missing-by-design Ausfallmuster unter Einbeziehung von Unit Nonresponse (Meinfielder (2013), S. 86)*

Parametervektor θ entspricht. Daraus folgern D’Orazio et al. (2006), dass Datensätze, die aus draws based on conditional predictive distributions resultieren, als approximativ repräsentativ für $f(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta)$ angesehen werden können. Für Datensätze, die durch conditional mean matching erhalten werden, folgern sie dies jedoch nicht, da die Varianz von \mathbf{Z} wegen mangelnder Variabilität der imputierten Werte unterschätzt wird.

Nicht-parametrische Mikroansätze

Die zwei vorhergehenden Methoden können auch als nicht-parametrische Varianten durchgeführt werden, indem eine nicht-parametrische Regressionsfunktion verwendet wird bzw. die Verteilung von $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ nicht-parametrisch geschätzt wird. Wofür es jeweils einige Möglichkeiten gibt. Hier werden jedoch andere nicht-parametrische Mikroansätze betrachtet und zwar sogenannte *hot deck Imputationsmethoden*.

Charakteristisch für diese Methoden ist, dass die fehlenden Werte durch beobachtete Werte ersetzt werden. Das heißt ein Nachteil des conditional mean matchings wird an dieser Stelle behoben. Was auf den ersten Blick auch für die Methoden spricht, ist, dass weder eine Verteilungsfamilie spezifiziert werden, noch eine Verteilungsfunktion geschätzt werden muss. Wie jedoch später noch deutlich wird, wird bei hot deck Methoden durchaus implizit eine Schätzung einer Verteilung oder einer bedingten Verteilungsfunktion vorausgesetzt (Vgl. D’Orazio et al. (2006), S. 35).

In der Praxis wird meist der statistische Matchingprozess anhand von hot deck Methoden als Imputation von einem ’Spender’-Datensatz (donor file) zu einem ’Empfänger’-Datensatz (recipient file) betrachtet. Das heißt zum Beispiel, dass die fehlenden Werte \mathbf{Z} in A (Empfänger) mittels der beobachteten Werte aus B (Spender) imputiert werden. Dazu wird der Zusammenhang zwischen der Variable \mathbf{Z} , die nur in B beobachtet wurde und der Variable \mathbf{X} , die in beiden Datensätzen enthalten ist, genutzt. Die Wahl

welchem Datensatz welche Rolle zufällt, hängt, wie D’Orazio et al. (2006) erläutern, von mehreren Faktoren ab. Wichtig ist dabei, dass man davon ausgehen kann, dass die beiden Datensätze aus der gleichen Verteilung mit der Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ stammen. Sind die Stichprobengrößen erheblich unterschiedlich, wird der kleinere Datensatz als Empfänger-Datensatz genommen, da sonst nicht ausreichend Spender vorhanden wären. Im Folgenden wird immer A als Empfänger-Datensatz und B als Spender-Datensatz betrachtet, das heißt \mathbf{Z} soll mit Hilfe der beobachteten Werte aus B in A imputiert werden.

Nach Klärung dieser Begrifflichkeiten, sollen nun diese drei verschiedenen hot deck Verfahren, *random hot deck*, *rank hot deck* und *distance hot deck*, die beim statistischen Matching verwendet werden, nacheinander betrachtet werden.

Die random hot deck Methode

Die random hot deck Methode wählt für jeden fehlenden Wert aus dem Empfängerdatensatz zufällig einen Eintrag aus dem Spenderdatensatz aus. Meistens werden die statistischen Einheiten aus A und B anhand einer oder mehrerer kategorialer Variablen \mathbf{X} , die in beiden Datensätzen erhoben wurden, in Klassen eingeteilt, so dass für die zufällige Auswahl eines Spenders nur die Einträge in Frage kommen, die in der gleichen Klasse wie der Empfänger sind. Dabei entspricht die random hot deck Methode mit getrennten Klassen einer Schätzung der bedingten Verteilung von \mathbf{Z} gegeben \mathbf{X} in B und einer zufälligen Ziehung daraus. Und random hot deck ohne Unterteilung in Klassen entspricht einer Schätzung der marginalen Verteilung von \mathbf{Z} in B , was implizit annimmt, dass \mathbf{Z} und \mathbf{X} unabhängig sind (Vgl. D’Orazio et al. (2006), S. 38-39).

Die rank hot deck Methode

Die rank hot deck Methode kann angewendet werden, wenn eine ordinale Matchingvariable X vorhanden ist. In diesem Fall wird ausgenutzt, dass die Werte von X in eine Reihenfolge gebracht werden können. Dazu werden die Einträge in beiden Datensätzen anhand der Werte von X sortiert. Ist im einfachsten Fall die Stichprobengröße von dem Spenderdatensatz B ein Vielfaches der Stichprobengröße von dem Empfängerdatensatz A , so erfolgt das statistische Matching durch Verknüpfung von Einträgen mit dem gleichen Rang (Vgl. D’Orazio et al. (2006), S. 39). Sind die Stichprobengrößen keine Vielfachen voneinander, werden die empirischen kumulativen Verteilungsfunktionen von X im Empfängerdatensatz

$$\hat{F}_X^A(x) = \frac{1}{n_A} \sum_{a=1}^{n_A} I(x_A \leq x), \quad x \in \mathcal{X}$$

und im Spenderdatensatz

$$\hat{F}_X^B(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(x_B \leq x), \quad x \in \mathcal{X}$$

herangezogen und jedem $a = 1, \dots, n_A$ wird dann der Eintrag b^* aus B zugeordnet, für den gilt

$$|\hat{F}_X^A(x_a) - \hat{F}_X^B(x_{b^*})| = \min_{1 \leq b \leq n_B} |\hat{F}_X^A(x_a) - \hat{F}_X^B(x_b)|$$

(Vgl. D’Orazio et al. (2006), S. 39).

Die distance hot deck Methode

Bei der distance hot deck Methode wird jedem Eintrag aus dem Empfängerdatensatz der Eintrag aus dem Spenderdatensatz zugeordnet, der den kleinsten Abstand (häufig die Mahalanobis Distanz, andere Distanzmaße sind aber auch möglich) bezüglich der Matchingvariablen \mathbf{X} hat. Etwas anschaulicher bedeutet das im einfachsten Fall, mit lediglich einer einzigen stetigen Matchingvariablen X , dass der Spender b^* für den a -ten Eintrag aus A so gewählt wird, dass für die Distanz d_{ab^*} gilt

$$d_{ab^*} = |x_a - x_{b^*}| = \min_{1 \leq b \leq n_B} |x_a - x_b| \quad (13)$$

(Vgl. D’Orazio et al. (2006), S. 41). Finden sich mehrere Spender mit dem gleichen Abstand, wird im Allgemeinen einer davon zufällig ausgewählt. Nach dieser Definition kann jeder Eintrag aus dem Spenderdatensatz B mehr als ein Mal als Spender verwendet werden. Deshalb wird es auch *unconstrained distance hot deck* genannt.

Eine andere Variante ist das *constrained distance hot deck*, bei dem jeder Eintrag aus B nur ein Mal als Spender ausgewählt werden darf. Hier ist es erforderlich, dass $n_A \leq n_B$ gilt, damit für jeden Eintrag aus A ein Eintrag aus B zur Verfügung steht. Im einfachsten Fall $n_A = n_B$ sollte die Spenderzuordnung so sein, dass

$$\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (d_{ab} w_{ab}) \quad (14)$$

unter folgenden Nebenbedingungen minimiert wird:

$$\sum_{b=1}^{n_B} w_{ab} = 1, \quad a = 1, \dots, n_A, \quad (15)$$

$$\sum_{a=1}^{n_A} w_{ab} = 1, \quad b = 1, \dots, n_B, \quad (16)$$

mit $w_{ab} \in \{0; 1\}$, wobei $w_{ab} = 1$, wenn das Paar (a, b) gematcht wurde und $w_{ab} = 0$, wenn a und b nicht einander zugeordnet wurden (Vgl. D’Orazio et al. (2006), S. 42).

Gibt es mehr Spender als Empfänger ($n_B > n_A$) sehen die Nebenbedingungen etwas anders aus:

$$\sum_{b=1}^{n_B} w_{ab} = 1, \quad a = 1, \dots, n_A, \quad (17)$$

$$\sum_{a=1}^{n_A} w_{ab} \leq 1, \quad b = 1, \dots, n_B, \quad (18)$$

mit $w_{ab} \in \{0;1\}$ und implizieren, dass $\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} w_{ab} = n_A$ gilt (Vgl. D’Orazio et al. (2006), S. 42).

D’Orazio et al. (2006) geben als Hauptvorteil der constrained Variante gegenüber der unconstrained an, dass die marginale Verteilung von der imputierten Variable \mathbf{Z} im gematchten Datensatz erhalten bleibt. Als Nachteil erwähnen sie, dass die durchschnittliche Distanz zwischen Spender- und Empfängerwert in der Matchingvariable X erwartungsgemäß größer ist als bei der unconstrained Variante. Und genau hier liegt der Ursprung für matching noise bei der distance hot deck Methode.

Auch bei dieser Methode kann zusätzlich eine Unterteilung in Klassen erfolgen, so dass nur die Distanzen zwischen Spender und Empfänger für die Einträge betrachtet werden, die in der gleichen Klasse bezüglich einer weiteren, kategorialen Matchingvariable X sind.

D’Orazio et al. (2006) weisen noch darauf hin, dass die distance hot deck Methode mit Vorsicht zu genießen ist, da auch hier der neu erzeugte Datensatz einem datengenerierenden Modell folgen kann, das nicht dem wahren datengenerierenden Modell entspricht. Das liegt daran, dass die (unconstrained) distance hot Methode einer nicht-parametrischen Regressionsschätzung anhand k NN mit $k = 1$ gleichkommt (Vgl. D’Orazio et al. (2006), S. 43-44).

An dieser Stelle sollte man sich noch mal insgesamt die Frage stellen, ob die Daten, die durch die hot deck Methoden erzeugt werden, wirklich aus der wahren, aber unbekanntem Verteilung mit Dichtefunktion $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ stammen. Für endliche Stichproben kann gezeigt werden, dass alle hot deck Methoden Datensätze erzeugen, die aus einer Verteilung stammen, die sich von der wahren unterscheidet (Vgl. D’Orazio et al. (2006), S. 46). Lässt man die Stichprobengröße gegen Unendlich laufen, nähert sich die Verteilung bei distance und rank hot deck Methoden der wahren Verteilung an.

2.3.3 Gemischte Methoden

Häufig werden auch parametrische und nicht-parametrische Methoden kombiniert, in dem Sinne, dass anfangs ein parametrisches Modell angenommen wird, welches anhand des parametrischen Makroansatzes geschätzt wird und dann wird ein künstlicher Datensatz mit Hilfe einer nicht-parametrischen Mikromethode erzeugt. Ziel ist es dadurch die positiven Eigenschaften der beiden Ansätze zu kombinieren, die Sparsamkeit von parametrischen Modellen im Vergleich zu nicht-parametrischen und die Robustheit bezüglich Fehlspezifikationen bei Modellen von nicht-parametrischen Methoden (Vgl. D’Orazio et al. (2006), S. 47). Die gemischten Methoden bestehen somit aus zwei Schritten. Im ersten Schritt werden die Parameter des parametrischen Modells geschätzt. Und im zweiten Schritt verwendet man eines der hot deck Verfahren bedingt auf die geschätzten Parameter aus dem ersten Schritt.

Stetige Variablen

Bei stetigen Variablen bestehen nach die gemischten Modelle aus den folgenden drei Schritten (Vgl. D’Orazio et al. (2006), S. 47):

- i) Die Parameter der Regression von \mathbf{Z} auf \mathbf{X} werden anhand Datensatz B geschätzt.
- ii) Basierend auf dieser geschätzten Regressionsfunktion wird für jedes $a = 1, \dots, n_A$ ein vorläufiger Wert $\tilde{\mathbf{z}}_a$ erzeugt.
- iii) Unter Berücksichtigung des vorläufigen Wertes $\tilde{\mathbf{z}}_a$ wird für jedes $a = 1, \dots, n_A$ ein beobachteter Wert \mathbf{z}_{b^*} , mit b^* aus B , für den a -ten Eintrag aus A mittels einer geeigneten distance hot deck Methode imputiert.

Kategoriale Variablen

Liegen kategoriale Daten vor, bestehen laut D’Orazio et al. (2006) die gemischten Modelle aus den folgenden zwei Schritten, wobei die drei vorliegenden kategorialen Variablen X , Y und Z hier univariat sein sollen, mit den Laufindizes i , j und k :

- i) Die erwarteten Zellhäufigkeiten werden anhand des loglinearen Modells, das zur CIA Annahme passt, geschätzt.
- ii) Mit Hilfe einer hot deck Methode werden passende Z Werte aus B für jeden Eintrag aus A ausgesucht. Ein möglicher Wert, der für den a -ten Wert gemäß (i, j, k) passend wäre, wird nur genommen, falls die geschätzte Zellhäufigkeit \tilde{n}_{ijk} nicht überschritten wird, andernfalls muss eine anderer Spender für den a -ten Eintrag gesucht werden.

3 Anwendungsbeispiel Lebensqualität

In diesem Abschnitt soll ein Anwendungsbeispiel vorgestellt werden um zum einen zu sehen in welchen Situationen statistisches Matching zum Einsatz kommen kann und zum anderen dessen Resultate zu sehen und zu bewerten. eurostat the european comission hat durch die beiden Autoren Aura Leulescu und Mihaela Agafitei 2013 ein Paper mit dem Titel "Statistical matching: a model based approach for data integration" herausgebracht, das unter anderem eine Pilotstudie beschreibt zum Thema statistisches Matching.

3.1 Ausgangslage

Kurz zu den Hintergrundinformationen dieser Studie. Wie durch einige jüngste europäische Initiativen deutlich wird, besteht ein immer größer werdendes soziales und politisches Interesse Wohlbefinden und Lebensqualität der Bevölkerung zu messen.

Das statistische Matching erscheint hier sehr nützlich, um die Informationen über Lebensqualität aus unterschiedlichen unabhängigen Quellen miteinander verbinden zu können, anstatt neue Erhebungen durchzuführen mit erheblich längeren Fragebögen um alle Dimensionen gemeinsam zu erfassen. Die Pilotstudie stellt nun einen ersten Test dar um festzustellen, ob der resultierende gematchte Datensatz sich für weiterführende Analysen eignet. Bei den zwei Datenquellen handelt es sich um the European Union Statistics on Income and Living Conditions (EU-SILC) und um the European Quality of Life Survey (EQLS).

Der EU-SILC Datensatz übernimmt hier die Rolle des Empfängers. Dieser Datensatz wurde von the European Commission and the Member States erhoben um Armut und soziale Ausgrenzung zu erfassen. Er enthält vor allem Dimensionen zu ökonomischem Wohlbefinden und den sogenannten AROPE-Index (People at-risk-of-poverty and social exclusion), der aus den folgenden drei Hauptindikatoren besteht. Die At-risk-of-poverty rate misst den Anteil der Personen, dessen verfügbares Einkommen sich unterhalb der Armutsgrenze befindet (Vgl. Leulescu und Agafitei (2013), S. 29). Die Severe material deprivation rate misst den Anteil Personen, die sich mindestens vier von neun abgefragten materiellen Grundbedürfnissen nicht leisten können (Vgl. Leulescu und Agafitei (2013), S. 29). Das sind zum Beispiel eine Woche im Jahr den Urlaub nicht zu Hause zu verbringen, sich jeden zweiten Tag ein Essen mit Fleisch, Huhn, Fisch oder vegetarisch ähnlichen Ersatzprodukten leisten können oder sich eine Waschmaschine leisten können. Und die Low work intensity rate misst den Anteil Personen, die in einem Haushalt leben, in dem Erwachsene weniger als 20% ihres möglichen Arbeitspotentials (bzgl. des Wertes eines Referenzjahres) arbeiten (Vgl. Leulescu und Agafitei (2013), S. 29). Hier geht es also um das Thema Unterbeschäftigung. Außerdem wurden noch weitere Variablen bezüglich den Wohnbedingungen, der Arbeit, der Gesundheit, der Demografie und der Bildung erhoben. Um die Lebensqualität in diesem Datensatz noch besser abzubilden, sollen Variablen imputiert werden, die Konzepte wie das emotionale Wohlbefinden, die soziale Teilhabe und das Vertrauen in Institutionen erfassen. Dazu wird der zweite Datensatz benötigt.

Der EQLS-Datensatz ist somit in dieser Pilotstudie der Spender. Er wurde von EuroFound erhoben und enthält 160 Indikatoren für emotionales Wohlbefinden die Bereiche wie berufliche und soziale Netzwerke, Lebenszufriedenheit, Zufriedenheit und Zugehörigkeitsgefühl und Teilnahme am gesellschaftlichen Leben abdecken. Dieser Datensatz kann somit den vorhergehenden durch wertvolle subjektive Indikatoren bereichern. Dabei wurden die folgenden vier Variablen als die Zielvariablen des Matchings in den EU-SILC Datensatz bezeichnet. Die allgemeine Lebenszufriedenheit (Overall life satisfaction), das Vertrauen in Institutionen (Trust in institutions), das Gefühl von Anerkennung (Recognition) und das Gefühl sozialer Ausgeschlossenheit (Social exclusion). Es gibt einige Variablen, die in beiden Datensätzen erfasst wurden. Dazu gehören Variablen aus den Bereichen Demografie, Haushaltzusammensetzung, Arbeit, Gesundheit,

Wohnstätte, materieller Mangel, Umwelt und Einkommen.

| | X | Y | Z |
|---------|-----------------|-------------|--------------------------|
| EU-SILC | Demografie usw. | AROPE-Index | |
| EQLS | Demografie usw. | | Lebenszufriedenheit usw. |

Tabelle 1: *Datensituation der beiden verwendeten Datensätze EU-SILC und EQLS zusammengefasst dargestellt.*

Diese Daten wurden von insgesamt 27 EU-Ländern erhoben. In der Pilotstudie wird das statistische Matching mittels zwei verschiedener Methoden für die beiden Länder Finnland und Spanien durchgeführt und verglichen.

3.2 Statistisches Matching

Im Folgenden werden nun die vier Schritte des statistischen Matchings beschrieben. Dazu gehören die Harmonisierung und Abstimmung der Quellen, die Analyse der Erklärungskraft der gemeinsamen Variablen, die Auswahl der Matchingmethoden und die Evaluation der Ergebnisse. Die ersten beiden Schritte beschäftigen sich mit der Auswahl geeigneter Matchingvariablen, zwei sehr wichtige Schritte, da sie die Grundlage bilden, auf der das statistische Matching stattfindet. Geeignet bedeutet hier, dass die gemeinsamen Variablen zum einen kohärent sein müssen und zum anderen eine hohe Aussagekraft bezüglich der zu imputierenden Variablen haben sollen.

Im **ersten Schritt** geht es darum abzuklären welche gemeinsamen Variablen der beiden Datensätze kohärent genug sind, um als Matchingvariablen in Frage zu kommen. Um zu überprüfen welche der gemeinsamen Variablen auch tatsächlich das gleiche erfasst haben, wurden die Frageformulierungen, die Definitionen der gemessenen Konzepte, die Messskalen sowie die Richtlinien der Messungen betrachtet. Denn hier können sich bereits große Unterschiede verbergen, die einen Vergleich vermeintlich gleicher Variablen unbrauchbar machen. Außerdem wurden die marginalen Verteilungen anhand von 95% Konfidenzintervallen verglichen und geeignete statistische Tests gemacht, um übereinstimmende Variablen zu finden. Dabei sind schon einige Variablen als potentielle Matchingvariablen ausgeschieden. Wobei in den beiden Ländern nicht unbedingt dieselben Variablen ausgeschieden sind. Die ein oder andere Variable konnte auch durch geeignete Transformationen beibehalten werden.

Im **zweiten Schritt** wird überprüft welche der übrigen Variablen auch einen Zusammenhang mit den Zielvariablen haben, die imputiert werden sollen, der stark genug ist, damit ein Matchen anhand dieser Variablen sinnvoll ist. Dazu wurden zuerst paarweise Korrelationen zwischen den gemeinsamen Variablen und den Zielvariablen berechnet und Tests mit der Nullhypothese "kein Zusammenhang bzw. Unabhängigkeit" gemacht. Anschließend wurde betrachtet welche der gemeinsamen Variablen einen starken Zusammenhang mit einem beträchtlichen Anteil der Zielvariablen hat. Der

Großteil der gemeinsamen Variablen konnte nicht beide Voraussetzungen, Kohärenz und hohe Erklärungskraft, erfüllen, so dass bei strenger Auswahl die folgenden Matchingvariablen für Spanien und Finnland vorerst übrig bleiben¹.

| Spanien | Finnland |
|---|--|
| Geschlecht | Geschlecht |
| Alter | Alter |
| NUTS 2 Region | NUTS 2 Region |
| Fleisch oder Fisch jeden 2. Tag möglich | Beschäftigungsstatus |
| selbst genannter Berufsstatus | monatliches Nettoeinkommen des Haushalts |
| Wohnbesitzverhältnis des Haushalts | genereller Gesundheitszustand (re-kategorisiert) |

Tabelle 2: Die übrigen Matchingvariablen für Spanien und Finnland, nach strenger Überprüfung der Kohärenz und Erklärungskraft (Vgl. Leulescu und Agafitei (2013), S. 30-35).

Viele Matchingmethoden basieren, wie im Theorieteil schon beschrieben, auf der Annahme bedingter Unabhängigkeit (CIA). Um diese Annahme bezüglich des AROPE-Index aus dem EU-SILC Datensatz und der Lebenszufriedenheit aus dem EQLS Datensatz zu überprüfen, wurde noch eine logistische Regression zur Erfassung von multivariaten Zusammenhängen berechnet. Die abhängigen Variablen waren die subjektiven Variablen, die imputiert werden sollten, welche offensichtlich dichotomisiert wurden. Die resultierende Erklärungskraft des Modells war, trotz guter Prädiktoren für die Zielvariablen, nicht besonders hoch. Der Fit zwischen vorhergesagter Wahrscheinlichkeiten und beobachteter Werte lag bei etwa 65% (Vgl. Leulescu und Agafitei (2013), S. 35). Das muss jedoch noch nicht bedeuten, dass die Annahme der bedingten Unabhängigkeit nicht zutreffen kann. Betrachtet man die resultierenden Odds ratios mit Lebenszufriedenheit als abhängige Variable, findet man einige Variablen, die einen starken Effekt haben, wie Gesundheit, Beziehungsstatus und Beschäftigungsstatus. Jedoch findet sich, selbst wenn auf diese Variablen kontrolliert wird, noch eine starke Korrelation zwischen Lebenszufriedenheit und allgemein gesagt schlechten wirtschaftlichen Bedingungen. Damit sind die Items für materiellen Mangel gemeint, die in beiden Datensätzen erhoben wurden. Da also auch unter Kontrolle einiger guter Prädiktoren aus den Reihen der gemeinsamen Variablen ein starker Zusammenhang zwischen Lebenszufriedenheit und einiger Items, die in den AROPE-Index mit eingehen, besteht, spricht das eher gegen die Annahme, dass die Lebenszufriedenheit bedingt auf die Matchingvariablen unabhängig von dem AROPE-Index ist (Vgl. Leulescu und Agafitei (2013), S. 35). Das heißt, dass davon ausgegangen werden muss, dass die CIA hier

¹NUTS = "Abk. für Nomenclature des Unités Territoriales Statistiques; hierarchische Systematik der statistischen Gebietseinheiten, vor mehr als 30 Jahren von Eurostat eingeführt, um für die Erstellung regionaler Statistiken für die Europäische Union eine einheitliche und konsistente territoriale Untergliederung zu schaffen." (Gabler Wirtschaftslexikon)

nicht zutrifft. Leulescu und Agafitei (2013) beschreiben Näherungsvariablen (proxy variables) als einzige Lösung für dieses Problem. Es sollen nun die drei Items aus dem AROPE-Index, die ebenfalls im EQLS Datensatz erhoben wurden, dazu verwendet werden die Schätzung der gemeinsamen Verteilung von Lebenszufriedenheit und AROPE Index zu verbessern. Um diesen Plan zu stützen wurden noch einige Analysen im Vorfeld gemacht. Diese zeigen, dass die drei Items eine hohe Vorhersagekraft (>90% Übereinstimmung zwischen Vorhersage und Beobachtung) für den gesamten Index haben. Das bedeutet, dass sie tendenziell den Zusammenhang zwischen AROPE-Index und den zu imputierenden Variablen abschwächen werden (Vgl. Leulescu und Agafitei (2013), S. 36-37).

Da selten beide Kriterien, Kohärenz und hohe Erklärungskraft, auf die gemeinsamen Variablen zutreffen, wurden schließlich die Imputationen mittels zwei Gruppen von Matchingvariablen gemacht. Die erste Gruppe enthält nur Variablen, bei denen ganz streng auf die Einhaltung beider Kriterien geachtet wurde (Siehe Tabelle 2). Und bei der zweiten Gruppe wurde die Kohärenz nicht ganz so strikt eingehalten, dafür wurde mehr Wert auf Variablen gelegt, die wichtig für die Einhaltung der CIA erschienen (Vgl. Leulescu und Agafitei (2013), S. 37).

Im **dritten Schritt** werden die Matchingmethoden ausgewählt um die Variablen zu den Themen Lebenszufriedenheit, Vertrauen in Institutionen und soziale Ausgrenzung von EQLS in EU-SILC zu imputieren. Leulescu und Agafitei (2013) haben sich für folgende zwei Methoden entschieden, dessen Ergebnisse anschließend verglichen werden. Für das Matching anhand der ersten Gruppe Matchingvariablen (siehe Tabelle 2) wurde die *distance hot deck Methode* gewählt und zwar die unconstrained Variante. Das heißt hier wird für jedes Objekt aus dem Empfängerdatensatz EU-SILC das Objekt aus dem Spenderdatensatz EQLS gesucht, das die geringste Distanz bezüglich der Werte der Matchingvariablen aufweist. Wobei es in diesem Fall erlaubt ist, dass Spender mehrmals zum Einsatz kommen, da die unconstrained Variante verwendet werden soll. Da hier qualitative Variablen, die alle zu binären Variablen transformiert wurden, vorliegen, können die Standard-Distanzmaße wie die euklidische Distanz oder die Mahalanobisdistanz, die für stetige Variablen sind, nicht angewandt werden. Leulescu und Agafitei (2013) greifen deshalb auf ein Distanzmaß für binäre Variablen zurück, das auf dem Koeffizienten von Dice (similarity coefficient) basiert und folgendermaßen definiert ist

$$D_{ij} = \sqrt{1 - S_{ij}}, \text{ wobei } S_{ij} = \frac{2a}{2a + b + c} \quad (19)$$

und a für die Anzahl der Indikatoren, für die $i = 1$ und $j = 1$ gilt, b für die Anzahl der Indikatoren, für die $i = 1$ und $j = 0$ gilt und c für die Anzahl der Indikatoren, für die $i = 0$ und $j = 1$ gilt, steht (Vgl. Leulescu und Agafitei (2013), S. 37-38).

Bei der zweiten Methode handelt es sich um ein gemischtes Modell, das sogenannte predictive mean matching. Im ersten Schritt werden die Parameter der Regression, hier

eine logistische Regression, von den Zielvariablen auf die Matchingvariablen anhand EQLS geschätzt. Diese Parameter werden anschließend dazu genutzt um vorläufige Schätzwerte der Zielvariablen für die Objekte aus EU-SILC zu erhalten. Schließlich wird für jeden Empfänger aus EU-SILC der Spender aus EQLS, der gemäß dem Distanzmaß der distance hot deck Methode am nächsten ist. Hier dient Gruppe zwei der Matchingvariablen als Menge der Kovariablen, die im ersten Schritt in die Regression mit eingehen. Diese besteht aus den folgenden Variablen:

- Alter, Geschlecht, Bildung, Familienstand
- Beschäftigungsstatus
- Gesundheitszustand
- 3 Items zum materiellen Mangel, über die Runden kommen können

Schließlich wird im **vierten Schritt** die Qualität der Ergebnisse beurteilt. Leulescu und Agafitei (2013) analysieren die Qualität der Ergebnisse anhand der vier Zielvariablen Lebenszufriedenheit, Anerkennung im Leben, Vertrauen in die Presse und Vertrauen in die Regierung. Die erste Feststellung ist, dass es beiden Methoden gelingt die Randverteilungen der Variablen vor und nach des Matchings vernünftig zu erhalten (siehe Abb. 3).

Es ist allerdings auch festzuhalten, dass die Ergebnisse bezüglich der gemeinsamen Verteilung der nicht gemeinsam beobachteten Variablen sich doch eher voneinander unterscheiden. Um die Unterschiede besser darstellen zu können haben Leulescu und Agafitei (2013) die vier, hier betrachteten, Zielvariablen in binäre Variablen transformiert (z.B.: hohe/niedrige Lebenszufriedenheit $>/\leq 6$). Mit Hilfe von Balkendiagrammen haben sie dann die Unterschiede zwischen der Gesamtpopulation und der Untergruppe der Personen mit materiellem Mangel in den vier Zielvariablen dargestellt (siehe Abb. 4). Dabei weisen sie vor allem darauf hin, dass für Spanien in den Variablen Lebenszufriedenheit und Anerkennung im Leben kein deutlicher Unterschied zwischen der gesamten Population und der Teilpopulation zu erkennen ist, wenn man die Ergebnisse der distance hot deck Methode betrachtet. Dies ist aber eher darauf zurückzuführen, dass hier im Vorfeld die Annahme bedingter Unabhängigkeit getroffen wurde und nicht darauf, dass es keinen Zusammenhang zwischen Lebenszufriedenheit bzw. Anerkennung und den zwei betrachteten Gruppen gibt. Denn unter Verwendung von proxy Variablen im Rahmen der zweiten Methode, wird durchaus deutlich, dass Personen, die materiellen Mangel erleiden, eher dazu tendieren unzufriedener mit ihrem Leben zu sein (Vgl. Leulescu und Agafitei (2013), S. 40).

Schließlich haben Leulescu und Agafitei (2013) noch Analysen zur Unsicherheit der Schätzungen gemacht. Dazu haben sie Unsicherheitsintervalle für die Schätzer der

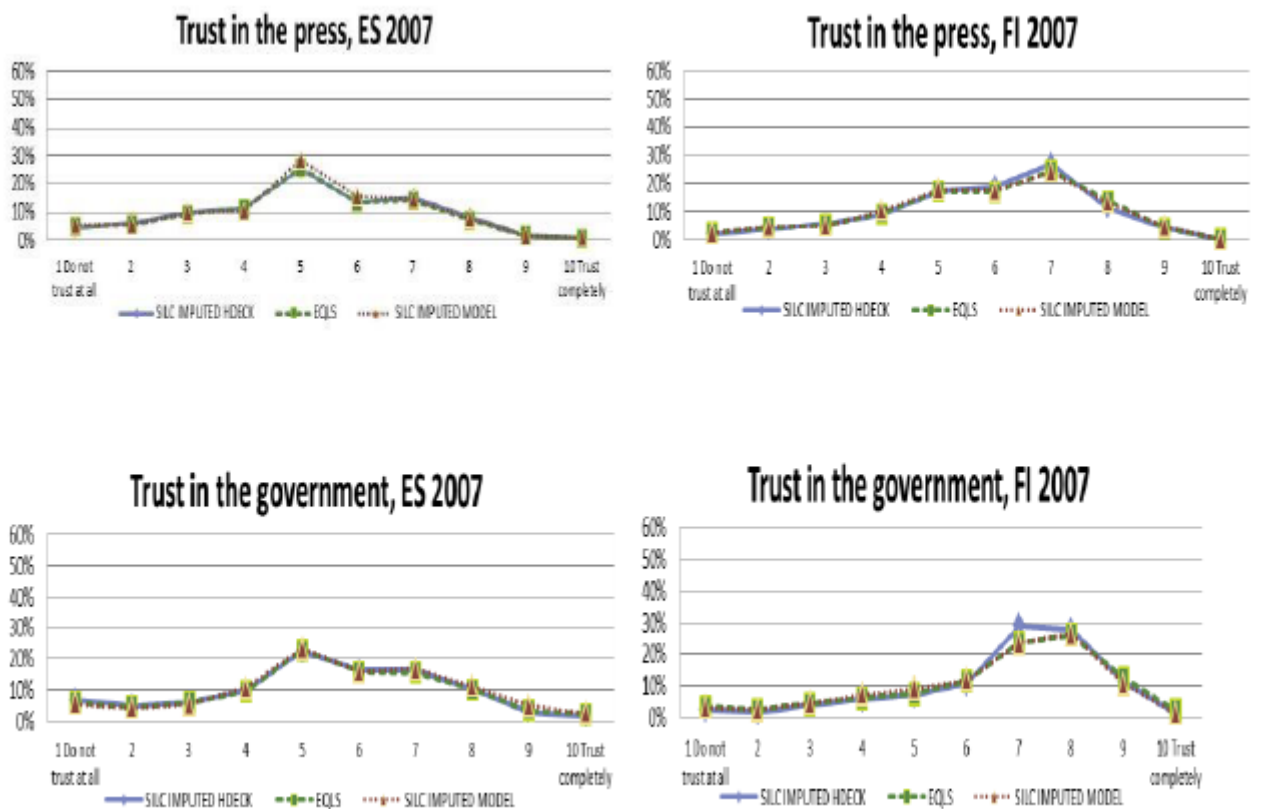


Abb. 3: Randverteilungen des beobachteten Datensatzes EQLS und der gematchten Datensätze mit den beiden Methoden distance hot deck und dem modellbasierten predictive mean matching, hier nur für das Vertrauen in Presse und Regierung dargestellt (Vgl. Leulescu und Agafitei (2013), S. 39)



Abb. 4: Vergleich der Verteilungen der Zielvariablen und der Teilpopulation mit großem materiellem Mangel und der Gesamtpopulation (Leulescu und Agafitei (2013), S. 40)

gemeinsamen Verteilung von Lebenszufriedenheit (auf einer Skala von 1-10) und extremen materiellem Mangel (ja/nein) berechnet. Allgemein ist die Intervallgröße akzeptabel. Jedoch berichten die Autorinnen, dass die Intervalle für kleinere Gruppen, zum Beispiel für die einzelnen Abstufungen des materiellen Mangels, nicht informativ sind (Vgl. Leulescu und Agafitei (2013), S. 42).

3.3 Fazit

Aus der Pilotstudie können einige wichtige Erkenntnisse gezogen werden. Erstens ist der Erhalt der Randverteilung der imputierten Variablen beim statistischen Matching zwar recht leicht, jedoch ist dies keine Garantie für gute Qualität (Vgl. Leulescu und Agafitei (2013), S. 42-43). Zweitens sollte die Wahl der Matchingvariablen sorgfältig und gut überlegt getroffen werden, da sie entscheidend für den Erfolg des statistischen Matchings ist. Denn fehlen wichtige Prädiktoren werden unter CIA die Zusammenhänge zwischen den gemeinsam beobachteten Variablen eher unterschätzt (Vgl. Leulescu und Agafitei (2013), S. 43).

Hier waren die Daten bereits vorhanden und man konnte an der Datenlage nichts mehr ändern, worunter die Qualität des gematchten Datensatzes sicher gelitten hat, da einige gemeinsame Variablen aus verschiedenen Gründen als Matchingvariablen ausgeschieden sind. Bei zukünftigen Datenerhebungen hat man jedoch die Chance noch auf das ein oder andere zu achten, so dass mehr potentielle Matchingvariablen zur Verfügung stehen könnten. Daher wünschen sich Leulescu und Agafitei (2013) mehr Kommunikation zwischen den einzelnen Studien um Standardfragen besser harmonisieren zu können.

4 Zusammenfassung

Nach Klärung des allgemeinen Sachverhaltes beim statistischen Matchings wurde ein begrenzte Auswahl an Methoden vorgestellt, die alle der CIA unterliegen. Dabei ist besonders festzuhalten, dass die CIA eine sehr starke und restriktive Annahme ist, dessen man sich bewusst sein sollte, wenn man derartige Methoden verwendet. So sind auch die Ergebnisse des statistischen Matchings selbst, sowie die darauf folgenden Analysen anhand des resultierenden Datensatzes, kritisch zu betrachten und mit Vorsicht zu genießen. Im Anwendungsbeispiel ist auch nochmal deutlich geworden wie wichtig die Wahl der Matchingvariablen ist, aber auch wie schwer es sein kann überhaupt geeignete Variablen zu finden. Da das statistische Matching aber sicher auch in Zukunft noch viele Anwender finden wird, wenn nicht sogar noch mehr als bisher, werden es da vermutlich noch einige Neuerungen auf diesem Gebiet geben. So gibt es bereits einige Methoden, die die CIA umgehen, als Alternative zu den hier betrachteten Methoden. Und es ist durchaus vorstellbar, dass bei zukünftigen Erhebungen bereits an die Möglichkeit des Matchings gedacht wird und damit eine verbesserte Datengrundlage

geschaffen wird, die mithelfen kann Ergebnisse mit besserer Qualität zu erhalten.

5 Literaturverzeichnis

D’Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley und Sons, Ltd., Sussex, England, Kap. 1 und 2.

Fahrmeir, L., Kunstler, R., Pigeot, I., Tutz, G., (2010). *Statistik: Der Weg zur Datenanalyse*. Springer Verlag, Berlin, 7.Auflage, S. 371-372.

Leulescu, A. und Agafitei, M. (2013). *Statistical matching: a model based approach for data integration*. Publications Office of the European Union, Luxemburg, Kap. 2.

Meinfelder, F. (2013). *Datenfusion: Theoretische Implikationen und praktische Umsetzung*. In: Riede, T., Bechtold, S. und Ott, N. (Hrsg.), *Weiterentwicklung der amtlichen Haushaltsstatistiken*, SCIVERO Verlag, Berlin, S. 83-98.

Gabler Wirtschaftslexikon, Springer Gabler Verlag (Herausgeber), Stichwort: NUTS, online im Internet (22.11.2014): <http://wirtschaftslexikon.gabler.de/Archiv/11703/nuts-v8.html>