Thema:

Datenfusion von EU-SILC und HBS: Vergleich zwischen Random Hot-Deck und Predictive Mean Matching im Rahmen einer Simulationsstudie

Masterarbeit

im Studiengang

Survey-Statistik

in der Fakultät

Sozial- und Wirtschaftswissenschaften

der Otto-Friedrich-Universität Bamberg

Verfasser: Jannik Schaller

Prüfer: Dr. Martin Messingschlager

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

In der amtlichen Statistik der Europäischen Union und ihrer Mitgliedstaaten wird im Rahmen des Projekts „Income, Consumption and Wealth“ (kurz: ICW) seit einigen Jahren versucht, die gemeinsame Verteilung von Einkommen, Konsumausgaben und Vermögen der privaten Haushalte zu messen. Im Besonderen ist hiermit eine OECD-Eurostat-Expertengruppe betraut, deren Arbeit auf Empfehlungen der „Commission on the Measurement of Economic Performance and Social Progress“ von 2009 zurückgeht (Balestra 2018). Zur Beurteilung des sozialen und wirtschaftlichen Lebensstandards der privaten Haushalte in der EU empfiehlt der Bericht der betreffenden Kommission unter anderem, die Determinanten Einkommen, Konsumausgaben und Vermögen gleichermaßen zu betrachten und zu untersuchen (Stiglitz et al. 2009). Momentan existiert in der amtlichen Statistik jedoch (noch) keine befriedigende Datenquelle, die alle drei Komponenten gemeinsam erfasst. Dementsprechend muss die Beurteilung der gemeinsamen Verteilung von Einkommen, Konsumausgaben und Vermögen durch anderweitige statistische Verfahren vorgenommen werden.

Eine Möglichkeit stellt dabei die Fusionierung und Verknüpfung entsprechender Datenquellen dar, die die jeweiligen Determinanten umfassend widerspiegeln. Dabei stehen drei Datenquellen zur Verfügung, die jeweils eine der drei Komponenten besonders detailliert erfassen: Für *Einkommen* die Studie „European Union Statistics on Income and Living Conditions“ (kurz: *EU-SILC*), für die *Konsumausgaben* der „Household Budget Survey“ (kurz: *HBS*[1]) sowie für das *Vermögen* der „Household Finance and Consumption Survey“ (kurz: *HFCS*). Der erste Schritt zu einer einheitlichen Datenquelle ist dabei die initiale Fusionierung von EU-SILC und HBS, also die gemeinsame Betrachtung von Einkommen und Konsumausgaben der privaten Haushalte (Lamarche 2017[2]). Damit verbindet die amtliche Statistik in Europa zusätzlich das Ziel, Armutsrisiken und Armutsindikatoren präziser messen zu können, als mit dem bisherigen Messinstrument, welches sich weitgehend nur auf das Einkommen selbst beschränkt und die Konsumausgaben außer Acht lässt (Serafino und Tonkin 2017).

Dementsprechend widmet sich auch die vorliegende Arbeit der Datenfusion von EU-SILC und HBS, um den primären und initialen Fokus der amtlichen Statistik, der sich zunächst besonders auf die gemeinsame Verteilung von Einkommen und Konsumausgaben bezieht, aufzugreifen. Die Arbeit baut dabei auf dem aktuellen Forschungsstand von Eurostat, dem Statistischen Amt der Europäischen Union, auf. Um EU-SILC und HBS zu fusionieren, ver-

[1] In Deutschland entspricht der HBS der Einkommens- und Verbrauchsstichprobe (kurz: *EVS*).

[2] Beim Artikel von Lamarche (2017) handelt es sich um eine vorläufige Version.

folgt Eurostat die Ergänzung des vorhandenen EU-SILC-Datensatzes um die spezifischen, im HBS enthaltenen Konsumvariablen, woraus ein vollständiger, fusionierter Mikrodatenfile aus Einkommens- und Konsumausgaben resultieren soll. Dementsprechend stellt EU-SILC den Empfänger- beziehungsweise Rezipientendatenfile dar, dem Informationen zu den Konsumausgaben hinzugefügt werden sollen. Der HBS ist wiederum der Spender- beziehungsweise Donorendatenfile, aus dem die entsprechenden Konsuminformationen stammen. Abbildung 1 stellt die daraus resultierende Fusionssituation dar. Zur Notation sei erwähnt, dass $\mathbf{Y}$ die Menge der spezifischen, lediglich in EU-SILC vorhandenen Variablen widerspiegelt, während $\mathbf{Z}$ die Menge an spezifischen Variablen bezeichnet, die lediglich im HBS enthalten sind. Jedoch stellen im fusionierten Datensatz die $\mathbf{Z}$-Variablen aus dem HBS eine durch den Fusionsprozess bedingte, künstliche Verteilung dar und sind demnach mit $\widetilde{\mathbf{Z}}$ gekennzeichnet. $\mathbf{X}$ bezeichnet die Menge an Variablen, die die beiden Datenfiles gemeinsam haben, beispielsweise das Alter der befragten Referenzperson eines Haushalts (Serafino und Tonkin 2017).

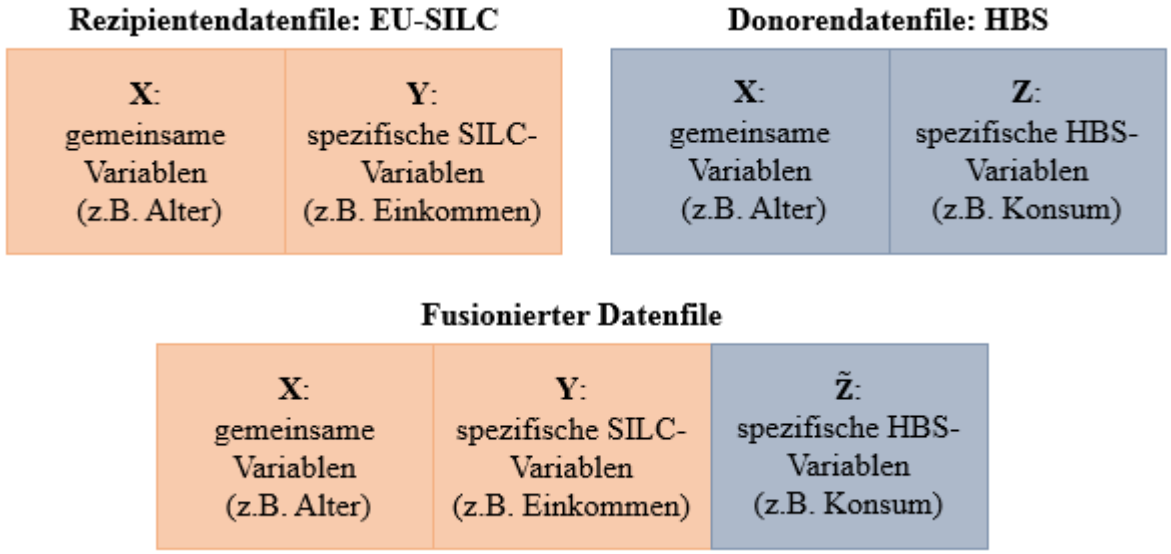

Abbildung in Anlehnung an Serafino und Tonkin (2017: 10).

Abbildung 1: Datenfusionssituation von EU-SILC und HBS

Im Rahmen einer Datenfusion werden die zu fusionierenden Datenfiles anhand der gemeinsamen $\mathbf{X}$-Variablen zusammengeführt, wobei hierfür verschiedene Methoden existieren. Um einen wie in Abbildung 1 dargestellten, fusionierten Mikrodatenfile zu generieren, hat Eurostat bereits eine Fusionierung der Daten von EU-SILC und HBS, jeweils aus dem Jahre 2015, vorgenommen und diese in Form von `R`-Programmcodes dem Statistischen Bundesamt in Deutschland zur Verfügung gestellt. Hierbei ist die amtliche Statistik in Deutschland, namentlich das Statistische Bundesamt, an einer wissenschaftlichen Bewertung des von Eu-

rostat verwendeten Fusionsverfahrens, insbesondere mit Blick auf die Fusionierung der deutschen EU-SILC- und HBS-Daten von 2015, interessiert. Dies versucht die vorliegende Arbeit zu leisten. Hierbei ist zunächst anzumerken, dass Eurostat eine univariate Datenfusion vorgenommen hat und lediglich *eine* Konsumvariable aus dem HBS, namentlich die Gesamtkonsumausgaben des Haushalts, mit der Rezipientenstichprobe, EU-SILC, verknüpft. Die von Eurostat dabei angewandte Fusionsmethode entspricht vom Prinzip her einem Random Hot-Deck-Verfahren (D'Orazio et al. 2006: 37-39; Lamarche 2017).

Aus statistisch-wissenschaftlicher Sicht kann jedoch in Zweifel gezogen werden, dass der von Eurostat verwendete Fusionsalgorithmus tatsächlich zu validen Ergebnissen führt und die gemeinsame Verteilung von Einkommen und Konsum adäquat widerspiegelt. Der Hauptkritikpunkt des Eurostat-Verfahrens besteht darin, dass jede der ausgewählten, in EU-SILC und HBS gemeinsam beobachteten **X**-Variablen mit *gleichem* Gewicht in den Fusionsprozess eingeht, ungeachtet der Frage, ob die entsprechenden Merkmale überhaupt die *gleiche* Relevanz für die Fusionierung aufweisen. Dieses Problem könnte das Fusionsergebnis negativ beeinflussen und die interessierende, gemeinsame Verteilung von Einkommen und Konsum verzerrt widerspiegeln. Da sich jedoch die amtliche Statistik in Europa von der gemeinsamen Betrachtung von Einkommen und Konsumausgaben, wie oben beschrieben, einerseits ein tieferes Verständnis des sozialen und wirtschaftlichen Lebensstandards in der EU, andererseits eine präzisere Erfassung von Armutsindikatoren erhofft, würde ein verzerrtes Fusionsergebnis fehlerhafte Schlussfolgerungen nach sich ziehen (Stiglitz et al. 2009; Serafino und Tonkin 2017). Daher versucht die vorliegende Arbeit das Eurostat-Verfahren, sofern möglich, mithilfe eines alternativen Verfahrens zu optimieren. Diesbezüglich könnte Predictive Mean Matching (kurz: PMM) eine vielversprechende Fusionsalternative darstellen, da dieses Verfahren die potentiell ungleiche Relevanz der für den Fusionsprozess ausgewählten, gemeinsamen **X**-Variablen berücksichtigt (Rubin 1986; Little 1988; Koller-Meinfelder 2009: 33-34; Meinfelder 2013).

Dementsprechend besteht das Ziel der vorliegenden Arbeit darin, die Random Hot-Deck-Methode von Eurostat wissenschaftlich zu beurteilen und zu untersuchen, inwiefern Predictive Mean Matching dazu imstande ist, das Eurostat-Verfahren zu optimieren. Dies geschieht im Rahmen einer Simulationsstudie, in der das „wahre" Fusionsergebnis bekannt ist und anhand derer Random Hot-Deck gegen Predictive Mean Matching getestet werden kann. Als Basis für die Simulationsstudie dienen Daten von EU-SILC aus dem Jahre 2013, die als Public Usefiles frei zugänglich und verfügbar sind. Im Zuge der Simulation werden in die-

ser Arbeit zwei HBS-Konsumvariablen mit EU-SILC verknüpft, um aufzuzeigen, dass auch eine multivariate Fusion unter Verwendung von Random Hot-Deck und Predictive Mean Matching möglich ist, während Eurostat lediglich den univariaten Fall betrachtet, also nur *ein* spezifisches Konsummerkmal aus dem HBS in EU-SILC hineinfusioniert. Die multivariate Datenfusion kann im Rahmen der Simulationsstudie, die Random Hot-Deck und Predictive Mean Matching vergleichen soll, integriert und exemplarisch anhand des bivariaten Falls erfolgen. Dies ergänzt den formulierten Untersuchungsauftrag der vorliegenden Arbeit, der sich zusammenfassend in folgender Fragestellung widerspiegelt: *Kann Predictive Mean Matching das von Eurostat verwendete Random Hot-Deck-Verfahren zur Datenfusion von EU-SILC und HBS optimieren?*

Die Beantwortung der Fragestellung soll entlang folgender Vorgehensweise gewährleistet werden: In Kapitel 2 wird ein theoretischer Überblick über die Datenfusionsthematik gegeben, der neben einer begrifflichen Einordnung der Datenfusion als Problem fehlender Daten die zentrale Annahme sowie die entsprechenden Validitätsstufen einer Datenfusion beleuchtet. Darauffolgend liefert Kapitel 3 zunächst einen kurzen Überblick über die vorhandenen und in der bisherigen Forschung sowie bei Eurostat besonders diskutierten Datenfusionsansätze, bevor eine detaillierte Erläuterung der für diese Arbeit relevanten Verfahren, Random Hot-Deck und Predictive Mean Matching, erfolgt, deren theoretische Implikationen diskutiert und in eine Arbeitshypothese zu überführen sind. Insofern integriert sich in den Kapiteln 2 und 3 parallel ein Überblick über den Forschungsstand, sowohl mit Bezug zu Datenfusionen allgemein, als auch hinsichtlich konkreter Fusionsalgorithmen mit besonderem Fokus auf Random Hot-Deck und Predictive Mean Matching. Anschließend wird in Kapitel 4 das Simulationsdesign zur Überprüfung der formulierten Arbeitshypothese beschrieben, wobei es insbesondere die vorhandene Datenbasis, die daraus auszuwählenden Variablen, die konkrete Simulationsmethode, die einer Monte-Carlo-Simulation entspricht, sowie die Programmgrundlage zu erläutern gilt. In Kapitel 5 werden die Ergebnisse der durchgeführten Simulationsstudie dargelegt, bewertet und, auch mit Blick auf Handlungsperspektiven für die amtliche Statistik, diskutiert, bevor diese in Kapitel 6 zusammengefasst, eingeordnet und kritisch reflektiert werden sowie die Beantwortung der Fragestellung gewährleisten sollen.

2 Datenfusion: Ein Überblick

2.1 Datenfusion als spezifisches Datenausfallmuster

Um sich dem formulierten Untersuchungsauftrag anzunähern, ist zunächst eine wissenschaftlich konsistente Definition des Begriffs der Datenfusion erforderlich. Denn häufig werden Datenfusionen, insbesondere in der Markt- und Sozialforschung, verkürzt mit einer „Verschmelzung“ (siehe z.B. Koschnick 1995: 309) über Nearest-Neighbour-Verfahren beziehungsweise „Statistische Zwillinge“ (siehe z.B. Bacher 2002: 40) gleichgesetzt, also mit Fusionsmethoden, die entlang von Distanzmaßen Datenquellen über möglichst ähnliche Beobachtungen versuchen zu fusionieren. Beides stellen jedoch lediglich mögliche Verfahren dar, um Datenfusionen durchzuführen; ebenso sagen derartige Definitionsversuche nichts über die Analyseziele einer Datenfusion aus (Meinfelder 2013; Cielebak und Rässler 2014).

Daher existiert in der statistischen Literatur eine allgemeinere und für Statistikzwecke zielführende Definition, der beispielsweise Rässler (2002), Meinfelder (2013) sowie Cielebak und Rässler (2014) folgen und worauf auch die vorliegende Arbeit zurückgreift: Demnach wird eine Datenfusion als *spezifisches Datenausfallmuster* definiert, das sich durch „Untereinanderbinden“ zweier oder mehrerer Datenquellen, die unabhängig voneinander entstanden sind, ergibt. Im Falle von zwei zu fusionierenden Datenquellen, nennen wir sie vereinfacht A und B, existieren somit eine Menge an gemeinsamen Variablen $\mathbf{X}$, die in beiden Datensätzen vorhanden sind, eine Menge an spezifischen Variablen $\mathbf{Y}$, die lediglich in Datenquelle A vorkommen sowie eine Menge an spezifischen Variablen $\mathbf{Z}$, die nur in Datenquelle B vorhanden sind. Das Analyseziel einer Datenfusion bezieht sich dabei stets auf die *gemeinsame* Verteilung der Menge an *nicht gemeinsam* beobachteten Variablen $\mathbf{Y}$ und $\mathbf{Z}$. In Abbildung 2 ist eine solche Datenfusionssituation sowie das entsprechende Datenausfallmuster dargestellt (Rässler 2002: 7-8; Meinfelder 2013; Cielebak und Rässler 2014).

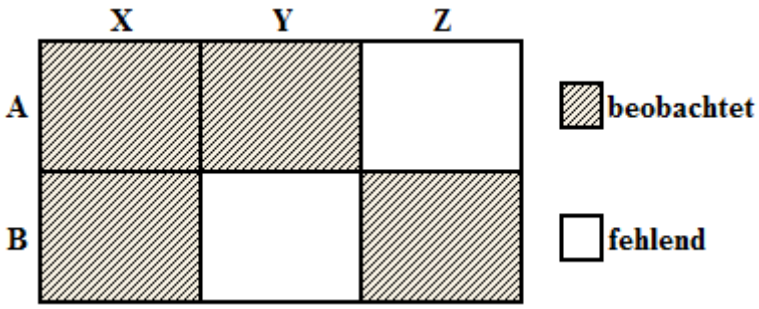

Abbildung in Anlehnung an Meinfelder (2013: 85).

Abbildung 2: Datenfusion als spezifisches Datenausfallmuster

Durch das „Untereinanderbinden“ der Datenquellen entsteht ein durch den Forscher absichtlich herbeigeführtes, spezifisches Datenausfallmuster, weshalb es in der Literatur als „Missing-by-design“ (siehe z.B. Koller-Meinfelder 2009: 9) kategorisiert, häufig aber auch unter direktem Bezug zur Datenfusion als „File Matching Pattern“ (siehe z.B. Raghunathan 2016: 4) bezeichnet wird. Dabei ist der Datenausfallmechanismus[3], dessen Ausdifferenzierung auf Rubin (1976) sowie auf Little und Rubin (2002: 11-19) zurückgeht, Missing Completely At Random (MCAR), solange die Datenquellen A und B Zufallsstichproben derselben Grundgesamtheit darstellen (D’Orazio et al. 2006: 4-7; Meinfelder 2013). Damit wäre der Datenausfall absolut zufällig und ignorierbar[4]. Die Ignorierbarkeit begründet sich vor allem damit, dass likelihood-basierte Inferenz, die besonders dazu dient, interessierende Parameter $\boldsymbol{\theta}$, beispielsweise Korrelationen oder Steigungsparameter von Regressionen, zu schätzen, weiterhin möglich ist, wenn mindestens MAR vorliegt, also die fehlenden Werte lediglich von den beobachteten Daten abhängen[5] (Spieß 2008: 10-12).

Hinsichtlich des Datenausfallmechanismus’ ist jedoch zu beachten, dass die zu fusionierenden Datenquellen oftmals nur vermeintlich Stichproben derselben Grundgesamtheit darstellen, wodurch ein MCAR-Ausfallmechanismus häufig zu bezweifeln ist. Denn im Falle

[3]Beim Datenausfallmechanismus wird zwischen Missing Completely At Random (MCAR), Missing At Random (MAR) und Missing Not At Random (MNAR) unterschieden: Bei MCAR hängt der Datenausfall weder von den beobachteten, noch von den fehlenden Werten ab. Unter MAR sind die fehlenden Daten lediglich von den beobachteten, jedoch nicht von den fehlenden Werten abhängig. Bei MNAR hängen die fehlenden Daten sowohl von den beobachteten, als auch von den fehlenden Werten ab. Folgendes Beispiel verdeutlicht die Unterscheidung: Betrachten wir lediglich die Variablen Alter und Einkommen, wobei für das Alter vollständige Beobachtungen vorliegen, bei der Einkommensvariable hingegen 30 % der Beobachtungen fehlen. Hängt der Datenausfall beim Einkommen weder vom Alter, noch von der Höhe des Einkommens ab, liegt MCAR vor. Sind die fehlenden Einkommenswerte vom Alter, jedoch nicht vom Einkommen selbst abhängig, ist der Datenausfall MAR. MNAR liegt vor, wenn sowohl das Alter, als auch die Einkommenshöhe mit den fehlenden Einkommenswerten zusammenhängen (Little und Rubin 1991).

[4]Für eine detaillierte Erläuterung zur Ignorierbarkeit siehe z.B. Rässler (2002: 75-78).

[5]Sofern zusätzlich die Distinktheit der Parameter gewährleistet ist, sie also in keiner funktionalen Beziehung zueinander stehen (Rubin 1976; Spieß 2008: 10).

von den zu fusionierenden Daten zugrundeliegender Unit-Nonresponse[6], welche durch *unterschiedliche* Datenausfallmechanismen induziert ist, wäre MCAR, also ein komplett zufälliger Datenausfall, nicht mehr gegeben (Koller-Meinfelder 2009: 9-10; Meinfelder 2013). So könnten beispielsweise bei einer Umfrage zur Internetnutzung vermehrt ältere Probanden, bei einer Erhebung zur Nutzung von Kosmetikartikeln eher männliche Probanden die Befragungsteilnahme schlicht deshalb verweigern, weil sie zu wenig Erfahrung mit der Internetnutzung beziehungsweise mit Kosmetikartikeln haben – der Datenausfall in Abbildung 2 wäre nicht mehr zufällig. D'Orazio und Kollegen (2006) weisen darauf hin, dass die Stichproben bereits dann nicht mehr der gleichen Grundgesamtheit entstammen, wenn ihre Ziehung zu unterschiedlichen Zeitpunkten stattfindet, was ebenso einem zufallsbasierten Datenausfall zuwider käme (D'Orazio et al. 2006: 4). Mit der Annahme der bedingten Unabhängigkeit (CIA), die im folgenden Abschnitt thematisiert wird, wird jedoch automatisch MAR, also ein ignorierbarer Datenausfall angenommen, da die CIA den MAR-Datenausfall inkludiert (Koller-Meinfelder 2009: 10; Meinfelder 2013).

Mit Blick auf den konkreten Fusionsprozess wird in der Praxis meist lediglich einer der beiden in Abbildung 2 dargestellten, fehlenden Variablenblöcke durch Fusionsverfahren ergänzt (siehe z.B. Rässler 2002: 17-18). Sofern etwa nur die fehlenden **Z**-Variablen in Datenfile A mithilfe der beobachteten **Z**-Variablen aus Datenquelle B imputiert werden, stellt A den Rezipientendatenfile und B den Donorendatenfile dar, wobei die **Z**-Variablen aus Datenquelle B in den Datenfile A hineinfusioniert werden (siehe z.B. Rässler 2002: 17-18). Dieses Schema lässt sich auf das Analyseziel der vorliegenden Arbeit übertragen, indem Datenfile A EU-SILC entspricht und besonders detailliert die Einkommensangaben (**Y**) erfasst, während die Datenquelle B den HBS darstellt, der wiederum die Konsumausgaben (**Z**) umfangreich widerspiegelt. Diese Konsuminformationen **Z** sollen im Rahmen der Datenfusion in den bestehenden EU-SILC-Datensatz hineinfusioniert werden, was auch dem Ziel der vorliegenden Arbeit entspricht.

Hinsichtlich der Notation sei darauf verwiesen, dass $\mathbf{X} = (X_1, \ldots, X_p)$, $\mathbf{Y} = (Y_1, \ldots, Y_{p_{silc}})$ und $\mathbf{Z} = (Z_1, \ldots, Z_{p_{hbs}})$ jeweils Matrizen darstellen, gekennzeichnet durch fettgedruckte Großbuchstaben, die wiederum aus Spaltenvektoren, gekennzeichnet durch Großbuchstaben, bestehen. **X** ist dabei die Menge die Menge der p gemeinsamen Variablen in beiden Datenfiles, **Y** bezeichnet die Menge der p_{silc} spezifischen Einkommensvariablen aus EU-SILC, **Z** wie-

[6] Unter Unit-Nonresponse wird der *vollständige Antwortausfall* einer Beobachtungseinheit verstanden, beispielsweise in Folge einer Teilnahmeverweigerung. Bei Item-Nonresponse nehmen Beobachtungseinheiten an der Datenerhebung teil, verweigern jedoch die Beantwortung *einzelner* Fragen.

derum die der p_{hbs} spezifischen Konsumvariablen aus dem HBS.

Unter terminologischen Gesichtspunkten sei noch erwähnt, dass der Begriff der Datenfusion dem in der englischsprachigen Literatur häufig verwendeten Begriff „Statistical Matching“ vorgezogen wird. Denn der Terminus „Statistical Matching“ wird häufig verkürzt mit Nearest-Neighbour-Verfahren assoziiert (Meinfelder 2013). Da diese jedoch lediglich spezifische Fusionsverfahren darstellen, wird in der Forschung mittlerweile, auch in der englischsprachigen Literatur, der Begriff Datenfusion (beziehungsweise Data Fusion) vermehrt präferiert (Cielebak und Rässler 2014).

Abschließend wurde deutlich, dass Datenfusionen durch ein spezifisches Datenausfallmuster gekennzeichnet sind, woraus sich ein gesondertes Problem fehlender Daten ableitet, welches es mittels verschiedener Fusionsverfahren zu beheben gilt. Dies erfolgt in „eine Fusionsrichtung“ (Meinfelder 2013: 89), indem in der vorliegenden Arbeit die **Z**-Variablen in Datenfile A beziehungsweise in EU-SILC ergänzt werden sollen.

2.2 Zentrale Annahme: Conditional Independence Assumption (CIA)

Während somit nun der begriffliche und definitorische Rahmen einer Datenfusion erläutert wurde, ist nun auf die zentrale Annahme einzugehen, die einer Datenfusion zugrunde liegt: Die Conditional Independence Assumption (kurz: CIA). Diese implizite Annahme der bedingten Unabhängigkeit im Rahmen von Datenfusionen wurde zuerst von Sims (1972) in seinem Kommentar zu Okner (1972) formuliert und insbesondere von Rodgers (1984) ausführlicher diskutiert. Die CIA besagt, dass die spezifischen und nicht gemeinsam beobachteten Variablen **Y** und **Z**, gegeben den gemeinsamen Variablen **X**, voneinander unabhängig sind. Wenn davon ausgegangen wird, dass die **Z**-Variablen in einen bestehenden Datenfile aus **X**- und **Y**-Variablen hineinfusioniert werden, was in der anstehenden Simulationsstudie geschehen soll, ergibt sich für den fusionierten Datensatz die künstliche Verteilung

$$\tilde{f}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(x,y,\tilde{z}) \quad = \quad \underbrace{f_{\mathbf{X},\mathbf{Y}}(x,y)}_{\substack{\text{Verteilung}\\ \text{Rezipienten-}\\ \text{file}}} \quad \cdot \quad \underbrace{f_{\mathbf{Z}|\mathbf{X}}(z|x)}_{\substack{\text{Bedingte}\\ \text{Verteilung}\\ \text{Donorenfile}}}, \tag{1}$$

die das Produkt der ursprünglichen Verteilung des Rezipientendatensatzes und der auf **X** bedingten Verteilung des Donorendatenfiles widerspiegelt (siehe Rässler 2002: 21; Meinfelder 2013: 85). Die künstliche Varianz-Kovarianzmatrix zwischen den spezifischen, nicht

gemeinsam beobachteten **Y**- und **Z**-Variablen ergibt sich dabei durch

$$\widetilde{\text{cov}}(\mathbf{Y},\mathbf{Z}) = \text{cov}(\mathbf{Y},\mathbf{Z}) - \text{E}\left(\text{cov}(\mathbf{Y},\mathbf{Z}|\mathbf{X})\right), \tag{2}$$

wobei unter der CIA

$$\text{E}\left(\text{cov}(\mathbf{Y},\mathbf{Z}|\mathbf{X})\right) = 0 \tag{3}$$

gilt (siehe Rässler 2002: 23-24; Meinfelder 2013: 86). Dementsprechend entspricht die fusionierte Varianz-Kovarianzmatrix $\widetilde{\text{cov}}(\mathbf{Y},\mathbf{Z})$ der wahren Kovarianz $\text{cov}(\mathbf{Y},\mathbf{Z})$, sofern sich die auf **X** bedingten Kovarianzen beziehungsweise, im standardisierten Setting, die auf **X** bedingten Korrelationen zwischen **Y** und **Z** im Mittel aufheben, also 0 sind (Rässler 2002: 24).

Im Kontext von Datenfusionen muss die CIA deshalb unterstellt werden, weil herkömmliche Fusionsalgorithmen stets bedingte Unabhängigkeit im fusionierten Datenfile herstellen[7]. Auch die in dieser Arbeit thematisierten Verfahren, Random Hot-Deck und Predictive Mean Matching, erfordern die Annahme bedingter Unabhängigkeit (D'Orazio et al. 2006: 37-39, 47-49). Doch wie in (2) und (3) ersichtlich ist, wird die Korrelation zwischen den nicht gemeinsam beobachteten Variablen **Y** und **Z** im fusionierten Datenfile nur dann korrekt wiedergegeben, wenn **Y** und **Z**, gegeben **X**, tatsächlich im Mittel unkorreliert sind. Andernfalls sind kohärente Schlussfolgerungen aus der im Rahmen der Datenfusion künstlich hergestellten Verteilung in (1) unbrauchbar und fehlerhaft (Kiesl und Rässler 2005, 2006).

Daher wäre es unabdingbar, zu überprüfen, beispielsweise im Rahmen eines statistischen Tests, ob die Annahme der bedingten Unabhängigkeit für die betreffende Grundgesamtheit, über die eine Aussage getroffen werden soll, haltbar ist. Doch da **Y** und **Z** in den zu fusionierenden Stichproben nie gemeinsam beobachtet wurden, was als Identifikationsproblem der Datenfusion bezeichnet wird, ist eine solche Überprüfung nicht möglich – die CIA ist also nicht testbar (Kiesl und Rässler 2005). Dies stellt für den Fusionsprozess ein schwerwiegendes Problem dar, dessen Dilemma Kiesl und Rässler (2005) treffend zusammenfassen:

> *„Herkömmliche Fusionsalgorithmen stellen immer bedingte Unabhängigkeit her, obwohl diese Annahme im besten Fall nicht überprüfbar, im schlechtesten Fall unhaltbar ist“*
> (Kiesl und Rässler 2005: 24).

[7] Für eine Illustration siehe Kiesl und Rässler (2005).

Jedoch ist es zumindest möglich, eine Eingrenzung der Korrelationen zwischen den unbeobachteten Variablenblöcken $\mathbf{Y}$ und $\mathbf{Z}$ über Fréchet-Hoeffding-Grenzen vorzunehmen (siehe dazu Kiesl und Rässler 2005, 2006).

Für die in dieser Arbeit thematisierte Datenfusion von EU-SILC und HBS, die dazu dienen soll, die gemeinsame Verteilung von Einkommen und Konsum zu betrachten, erscheint die CIA ebenso eine unrealistische Annahme zu sein. Folgende vereinfachte, univariate Situation illustriert dies: Nehmen wir an, EU-SILC und HBS bestehe lediglich aus jeweils zwei erhobenen Merkmalen beziehungsweise Variablen. In EU-SILC wurde das Alter der Referenzperson ($\mathbf{X} = X$) sowie das Haushaltsnettoeinkommen ($\mathbf{Y} = Y$) erhoben, im HBS wiederum ebenfalls das Alter der Referenzperson ($\mathbf{X} = X$) sowie die Gesamtkonsumausgaben des Haushalts ($\mathbf{Z} = Z$). Sofern durch eine Datenfusion ein wie in Abbildung 1 dargestellter, fusionierter Datenfile aus $\mathbf{X}$, $\mathbf{Y}$ und $\widetilde{\mathbf{Z}}$ entsteht, wäre die auf das Alter (X) bedingte Korrelation zwischen dem Haushaltsnettoeinkommen (Y) und der künstlichen Verteilung der Gesamtkonsumausgaben ($\widetilde{Z}$) gleich 0 (Kiesl und Rässler 2005). Doch ist es höchst unwahrscheinlich, dass für Individuen mit gleichem Alter kein Zusammenhang zwischen Einkommen und Konsum besteht. Genau dies würde jedoch die CIA implizit annehmen. Darüber hinaus ist es hinsichtlich des Datenausfallmechanismus' unrealistisch, dass die Stichproben EU-SILC und HBS Zufallsstichproben der exakt selben Grundgesamtheit darstellen und demnach ein ignorierbarer Datenausfall vorläge. Wie bereits erwähnt, müsste dafür beispielsweise die Stichprobenziehung von EU-SILC und HBS zum gleichen Zeitpunkt erfolgen (D'Orazio et al. 2006: 4). Jedoch ist dies schon allein aufgrund der Erhebungsstrukturen beider Stichproben unplausibel und somit äußerst unrealistisch (siehe z.B. Statistisches Bundesamt 2016a, 2016b). Ein ignorierbarer Datenausfall wäre also deutlich in Frage gestellt. Dennoch wird mit der CIA die Ignorierbarkeit implizit angenommen, da sie die MAR-Annahme beinhaltet. Somit wird deutlich, dass bei der Datenfusion von EU-SILC und HBS stichhaltige Indizien vorliegen, dass die Annahme der bedingten Unabhängigkeit nicht, oder allenfalls nur eingeschränkt, haltbar ist, was fehlerhafte Schlussfolgerungen induzieren könnte. Dieser Indizienlage und der damit verbundenen Problematiken sollte sich die amtliche Statistik hinsichtlich ihres Datenfusionsvorhabens stets bewusst sein.

Um die mit der CIA auftretenden Probleme zu berücksichtigen, wird in der bisherigen Datenfusionsforschung häufig das Hinzuziehen von Hilfsinformationen empfohlen (siehe z.B. Singh et al. 1993; Fosdick et al. 2015). Diese können einerseits mikrobasiert aus einem dritten Datenfile C, in dem alle drei Variablenblöcke $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ oder zumindest die spezifischen

Variablen $(\mathbf{Y}, \mathbf{Z})$ gemeinsam beobachtet wurden, andererseits makrobasiert entlang von früher publizierten Parametergrößen bezüglich den unbekannten Verteilungen $(\mathbf{Y}, \mathbf{Z}|\mathbf{X})$ oder $(\mathbf{Y}, \mathbf{Z})$ gewonnen werden (Singh et al. 1993). Die verfügbaren Hilfsinformationen sind dann in den Fusionsprozess miteinzubeziehen (siehe D'Orazio et al. 2006: 65-95).

Bezüglich deren Einbindung in den Fusionsprozess stellt etwa der *glue*-Ansatz von Fosdick et al. (2015) eine moderne und vielversprechende Alternative dar. Fosdick und Kollegen (2015) versuchen, zumindest für einzelne **Y**- und **Z**-Variablen gemeinsame Beobachtungen aus weiteren Datenquellen miteinzubeziehen. Denn häufig besteht das Problem, dass entsprechende Hilfsinformationen, sofern überhaupt vorhanden, nicht für alle **Y**- und **Z**-Variablen zur Verfügung stehen. Sofern jedoch solche Informationen zumindest teilweise verfügbar sind, ist die Idee von Fosdick et al. (2015), vereinfacht gesprochen, dass durch das „Untereinanderbinden" der zu fusionierenden Datenquellen A und B und der Erweiterung um die zur Verfügung stehenden Hilfsdaten das spezifische Datenausfallmuster aus Abbildung 2 ein wenig diffuser wird, was die CIA etwas auflöst, sodass sie dann weniger stark ins Gewicht fällt. Die Fusionierung selbst erfolgt dann mit bayesianischen Modellen und Multipler Imputation (Fosdick et al. 2015).

Derartige Alternativen, die die Problematiken mit der CIA eindämmen, könnten auch für die amtliche Statistik von Interesse sein – eine Implementierung dessen ist jedenfalls, sofern möglich, nahezulegen. Wie bereits erwähnt, wird jedoch für die in vorliegender Arbeit verwendeten Ansätze, Random Hot-Deck und Predictive Mean Matching, die CIA unterstellt. Die damit verbundenen Problematiken wurden diskutiert und sind bei der späteren Ergebnisinterpretation der anstehenden Simulationsstudie zu berücksichtigen.

2.3 Vier Validitätsstufen einer Datenfusion

Für die Bewertung eines solchen Datenfusionsergebnisses diskutiert die Fusionsforschung vier Validitätsstufen, die in diesem Abschnitt kurz vorgestellt werden. Die Kategorisierung und Initiierung der vier Validitätsstufen geht auf Rässler (2002: 29-32) zurück. Sofern entlang der Abbildung 2 und dem dieser Arbeit zugrundeliegenden Untersuchungsauftrag davon ausgegangen wird, dass, unter allgemeiner Notation, B die Donorenstichprobe darstellt, deren p_B spezifische **Z**-Variablen in den bestehenden Rezipientendatenfile A hineinfusioniert werden sollen, stellen sich die vier Validitätsstufen wie folgt dar (Rässler 2002: 29-32):

1. Erhalt der individuellen Werte der spezifischen **Z**-Variablen des Donorendatenfiles. Somit gilt für jeden der p_B fusionierten Spaltenvektoren: $\tilde{z}_i, \ldots, \tilde{z}_{n_A} = z_i, \ldots, z_{n_A}$ für $i = 1, \ldots, n_A$, wobei n_A die Stichprobengröße des Rezipientendatenfiles A darstellt.

2. Erhalt der gemeinsamen Verteilung aller Variablenmengen $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, wodurch die künstliche Verteilung im Fusionsdatenfile der wahren Verteilung entspricht: $\tilde{f}_{\mathbf{X},\mathbf{Y},\mathbf{Z}} = f_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$.

3. Erhalt der Korrelationsstrukturen, wodurch die wahren Korrelationen zwischen $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ den Korrelationen im fusionierten Datenfile entsprechen: $\widetilde{\mathrm{cov}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \mathrm{cov}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

4. Erhalt der ursprünglichen Verteilungen des Donorendatenfiles, wodurch der Fusionsdatenfile zumindest die marginale Verteilung von **Z** sowie die gemeinsame Verteilung $(\mathbf{X}, \mathbf{Z})$ aus dem Donorendatenfile widerspiegelt: $\tilde{f}_{\mathbf{Z}} = f_{\mathbf{Z}}$ und $\tilde{f}_{\mathbf{X},\mathbf{Z}} = f_{\mathbf{X},\mathbf{Z}}$.

Anzumerken ist zunächst, dass herkömmliche Fusionsalgorithmen die Validitätsstufen 2 und 3 nur unter der CIA erreichen können, also wenn **Y** und **Z**, gegeben **X**, im Mittel unkorreliert sind. Zudem kann lediglich die Überprüfung der vierten Validitätsstufe mit empirischen Daten erfolgen, da im Donorendatenfile die marginale Verteilung zwischen **X** und **Z** bekannt ist. Zur Evaluierung aller weiteren Stufen bedarf es einer Simulationsstudie oder zusätzlicher Hilfsinformationen. Daher ist es bei reell durchgeführten Datenfusionen ratsam, mindestens zu überprüfen, ob sich die bereits im Donorendatenfile beobachtete Verteilung von **X** und **Z** im Fusionsdatenfile adäquat widerspiegelt, was etwa, je nach Skalenniveau, unter Zuhilfenahme von χ^2-Tests oder t-Tests verglichen werden könnte. Jedoch ist bei Verwendung dieser Tests das eigentlich wünschenswerte Ergebnis, dass keine signifikanten Verteilungsunterschiede zwischen Donoren- und Fusionsdatenfile auftreten, in der Nullhypothese verortet, womit das Problem einhergeht, dass sich aus Perspektive der Testtheorie mit einer Nichtannahme der H_0 eine wenig substanzielle Aussagekraft verbindet (Rässler 2002: 31-32; Kiesl und Rässler 2005, 2006).

Ferner ist anzumerken, dass, obwohl die erste Stufe zwar die formal höchste Validitätsebene darstellt, sie kein geeignetes Bewertungskriterium für eine Datenfusion ist. Denn einerseits wäre bei metrischen Variablen die Punktwahrscheinlichkeit, dass der imputierte Wert tatsächlich dem korrekten, aber unbekannten Wert entspricht, gleich 0. Besonders problematisch ist jedoch andererseits, dass durch das Ziel der Reproduktion der exakten, individuellen Werte der **Z**-Variablen nicht sichergestellt ist, dass deren tatsächliche Verteilung erhalten bliebe[8]. Bezüglich der zweiten Stufe ist wiederum zu konstatieren, dass der Erhalt der ge-

[8] Für eine Illustration siehe Kiesl und Rässler (2005)

meinsamen Verteilung für empirische Daten nur dann im Ansatz realistisch ist, wenn äußerst hohe Korrelationen zwischen den gemeinsamen **X**-Variablen und den jeweils spezifischen **Y**- und **Z**-Variablen vorlägen. Da solche starken Zusammenhänge jedoch sehr unwahrscheinlich sind, gilt der Erhalt der gemeinsamen Verteilung von **Y** und **Z** im Zuge einer Datenfusion als unrealistisch, wodurch eine Evaluation dessen als wenig vielversprechend angesehen wird (Kiesl und Rässler 2005, 2006; Cielebak und Rässler 2014).

Daher liegt für die Bewertung eines Datenfusionsergebnisses der besondere Fokus auf den Validitätsstufen 3 und 4. Vor allem die dritte Stufe, also der Erhalt der Korrelationsstrukturen, ist von elementarem Interesse, da Datenfusionen oftmals genau zu diesem Zwecke, zur Erfassung und inferenzstatistischen Analyse unbeobachteter Korrelationen zwischen **Y** und **Z**, durchgeführt werden (Kiesl und Rässler 2005). Die vierte Stufe, der Erhalt der bereits im Donorendatenfile beobachteten, marginalen Verteilung stellt eine Art Mindestanforderung an eine Datenfusion dar und sollte demnach ebenso evaluiert werden. Auch bezüglich der vierten Validitätsebene ist für inferenzstatistische Analysen von besonderem Interesse, inwiefern sich die im Donorendatenfile bereits beobachtete Korrelationsstruktur zwischen **X** und **Z** im Fusionsdatensatz widerspiegelt. Dementsprechend wird auch das Fusionsergebnis der anstehenden Simulationsstudie entlang der Validitätsstufen 3 und 4 evaluiert, wobei besondere Bedeutung und Aufmerksamkeit der dritten Stufe zukommt (Kiesl und Rässler 2006).

Abschließend wurde in diesem einführenden Überblickskapitel deutlich, dass Datenfusionen als Problem fehlender Daten mit einem spezifischen Datenausfallmuster anzusehen sind und ihnen mit der CIA eine problematische weil häufig verletzte und unrealistische Annahme zugrunde liegt. Ein Datenfusionsergebnis kann anhand von vier Bewertungskriterien, den eben diskutierten Validitätsstufen, evaluiert werden, wobei der Fokus auf den Stufen 3 und 4 liegen sollte.

3 Relevante Fusionsalgorithmen

Doch wie werden Datenfusionen nun in der Praxis durchgeführt? Um sich den konkreten Fusionsalgorithmen anzunähern, liefert dieses Kapitel zunächst einen kurzen Abriss zu klassischen, in der bisherigen Forschung sowie in den wissenschaftlichen Beiträgen von Eurostat diskutierten Fusionsansätzen und erläutert, welcher dieser Ansätze den hier thematisierten Verfahren, Random Hot-Deck und Predictive Mean Matching, zugrunde liegt. Neben dieser

Einordnung ist auch auf die relevante Unterscheidung zwischen Unconstrained Matching und Constrained Matching einzugehen. Anschließend, und weitaus detaillierter, werden die beiden in dieser Arbeit betrachteten Algorithmen zur Datenfusion von EU-SILC und HBS, Random Hot-Deck und Predictive Mean Matching, beschrieben sowie deren theoretische Implikationen diskutiert. Die daraus resultierende Quintessenz ist letztlich in eine Arbeitshypothese zu transferieren.

3.1 Überblick gängiger Fusionsansätze

Traditionell werden Datenfusionen mit Nearest-Neighbour-Methoden durchgeführt, die die Daten über Beobachtungen fusionieren, die sich in ihren gemeinsamen **X**-Variablen so ähnlich wie möglich sind (siehe z.B. Van der Putten et al. 2002; Kiesl und Rässler 2005). Neben dieser Methodenfamilie existieren in der Datenfusionsforschung jedoch auch rein verteilungsbasierte Ansätze, also parametrische Verfahren, die auf Regressionen von **Z** auf **X** im Donorendatenfile basieren und mithilfe der daraus resultierenden Regressionsparameter die fehlenden **Z**-Werte im Rezipientendatenfile imputiert werden (siehe z.B. D'Orazio et al. 2006: 14-25, 31-34; Gilula et al. 2006). Bei Nearest-Neighbour-Methoden wie bei parametrischen Verfahren wird in der Regel angestrebt, einen integrierten, fusionierten Mikrodatenfile aus $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ zu erhalten, was auch die amtliche Statistik verfolgt und dem Ziel dieser Arbeit entspricht.

Es können aber auch rein makrobasierte Ansätze zum Einsatz kommen. Diese versuchen, auf Basis der zu fusionierenden Datensätze lediglich die gemeinsame Verteilung $f_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(x,y,z)$, oder zumindest einzelne Eigenschaften davon, zu schätzen, anstatt einen vollständigen Mikrodatenfile zu generieren (D'Orazio et al. 2006: 2-3). Jedoch ist die anschließende Datenanalyse mit einem fusionierten Mikrodatenfile einfacher und plausibler, weshalb mikrobasierte Ansätze deutlich häufiger verwendet werden als Makromethoden (D'Orazio et al. 2006: 2-3). Während allen genannten Fusionsansätzen die Annahme der bedingten Unabhängigkeit (CIA) zugrunde liegt, dessen Problematiken in Kapitel 2.2 diskutiert wurden, existieren in der neueren Datenfusionsforschung auch Methoden, die die CIA, insbesondere mittels Hilfsinformationen, zu umgehen versuchen, wie etwa der bereits erwähnte *glue*-Ansatz von Fosdick et al. (2015).

Die beiden in dieser Arbeit relevanten Fusionsalgorithmen, Random Hot-Deck und Predictive Mean Matching, lassen sich hingegen den Nearest-Neighbour-Methoden zuordnen. Ran-

dom Hot-Deck stellt dabei einen nicht-parametrischen Fusionsansatz dar (D'Orazio et al. 2006: 37-39). Predictive Mean Matching verfolgt jedoch die Distanzmessung mittels auf Regressionen basierenden „intermediate value[s]" (D'Orazio et al. 2006: 47), also unter parametrischen Grundlagen. Da jedoch im Vergleich zu den parametrischen Verfahren die fehlenden **Z**-Werte nicht regressionsbasiert, sondern mit den jeweiligen Ausprägungen der über die intermediate values berechneten, maximal ähnlichsten Donorenbeobachtung in nicht-parametrischer Weise imputiert werden, kann Predictive Mean Matching zu einer weiteren Klasse von (mikrobasierten) Fusionsansätzen zugeordnet werden: Den semi-parametrischen Verfahren, die eine Mischmethode zwischen parametrischen und nicht-parametrischen Ansätzen darstellen (D'Orazio et al. 2006: 47; Meinfelder 2013).

Auch in bisherigen Working Papers von Eurostat zur Datenfusion von EU-SILC und HBS werden derartige „Mixed Methods" (siehe z.B. Leulescu und Agafitei 2013: 18; Webber und Tonkin 2013: 20; Serafino und Tonkin 2017: 15), neben nicht-parametrischen und parametrischen Ansätzen, als potentielle Fusionsverfahren diskutiert. Dass Eurostat mit dem Random Hot-Deck dennoch ein nicht-parametrisches Verfahren angewendet hat, ist einerseits etwas überraschend, da die Studie von Webber und Tonkin (2013) leichte Performancevorteile den semi-parametrischen Ansätzen zuschreibt (Webber und Tonkin 2013). Andererseits könnte das Implementieren von Random Hot-Deck auch gerade auf die *leichten* Performancevorteile zurückgeführt werden, da die entsprechenden Ergebnisse von Webber und Tonkin (2013) nur eine geringe Überlegenheit von semi-parametrischen Verfahren gegenüber nicht-parametrischen Ansätzen implizieren (Webber und Tonkin 2013). Die anstehende Simulationsstudie kann daher auch die diesbezügliche, in der Datenfusionsforschung von Eurostat verankerte Diskussion befruchten, indem die Fusionsperformance einer nicht-parametrischen Methode (Random Hot-Deck) sowie eines semi-parametrischen Verfahrens (Predictive Mean Matching) nochmals kompetitiv rezipiert wird.

Da beides Nearest-Neighbour-Methoden darstellen, ist noch anzumerken, dass diesbezüglich in der Forschung besonders zwischen Unconstrained Matching und Constrained Matching unterschieden wird (Rodgers 1984; Rubin 1986). Beim Unconstrained Matching kann eine Donorenbeobachtung beliebig oft mit einem Rezipienten verknüpft werden, während beim Constrained Matching eine Donorenbeobachtung lediglich für eine beschränkte Anzahl an Rezipienten, im Extremfall lediglich für eine Empfängerbeobachtung, verwendet werden darf (Kiesl und Rässler 2005). Beim Constrained Matching können einerseits die Randverteilungen des Donorendatenfiles besser reproduziert werden, was der vierten Validi-

tätsstufe zuträglich ist (Rässler 2002: 57-60; Kiesl und Rässler 2005). Andererseits führt dies dazu, dass der Fusionsdatensatz um zusätzliche Zeilen vergrößert wird und er somit einen höheren Stichprobenumfang als der zu ergänzende Rezipientendatenfile aufweist (Rässler 2002: 57-60). Beim Unconstrained Matching ist wiederum sichergestellt, dass die Rezipientenbeobachtungen tatsächlich mit ihrer ähnlichsten Donorenbeobachtung verbunden werden (Rässler 2002: 53). Auch Random Hot-Deck und Predictive Mean Matching verfolgen ein solches Unconstrained Matching, womit also die unbegrenzte Imputationsverwendung einer Donorenbeobachtung zugelassen wird.

Bevor die beiden Verfahren ausführlich beschrieben werden, bleibt festzuhalten, dass die bisherige Forschung, neben den Nearest-Neighbour-Methoden, eine Vielzahl weiterer Fusionsansätze diskutiert. Zu Beginn dieses Abschnitts erfolgte lediglich eine kurze Zusammenschau der verschiedenen, in der wissenschaftlichen Literatur relevanten Verfahren. Für einen ausführlichen Überblick klassischer Fusionsalgorithmen sei insbesondere auf Rässler (2002) und D'Orazio et al. (2006) verwiesen, wobei in Rässler (2002) zusätzlich alternative, bayesianische Fusionsalgorithmen diskutiert und analysiert, in D'Orazio et al. (2006) wiederum auch makrobasierte Ansätze thematisiert werden. Abschließend wurde deutlich, dass Random Hot-Deck und Predictive Mean Matching Nearest-Neighbour-Verfahren darstellen, die eine unbeschränkte Donorenverwendung (Unconstrained Matching) zulassen. Während Random Hot-Deck den nicht-parametrischen Ansätzen zugeordnet werden kann, spiegelt Predictive Mean Matching ein semi-parametrisches Verfahren wider.

3.2 Random Hot-Deck von Eurostat

Nun wird in diesem Abschnitt zunächst das von Eurostat angewendete Fusionsverfahren ausführlich erläutert. Wie sich der konkrete Datenfusionsalgorithmus von Eurostat gestaltet, kann dem R-Code von Eurostat, der dem Statistischen Bundesamt vorliegt, entnommen werden. Wie bereits deutlich wurde, entspricht das betreffende Fusionsverfahren einem erweiterten Random Hot-Deck, wie es in Lamarche (2017) beschrieben ist.

Allgemein bezeichnet ein Random Hot-Deck im Kontext von Datenfusionen zunächst das zufällige Zuweisen von Beobachtungen des Donorendatensatzes zu Beobachtungen des Rezipientendatenfiles. Die nicht vorhandenen $\mathbf{Z}$-Variablenwerte für jede Beobachtung der Rezipientenstichprobe werden dann durch die entsprechenden Variablenausprägungen der ihr zugewiesenen Donorenbeobachtung imputiert. Da dem Fusionsprozess keine Verteilungs-

grundlage obliegt, ist dies, wie bereits deutlich wurde, eine nicht-parametrische Vorgehensweise. Die Zufallszuweisung zwischen Rezipientenbeobachtung und Donorenbeobachtung erfolgt jedoch in der Regel innerhalb von homogenen Teilgruppen, also beispielsweise nur innerhalb des gleichen Geschlechts, sodass etwa einer männlichen beziehungsweise weiblichen Rezipientenbeobachtung auch nur männliche beziehungsweise weibliche Donorenbeobachtungen zufällig zugewiesen werden können (D'Orazio et al. 2006: 37).

Die Fusionsmethode von Eurostat basiert auf einer solchen zufälligen Zuweisung innerhalb homogener Teilgruppen, wobei dieser Zufallszuweisung einige Schritte vorausgehen. Das Eurostat-Verfahren stellt sich bezüglich der Datenfusion von EU-SILC und HBS für jeden zu fusionierenden Spaltenvektor Z_r (mit $r = 1, \ldots, p_{hbs}$) des Variablenblocks $\mathbf{Z}$ folgendermaßen dar: Zunächst werden alle gemeinsamen Variablen $X_1, \ldots, X_p$, die metrisches Skalenniveau aufweisen, kategorisiert. So wird etwa das Alter in grobe Altersklassen oder das Einkommen in Einkommensquintile transferiert. Alle $\mathbf{X}$-Variablen sind damit höchstens ordinalskaliert. Anschließend erfolgt mithilfe einer auf OLS-Regression basierenden Backward Selection eine Auswahl an für die Erklärung der zu fusionierenden Variable Z_r relevanten $\mathbf{X}$-Variablen. Die entsprechenden Regressionen werden dabei auf Grundlage des Donorendatenfiles, also des HBS berechnet. Anhand der a ausgewählten Variablen $X_1, \ldots, X_a$ erfolgt dann die Bildung einer Schichtungsvariable in beiden Datenfiles, die im Prinzip die kategorialen Ausprägungen „aneinanderhängt". Ein Beispiel: Angenommen, die $a = 2$ ausgewählten Variablen wären Geschlecht (X_1) und die Altersklasse (X_2), wobei für eine konkrete Beobachtung das Geschlecht den Wert 1, die Altersklasse den Wert 3 annimmt – die entsprechende Ausprägung auf der Schichtungsvariable wäre dann „13" (Lamarche 2017).

Um sicherzustellen, dass für jede der l Schichtungsausprägungen genügend Donoren in HBS ($s_{l,hbs}$) für die Rezipienten in EU-SILC ($s_{l,silc}$) zur Verfügung stehen, setzt Eurostat einen Schwellwert, sodass das Verhältnis zwischen Rezipienten und vorhandenen Donoren je Schichtungsausprägung in etwa dem Verhältnis der Stichprobenumfänge von EU-SILC (s_{silc}) und HBS (s_{hbs}) entspricht:

$$\frac{s_{l,silc}}{s_{l,hbs}} \quad \geq \quad c \cdot \frac{s_{silc}}{s_{hbs}} \,,$$

wobei die Konstante c von Eurostat als Faustregel auf $c = 3$ gesetzt wird (siehe Lamarche 2017: 5). Sofern der entsprechende Schwellwert von mindestens 90 % der Rezipienten-

beobachtungen nicht überschritten wird[9], was das Abbruchkriterium darstellt, werden die durch die Backward Selection ausgewählten Variablen $X_1, \ldots, X_a$ beibehalten und die oben erwähnte Schichtung wird durchgeführt. Andernfalls wird die Variablenauswahl solange iteriert, bis das Abbruchkriterium erreicht ist, wobei sich die maximale Teilmenge der durch die Backward Selection auszuwählenden **X**-Variablen je Iterationsschritt um 1 reduziert. Sobald dann die finalen Variablen $X_1, \ldots, X_a$ feststehen, werden innerhalb der gleichen Schichtungsausprägungen, die die homogenen Teilgruppen darstellen, zufällige Donorenbeobachtungen den Rezipientenbeobachtungen zugewiesen. So erfolgt etwa entlang obigem Beispiel eine Zufallszuweisung zwischen Rezipient und Donor lediglich unter den Beobachtungen, die für Geschlecht den Wert 1 und für die Altersklasse den Wert 3 aufweisen. Nach der Zufallszuweisung werden die Z_r-Werte der Rezipienten in EU-SILC durch die ihrer jeweils zugewiesenen Donoren aus HBS imputiert (Lamarche 2017).

Sollten Rezipienten überbleiben, für die entlang ihrer Werte von $X_1, \ldots, X_a$ noch kein merkmalsgleicher Donor zur Verfügung steht, wird die Variablenauswahl solange weiter iteriert, bis entlang der gebildeten Schichtungsausprägungen alle verbleibenden Rezipienten mit mindestens einem Donor übereinstimmen, wobei die Variablenauswahl nun ohne Berücksichtigung der obigen Schwelle erfolgt. Die maximale Variablenanzahl wird jedoch auch hier in jedem Iterationsschritt um 1 reduziert. Sobald für alle übergebliebenen Rezipienten mindestens ein merkmalsgleicher Donor verfügbar ist, wird die iterative Variablenauswahl abgebrochen, die oben beschriebene Schichtung anhand der ausgewählten Variablen durchgeführt und den Rezipienten merkmalsgleiche Donoren zufällig zugewiesen, wobei für jede Rezipientenbeobachtung der Z_r-Wert der zugewiesenen Donorenbeobachtung imputiert wird. Sodann liegt ein fusionierter Datenfile, bestehend aus den Variablenblöcken $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{Z}})$ mit $\widetilde{\mathbf{Z}} = \widetilde{Z}_r$, entlang der Abbildung 1 vor. Die Fusionierung der übrigen spezifischen **Z**-Variablen erfolgt analog (Lamarche 2017).

3.3 Predictive Mean Matching (PMM)

Nun soll im Rahmen der vorliegenden Arbeit untersucht werden, ob das eben beschriebene Eurostat-Verfahren zur Datenfusion von EU-SILC und HBS durch Predictive Mean Matching (kurz: PMM) optimiert werden kann, weshalb dieser Abschnitt das Fusionskonzept

[9] Im Falle von gleich großen Stichprobenumfängen von EU-SILC und HBS bedeutet die Schwelle übersetzt, dass auf einen Donoren höchstens drei Rezipienten fallen dürfen – für maximal 10 % der Rezipienten wird eine Abweichung dessen, also dass ein Donor auf mehr als drei Rezipienten fällt, toleriert.

mittels PMM erläutert. PMM geht ursprünglich auf Rubin (1986) und Little (1988) zurück und wurde für die Imputation stetiger Variablen entwickelt. Die Grundidee ist, dass für jeden fehlenden Wert einer Variablen ihr „predictive mean“ (Little 1988: 291), der sich mithilfe einer OLS-Regression berechnet, mit den predictive means der beobachteten Werte verglichen wird, wobei die Beobachtung mit dem ähnlichsten predictive mean als Donor fungiert, dessen realer Wert sodann den fehlenden Wert ersetzt (Rubin 1986; Little 1988). Die predictive means stellen dabei die bereits erwähnten und bei semi-parametrischen Verfahren relevanten intermediate values dar. Mittlerweile ist PMM ein häufig verwendetes Verfahren zur Imputation fehlender Werte in Datensätzen – im `R`-Package `mice` von Van Buuren (2018) und der gleichnamigen Funktion stellt es etwa die Default-Methode für die Ergänzung stetiger Variablen dar. Die Anwendung von PMM im Kontext von Datenfusionen wurde hingegen besonders in Koller-Meinfelder (2009) diskutiert.

Das PMM-Verfahren stellt sich bezüglich der Datenfusion von EU-SILC und HBS für jede Variable Z_r (mit $r = 1, \ldots, p_{hbs}$) wie folgt dar: Zunächst werden, wie beim Random Hot-Deck von Eurostat, relevante $\mathbf{X}$-Variablen über eine auf OLS-Regression basierenden Backward Selection der zu fusionierenden Variable Z_r auf die gemeinsamen Variablen $\mathbf{X}$ ausgewählt, wobei die metrischen $\mathbf{X}$-Variablen nicht kategorisiert werden müssen und somit jeder Spaltenvektor $X_1, \ldots, X_p$ sein ursprüngliches Skalenniveau beibehält (Meinfelder und Schnapp 2015). Anhand der Regressionsgleichung, die die a ausgewählten Variablen $X_1, \ldots, X_a$ beinhaltet, wird dann der predictive mean für jede Beobachtung in EU-SILC und HBS errechnet. Die Suche nach entsprechenden Donoren erfolgt nun unter Verwendung der von Little (1988) vorgeschlagenen Mahalanobis-Distanzfunktion:

$$D_{i,j} = (\hat{z}_i - \hat{z}_j)^T S^{-1}_{Z_r|\mathbf{X}} (\hat{z}_i - \hat{z}_j) \tag{4}$$

mit $i = 1, \ldots, n_{silc}$ und $j = 1, \ldots, n_{hbs}$, wobei $\hat{z}_i$ dem predictive mean der i-ten Beobachtung aus EU-SILC und $\hat{z}_j$ dem predictive mean der j-ten Beobachtung aus HBS entspricht (siehe Meinfelder 2013: 89). $S^{-1}_{Z_r|\mathbf{X}}$ stellt die inverse Varianz-Kovarianzmatrix der Residuen der Regression von Z_r auf $X_1, \ldots, X_a$ dar, anhand derer die Distanz in (4) gewichtet wird. Dadurch kommt jenen a ausgewählten Variablen $X_1, \ldots, X_a$, die eine gute Erklärungskraft bezüglich der zu fusionierenden Z_r-Variable aufweisen, ein stärkerer Einfluss auf die Gesamtdistanz zu, als jenen gemeinsamen Variablen, die Z_r weniger gut erklären können und die somit für die Berechnung der Gesamtdistanz geringer gewichtet werden (Koller-Meinfelder 2009: 33-34; Meinfelder 2013).

Nach Berechnung der Distanzen $D_{i,j}$ kann für jede Rezipientenbeobachtung in EU-SILC ihr maximal ähnlicher Donor ermittelt werden, welcher sich durch die geringste Gesamtdistanz $D_{i,j}$ definiert. Sodann wird für jeden Rezipienten der entsprechende, real beobachtete Z_r-Wert des maximal ähnlichen Donors aus HBS imputiert. Sofern ein Rezipient zu mehreren Donoren die minimalste Distanz aufweist, erfolgt eine zufällige Auswahl dieser „gleichähnlichen" Donoren. Auch hier liegt sodann ein fusionierter Datenfile aus $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{Z}})$ mit $\widetilde{\mathbf{Z}} = \widetilde{Z}_r$ vor, wobei die Fusionierung der übrigen $\mathbf{Z}$-Variablen analog erfolgt (Koller-Meinfelder 2009: 93-96).

3.4 Diskussion und theoretische Implikationen

Beide Verfahren, Random Hot-Deck wie PMM, versuchen, eine Datenfusion über maximal ähnliche Beobachtungen vorzunehmen und erlauben eine uneingeschränkte Donorenverwendung (Unconstrained Matching). Das von Eurostat angewandte Random Hot-Deck stellt dabei ein nicht-parametrisches Verfahren dar. PMM ist hingegen den semi-parametrischen Ansätzen zuzuordnen, da die Distanzmessung bei PMM über regressions- und damit verteilungsbasierte predictive means erfolgt. Beide Datenfusionsalgorithmen, Random Hot-Deck und PMM, unterscheiden sich dabei in zwei wesentlichen Punkten, die in diesem Abschnitt diskutiert werden und deren theoretische Implikationen in eine Arbeitshypothese zu überführen sind.

Erstens geht beim Random Hot-Deck von Eurostat, sobald ein finales Set an $X_1, \ldots, X_a$ Fusionsvariablen feststeht, jede dieser Variablen mit gleichem Gewicht in die Distanzberechnung beziehungsweise in die Suche nach Übereinstimmungen mit ein, ungeachtet ihres Erklärungsbeitrags auf die zu fusionierenden p_{hbs} Konsumvariablen $\mathbf{Z}$. Wenn etwa beim Eurostat-Verfahren eine durch die Backward Selection ausgewählte Fusionsvariable im Vergleich zu den anderen, ebenfalls ausgewählten gemeinsamen Variablen die Konsuminformationen $\mathbf{Z}$ schlechter erklären kann, so spielt dies bei der Distanzberechnung beziehungsweise bei der Suche nach Matches keine Rolle. Dementsprechend lässt das Random Hot-Deck außer Acht, dass trotz der ausgewählten Variablen $X_1, \ldots, X_a$ ein Optimierungsproblem dahingehend besteht, welche dieser Variablen für die Distanzberechnung beziehungsweise, bei Eurostat, für die Suche nach Übereinstimmungen besondere Relevanz aufweisen. PMM steuert die Distanz hingegen über $\hat{z}_i$ und $\hat{z}_j$, also entlang der predictive means. Dadurch wird das auftretende Optimierungsproblem durch Gewichtung anhand der Inversen der Varianz-Kovarianzmatrix der Residuen einer Regression von Z_r auf $X_1, \ldots, X_a$ berücksichtigt. Je bes-

ser beziehungsweise schlechter dabei die ausgewählten Fusionsvariablen $X_1, \ldots, X_a$ die Konsumangaben Z_r erklären können, desto stärker beziehungsweise schwächer ist ihr Einfluss in der Distanzberechnung (Koller-Meinfelder 2009: 33-34; Meinfelder 2013). Besonders aufgrund dieser Abstufung der gemeinsamen, ausgewählten Variablen $X_1, \ldots, X_a$, die das Random Hot-Deck von Eurostat nicht vornimmt, sollte PMM eine präzisere Identifikation maximal ähnlicher Beobachtungen gewährleisten und ein besseres Fusionsergebnis produzieren.

Das präzisere Finden eines geeigneten Donors wird von PMM noch aus einem weiteren Aspekt erwartet: Denn *zweitens* muss bei PMM keine Kategorisierung der metrischen **X**-Variablen vorgenommen werden, womit ihr ursprüngliches Skalenniveau erhalten bleibt. Dies erspart einerseits Arbeit und stellt daher eine pragmatische Lösung dar. Andererseits, und weitaus bedeutender als die Arbeitsersparnis, vermeidet PMM damit einen Informationsverlust, der beim Eurostat-Verfahren durch die Einteilung metrischer Merkmale in Kategorien resultiert, wodurch PMM auch in dieser Hinsicht eine präzisere Distanzberechnung vornimmt. Mit der von Eurostat vorgenommenen Kategorisierung geht einher, dass lediglich Nulldistanzen, also vermeintlich „exakte“ Matches erlaubt sind. Liegen derartige Übereinstimmungen nicht vor, verringert sich aufgrund der Iterationskomponente des Eurostat-Verfahrens die Anzahl der a ausgewählten **X**-Variablen, wodurch das Modell kleiner und somit ein Ausschluss von für den Fusionsprozess relevanten **X**-Variablen riskiert wird. Auch ist beim Eurostat-Verfahren zu beachten, dass exakte Matches nur vermeintlich vorliegen, jedoch durch die Kategorisierung metrischer Variablen exakte Übereinstimmungen äußerst unwahrscheinlich sind. Streng genommen ist die Punktwahrscheinlichkeit, die gleiche metrische Variablenausprägung in beiden Datenfiles, EU-SILC und HBS, vorzufinden, ohnehin gleich 0. Eine Distanzberechnung wäre daher angebracht, die jedoch durch die Kategorisierung übergangen wird. Dementsprechend ist von PMM auch in dieser Hinsicht eine präzisere Suche nach den maximal ähnlichsten Beobachtungen zu erwarten, was dem Fusionsergebnis ebenfalls positiv zugute kommen dürfte.

Zusammenfassend liegen die beiden wesentlichen Unterscheidungspunkte zwischen dem Random Hot-Deck von Eurostat und PMM in der Distanzberechnung: Einerseits, und für den Fusionsprozess wohl besonders entscheidend, geht beim Random Hot-Deck jede ausgewählte Fusionsvariable $X_1, \ldots, X_a$ mit *gleichem* Gewicht in die Suche nach maximal ähnlichen Donoren mit ein. Andererseits nimmt Eurostat zudem einen durch die Kategorisierung metrischer Variablen induzierten *Informationsverlust* in Kauf, weshalb das Random Hot-Deck lediglich Nulldistanzen erlaubt. PMM hingegen *gewichtet* erstens die ausgewählten, in EU-

SILC wie in HBS vorhandenen Fusionsvariablen $X_1, \ldots, X_a$ entlang ihrer Erklärungskraft auf die in EU-SILC zu fusionierenden Konsumvariablen $\mathbf{Z}$ und erlaubt zweitens eine Distanzberechnung *ohne* metrische Variablen kategorisieren zu müssen. Abschließend lässt sich somit folgende Arbeitshypothese formulieren, deren Überprüfung die anstehende Simulationsstudie gewährleisten soll: *Da von Predictive Mean Matching eine präzisere Distanzberechnung ausgeht, führt PMM zu einem besseren Fusionsergebnis als das Random Hot-Deck von Eurostat.*

4 Simulationsdesign

Somit liegt der Fragestellung dieser Arbeit, ob PMM das von Eurostat verwendete Random Hot-Deck-Verfahren optimieren kann, eine Arbeitshypothese zugrunde. Die Überprüfung der Hypothese geschieht, wie bereits erwähnt, im Rahmen einer Simulationsstudie, in der das „wahre“ Fusionsergebnis bekannt ist. In diesem Abschnitt wird daher das Design der Simulationsstudie erläutert, wobei zunächst auf die verfügbare Datenbasis und die dabei relevanten Variablen eingegangen wird. Anschließend ist die methodische Durchführung der anstehenden Monte-Carlo-Simulation zu beleuchten, bevor ein kurzer Abriss der Programmgrundlage sowie der relevanten `R`-Packages erfolgt.

4.1 Datenbasis: EU-SILC 2013 PUFs für Deutschland, Frankreich, Niederlande

Als Datengrundlage dienen zugangsbedingt Public Usefiles (PUFs) von EU-SILC aus dem Jahre 2013[10]. Diese stellen für die Stichprobenziehung der anstehenden Simulationsstudie (siehe Kap. 4.3) die künstliche und substituierte Grundgesamtheit dar. Um zu gewährleisten, dass diese Ersatzpopulation ausreichend groß ist, um Simulationszüge vornehmen

[10] Anderweitige Datenzugänge, etwa über Scientific Usefiles oder Gastwissenschaftlerarbeitsplätze, sind gemäß der Nutzungsbedingungen von Eurostat für Studierende nicht möglich. Entsprechende Anträge wurden abgelehnt. Die Public Usefiles stehen nur bis zum Jahre 2013 zur Verfügung. Neuere Daten (2014 bis heute) sind nicht als Public Usefiles zugänglich. Inhaltliche und erhebungstechnische Details der EU-SILC-Daten von 2013 sind für Simulationszwecke weniger relevant. Jedoch sei für interessierte Leser auf entsprechende Dokumentationen verwiesen, etwa auf Eurostat (2013a). Da die Datenerhebung den nationalen Statistikämtern der EU-Mitgliedstaaten obliegt, liegen für jedes Land eigene Qualitätsberichte und Dokumentationen vor. Für Deutschland siehe Statistisches Bundesamt (2016a). Für Frankreich und die Niederlande sind für 2013 keine derartigen Berichte verfügbar. Ein Überblick der länderspezifischen Qualitätsberichte, sofern vorhanden, ist unter `https://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports#` zu finden.

zu können, werden die EU-SILC-Datensätze für Deutschland ($N_{DE} = 9555$), Frankreich ($N_{FR} = 8645$) und die Niederlande ($N_{NL} = 7657$) gemeinsam betrachtet und über den R-Befehl `rbind()` zeilenmäßig untereinandergebunden. Daraus resultiert eine Fallzahl von $N = N_{DE} + N_{FR} + N_{NL} = 25857$ (EU-SILC 2013 PUF: DE, FR, NL).

Mit Blick auf die Datengrundlage sei angemerkt, dass Public Usefiles zwar für inhaltliche Forschungsanalysen ungeeignet sind, da sie absolut anonymisiert vorliegen und daher ein geringes Analysepotential aufweisen. Sofern sie jedoch lediglich als Daten- und Stichprobengrundlage für eine Simulationsstudie verwendet werden, ist der Rückgriff auf Public Usefiles, falls keine anderweitigen Datenzugänge möglich sind, aus statistisch-methodischer Sicht eine sinnvolle Alternative. Denn insbesondere ist darauf zu achten, dass der datengenerierende Prozess nicht auf Verteilungsfamilien, etwa der multivariaten Normalverteilung, aufbaut, da dies eine faire Beurteilung der Datenfusionsverfahren gefährdet. Denn Random Hot-Deck und Predictive Mean Matching könnten von unterstellten Verteilungen *unterschiedlich* profitieren. Insofern ist eine Simulation, die auf empirischen Daten aufbaut, deutlich sinnvoller, selbst wenn diese lediglich als Public Usefiles zugänglich sind. Damit sind zwar Einschränkungen in der inhaltlichen Aussagekraft verbunden, allerdings sind die Rückschlüsse der anstehenden Simulationsstudie ohnehin nicht inhaltlicher, sondern statistisch-methodischer Natur, wofür Public Usefiles aufgrund der Simulationskomponente geeignet sind. Auch die Tatsache, dass die dieser Arbeit zugrundeliegenden EU-SILC-Daten aus dem Jahre 2013 stammen, obwohl eigentlich eine Fusionierung der EU-SILC- und HBS-Daten von 2015 vorgenommen werden soll, ist für Simulationszwecke und statistisch-methodische Schlussfolgerungen unerheblich. Dementsprechend sind nun auf Basis der vorhandenen Datengrundlage, den EU-SILC PUFs für Deutschland, Frankreich und die Niederlande von 2013, Variablenblöcke $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ auszuwählen, mithilfe derer eine Datenfusion von EU-SILC und HBS simuliert werden kann.

4.2 Variablenauswahl

4.2.1 Auswahl der gemeinsamen X-Variablen

Hinsichtlich der gemeinsamen Merkmale $\mathbf{X}$ hat Eurostat bereits elf Variablen identifiziert, die in EU-SILC wie in HBS erhoben wurden. Davon verwendet Eurostat, je nach Land, eine bestimmte Teilmenge für die länderspezifische Datenfusion. Die Teilmenge bilden dabei jene gemeinsamen Merkmale $\mathbf{X}$, deren Verteilungen über beide Datensätze, EU-SILC und HBS,

vergleichbar und ausreichend ähnlich sind, was beispielsweise anhand der Hellinger-Distanz beurteilt wird (siehe Leulescu und Agafitei 2013; Lamarche 2017). Da die vorliegende Arbeit das Statistische Bundesamt wissenschaftlich unterstützen soll und der dortige Fokus logischerweise auf der Fusionierung der EU-SILC- und HBS-Daten für Deutschland liegt, werden in dieser Arbeit jene **X**-Variablen verwendet, die Eurostat für Deutschland herangezogen hat, was ebenfalls dem R-Code von Eurostat entnommen werden kann.

Eurostat verwendet für Deutschland sieben **X**-Variablen. Diese sollen nun durch die Public Usefiles abgedeckt oder zumindest mit Stellvertretervariablen substituiert beziehungsweise, sofern keine Substitution möglich ist, eigenhändig generiert werden. Tabelle 1 zeigt einen Überblick über die entsprechenden, in der anstehenden Simulationsstudie verwendeten $p = 7$ **X**-Variablen $X_1, \ldots, X_7$ inklusive des jeweiligen Wertebereichs und Skalenniveaus. Dabei wird ersichtlich, dass für das Random Hot-Deck von Eurostat alle Variablen kategorisiert vorliegen, bei PMM hingegen die Variablen Alter (X_2) und Einkommen (X_7) auf dem ursprünglichen, metrischen Skalenniveau verbleiben können.

Gemeinsame X-Variablen	**Wertebereich / Skalenniveau**	
	Eurostat	**PMM**
X_1: Activity Status of RP[a,b]	1 bis 7 / kategorial	1 bis 7 / kategorial
X_2: Age of RP[b]	1 bis 8 / kategorial	gem. X_2 / metrisch
X_3: Population Density Level	1 bis 3 / kategorial	1 bis 3 / kategorial
X_4: Type of Household[a,c,d]	1 bis 4 / kategorial	1 bis 4 / kategorial
X_5: Tenure Status	1 bis 5 / kategorial	1 bis 5 / kategorial
X_6: Main Source of Income[a]	1 bis 2 / kategorial	1 bis 2 / kategorial
X_7: Income	1 bis 5 / kategorial	gem. X_7 / metrisch

[a] Zusätzlich bilden hier die fehlenden Werte eine Kategorie (kodiert als 9);
[b] RP: „Reference Person“ (Befragte Person des gezogenen Haushalts);
[c] Approximiert durch Variable „Dwelling Type“;
[d] Eigentlicher Wertebereich 1 bis 5, Kategorie 5 ist jedoch leer.

Quelle: EU-SILC 2013 PUF: DE, FR, NL.

Tabelle 1: Übersicht der gemeinsamen **X**-Variablen

Der *Activity Status* (X_1) spiegelt den derzeitigen Tätigkeitsstatus (Selbstständiges/Nichtselbstständiges Beschäftigungsverhältnis, Rentner, Erwerbslos, Schüler, etc.) wider (Eurostat 2013a: 277). Dessen Generierung ist im R-Code von Eurostat nicht gänzlich nachzuvollziehen, wodurch der *Activity Status* eigenhändig aus Plausibilitätserwägungen generiert wird, wie in Anhang A dokumentiert. Das *Alter* (X_2) der Referenzperson des Haushalts ist

bei Eurostat in acht Altersklassen unterteilt. Das *Population Density Level* (X_3) bezeichnet die Bevölkerungsdichte der aktuellen Wohngegend und musste zufällig generiert werden, da diese Variable in den Public Usefiles leer ist und durch kein geeignetes, vorhandenes Erhebungsmerkmal substituiert werden kann (EU-SILC 2013 PUF: DE, FR, NL; Eurostat 2013a: 103). Bei der Zufallsgenerierung wurde jedoch darauf geachtet, dass für jedes Land die Häufigkeitsverteilungen der EU-SILC-Originaldaten von 2013 annähernd abgebildet sind, die wiederum dem onlinebasierten Data Explorer von Eurostat entnommen werden können (siehe Eurostat 2019: Data Explorer). Im Anhang A ist die Generierung dokumentiert. Der *Type of Household* (X_4) beinhaltet Informationen über die Anzahl der Personen und eventuell vorhandener Kinder im Haushalt, dessen Generierung durch die Public Usefiles ebenfalls nicht möglich ist. Er kann jedoch durch das vorhandene Merkmal „Dwelling Type“ (Eurostat 2013a: 169) approximiert werden, welches die Art der Unterkunft (Wohnhaus, Wohnung, etc.) und deren grobe Zimmeranzahl widerspiegelt. Der *Tenure Status* (X_5) vereinigt Informationen über das Eigentumsverhältnis der Wohneinheit (Alleiniger Besitzer, Mieter, etc.) und über die im Falle eines Mietverhältnisses anfallenden (klassierten) Mietkosten (Eurostat 2013a: 171, 178). *Main Source of Income* (X_6) betrachtet die Haupteinnahmequelle des Haushalts, wobei die binäre Variable eine Unterscheidung zwischen (1) Einkommen aus selbstständiger/nicht-selbstständiger Arbeit, aus Besitz, Eigentum sowie Vermögen und (2) Einkommen aus Renten, Sozialleistungen sowie sonstigen Transferleistungen vornimmt. Die Variable ist nicht im EU-SILC-Datenfile vorhanden, sondern wurde von Eurostat manuell berechnet. Ihre Generierung kann im `R`-Code von Eurostat jedoch nicht nachvollzogen werden. Es wurde daher entlang der Informationen und Ausführungen in Eurostat (2013a: 7, 306-309, 315-329) und Eurostat (2013b: 20, 27-28) versucht, eine sinnhafte, eigene Berechnung aus Plausibilitätserwägungen vorzunehmen, wie im Anhang A dokumentiert. Das *Income* spiegelt das „Total disposable household income“ (Eurostat 2013a: 208) wider und besteht bei Eurostat aus den fünf Einkommensquintilen.

Mit Blick auf die Variablen *Activity Status* (X_1), *Alter* (X_2) und *Main Source of Income* (X_6) ist zu beachten, dass diese auf Basis von Merkmalen des Personendatenfiles (p-file) generiert werden mussten (Eurostat 2013a: 3-7, 249, 277, 306-309, 315-329). Die Variablen X_4, X_5 und X_7 bauen auf dem Haushaltsdatenfile (h-file), X_3 wiederum auf dem Haushaltsregisterdatenfile (d-file) auf (Eurostat 2013a: 3-7, 103, 169, 171, 178, 208). Der Haushalts- (h-file) und Haushaltsregisterdatenfile (d-file) lassen sich über die Haushaltsidentifikationsnummer (Haushalts-ID) zusammenfügen. Da jedoch im Haushalts- und Haushaltsregisterdatenfile keine Personenidentifikationsnummer (Personen-ID) vorliegt, wurde diese länder-

spezifisch, also für Deutschland, Frankreich und die Niederlande separat und zufällig generiert, weshalb die personenbezogenen Variablenausprägungen von *Activity Status*, *Alter* und *Main Source of Income* zufällig den Beobachtungseinheiten der haushaltsbezogenen Datenfiles hinzugefügt wurden, um eine möglichst vergleichbare Fusionssituation entlang der von Eurostat verwendeten **X**-Variablen zu simulieren. Denn auch Eurostat fügt dem Haushaltsdatenfile (h-file) die entsprechenden Variablen aus dem Personendatenfile (p-file) und dem Haushaltsregisterdatensatz (d-file) hinzu, wobei die konkrete Datenfusion dann anhand des um die Variablen des p- und d-files erweiterten Haushaltsdatensatzes erfolgt. Die Simulation der Datenfusion von EU-SILC und HBS wird somit unter Einbeziehung von sieben **X**-Variablen, die so präzise wie möglich die von Eurostat für Deutschland verwendeten Fusionsvariablen repräsentieren sollen, gewährleistet.

4.2.2 Auswahl der spezifischen Variablen Y und Z

Für die nicht gemeinsam beobachteten Variablen **Y** und **Z**, die die besonders umfassenden Einkommensvariablen in EU-SILC (**Y**) sowie die detailliert erfassten Konsumausgaben (**Z**) in HBS widerspiegeln, sind ebenfalls Variablen aus den zugrundeliegenden EU-SILC-Daten von 2013 auszuwählen. Eurostat hat jedoch bisher lediglich *eine* spezifische HBS-Variable, namentlich die Gesamtkonsumausgaben des Haushalts, in EU-SILC hineinfusioniert. In dieser Arbeit sollen daher $\mathbf{Z} = (Z_1, Z_2)$, also $p_{hbs} = 2$ spezifische HBS-Variablen verwendet werden, um exemplarisch aufzuzeigen, dass die univariate Datenfusion (Fusionierung *einer* HBS-Variable) auch multivariat mit mehr als einer spezifischen HBS-Variable vorgenommen werden kann. Um die $p_{hbs} = 2$ **Z**-Variablen mit der Verteilung der spezifischen Einkommensvariablen in EU-SILC analysieren zu können, werden für EU-SILC ebenso $p_{silc} = 2$ spezifische Variablen $\mathbf{Y} = (Y_1, Y_2)$ ausgewählt.

Klar ist, dass eine inhaltliche Abdeckung mit den spezifischen, nicht gemeinsam beobachteten Variablen **Y** und **Z** allenfalls für die Einkommensvariablen aus EU-SILC möglich ist, weil die Datengrundlage ihrerseits aus EU-SILC besteht. Da jedoch keine inhaltlichen, sondern statistisch-methodische Schlussfolgerungen von Interesse sind, ist die inhaltliche Abdeckung nachrangig. Vielmehr ist für jene Variablen, die die Einkommensangaben aus EU-SILC, als auch für jene Merkmale, die die Konsumausgaben aus dem HBS widerspiegeln sollen, auf gleiches Skalenniveau zu achten. Da Einkommen und Konsum metrische Merkmale darstellen, ist es essentiell, hierfür *metrische* Variablen auszuwählen. Jedoch weisen eine Vielzahl der im Haushaltsdatenfile (h-file) vorhandenen, metrischen Variablen einen

hohen Anteil fehlender Werte auf, sind komplett leer oder beinhalten einen übermäßigen Anteil an 0-er-Ausprägungen, was unter anderem dem absoluten Anonymisierungsgrad der Public Usefiles geschuldet sein dürfte (EU-SILC 2013 PUF: DE, FR, NL). Daher bestehen die ausgewählten **Y**- und **Z**-Variablen aus jenen metrischen Merkmalen, die möglichst wenig fehlende Werte und 0-er-Ausprägungen aufweisen, um eine gehaltvolle Fusionssimulation zu ermöglichen. Dies reduziert bereits die verfügbaren Variablen erheblich, wobei dennoch vier metrische Merkmale mit den gewünschten Kriterien gefunden werden konnten, die als Approximation für (Y_1, Y_2) beziehungsweise (Z_1, Z_2) dienen.

Dabei wird für Y_1 die Variable „Total disposable household income before social transfers including old-age and survivor's benefits" (Eurostat 2013a: 208) sowie für Y_2 die Variable „Interest, dividends, profit from capital investments in unincorporated business" (Eurostat 2013a: 217) verwendet. Während hierbei eine inhaltliche Übereinstimmung mit Einkommen gegeben ist, ist bei den Substituten, die die Konsumvariablen in HBS widerspiegeln sollen, besonders das metrische Skalenniveau relevant, weshalb für Z_1 die Variable „Total household gross income" (Eurostat 2013a: 206) sowie für Z_2 das erhobene Merkmal „Total disposable household income before social transfers other than old-age and survivor's benefits" (Eurostat 2013a: 208) ausgewählt wird. Der Übersichtlichkeit wegen zeigt Tabelle 2 nochmals die spezifischen und für die Simulationsstudie verwendeten **Y**- und **Z**-Variablen inklusive ihres Skalenniveaus.

Spezifische EU-SILC-Variablen (Y)	**Skalenniveau**
Y_1: Total disposable household income before social transfers including old-age and survivor's benefits	metrisch
Y_2: Interest, dividends, profit from capital investments in unincorporated business	metrisch
Spezifische HBS-Variablen (Z)	**Skalenniveau**
Z_1: Total household gross income	metrisch
Z_2: Total disposable household income before social transfers other than old-age and survivor's benefits	metrisch

Quelle: EU-SILC 2013 PUF: DE, FR, NL.

Tabelle 2: Übersicht der spezifischen Variablen für EU-SILC (**Y**) und HBS (**Z**)

Somit ist die Datengrundlage, die auf den EU-SILC PUFs von 2013 für Deutschland, Frankreich und die Niederlande aufbaut, präsent. Daraus wurden sieben gemeinsame **X**-Variablen ausgewählt, die, zumindest approximativ, den von Eurostat verwendeten gemeinsamen Merk-

malen für Deutschland entsprechen. Für die nicht gemeinsam beobachteten, spezifischen Variablenblöcke $\mathbf{Y}$ und $\mathbf{Z}$ dienen jeweils zwei Substitute, die ihrerseits metrisches Skalenniveau, wie auch die tatsächlich zu fusionierenden Einkommens- und Konsumvariablen, aufweisen. Die daraus resultierende Datengrundlage stellt die künstliche Grundgesamtheit dar, in der die wahre, gemeinsame Verteilung von (Y_1, Y_2, Z_1, Z_2) bekannt ist. Aus dieser Datenbasis heraus kann nun die Datenfusion von EU-SILC und HBS simuliert werden.

4.3 Methode: Monte-Carlo-Simulation

Die methodische Durchführung der Simulationsstudie entspricht dabei einer Monte-Carlo-Simulation (kurz: MC-Simulation). Allgemein handelt es sich dabei um ein computerbasiertes und rechenintensives Experiment, bei dem Daten durch k Zufallsstichproben aus bekannten Verteilungen generiert beziehungsweise gezogen werden, um etwa die Performance verschiedener statistischer Verfahren beurteilen zu können (siehe z.B. Morris et al. 2019). Im Rahmen dieser Arbeit stellt sich die MC-Simulation folgendermaßen dar: Zunächst werden k Zufallsstichproben der Größe n gezogen, für die sodann das Datenausfallmuster einer Datenfusion (siehe Abb. 2) generiert wird, bevor die fehlenden Z_1- und Z_2-Werte im simulierten EU-SILC-Datenfile mittels beider Verfahren, Random Hot-Deck und PMM, ergänzt werden sollen.

Für jeden der k Zufallszüge wird dabei ein Jackknife-Sampling verwendet, das heißt, es erfolgt eine Ziehung *ohne* Zurücklegen (siehe z.B. Rodgers 1999). Anschließend sind die Daten der Stichprobe in zwei Datensätze zu zerteilen, sodass der eine Datenfile EU-SILC mit den beobachteten Variablen $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$ und $\mathbf{Y} = (Y_1, Y_2)$ ohne Informationen zu den $\mathbf{Z}$-Variablen darstellt, der andere wiederum HBS mit den beobachteten Variablen $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$ und $\mathbf{Z} = (Z_1, Z_2)$ repräsentiert, der wiederum keine Informationen über die $\mathbf{Y}$-Variablen beinhaltet. Ein „Untereinanderbinden“ beider Datenquellen führt zum spezifischen Datenausfallmuster einer Datenfusion, wie in Abbildung 2 dargestellt. Analog zu Eurostat sollen dann die fehlenden Merkmalsausprägungen für Z_1 und Z_2 im simulierten EU-SILC-Datenfile mittels der für diese Arbeit relevanten Fusionsverfahren, Random Hot-Deck und PMM, imputiert werden, wodurch sie im fusionierten Datenfile eine künstliche Verteilung $\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, \widetilde{Z}_2)$ widerspiegeln. Anschließend sind die Korrelationen zwischen $\mathbf{Y}$ und $\widetilde{\mathbf{Z}}$ sowie zwischen den metrischen $\mathbf{X}$-Variablen (X_2: Alter und X_7: Einkommen) und $\widetilde{\mathbf{Z}}$ zu berechnen und mit den wahren Korrelationen zu vergleichen, die wiederum aus der in Kapitel 4.1 beschriebenen und $N = 25857$ Individuen umfassenden

Ersatzpopulation bekannt sind. Dabei wird Single Imputation ($M = 1$) angewandt, woraus demzufolge Punktschätzer für die Korrelationen resultieren. Dieser Prozess, von der Stichprobenziehung über die Imputation bis zur Korrelationsberechnung, wird 1000 mal, also mit $k = 1000$ MC-Simulationszügen wiederholt. Damit kann letztlich die Performance beider Verfahren hinsichtlich der in Kapitel 2.3 beschriebenen, relevanten Validitätsstufen 3 und 4 beurteilt werden, wobei besonderes Augenmerk auf den Korrelationen zwischen $\mathbf{Y}$ und $\widetilde{\mathbf{Z}}$, also der dritten Validitätsstufe liegt.

Um im Zuge der MC-Simulationsstudie eine grobe Beurteilung dahingehend vorzunehmen, wie sensitiv sich das Verhältnis zwischen den Stichprobengrößen des Rezipientendatenfiles (n_{silc}) und des Donorendatenfiles (n_{hbs}) auf die Fusionsergebnisse beider Verfahren auswirken könnte, soll der den Simulationszügen zugrundeliegende Stichprobenumfang n variiert werden. Dabei ist besonders von Interesse, inwiefern sich eine übermäßige Anzahl an Donoren ($n_{silc} << n_{hbs}$) im Vergleich zu einem gleichmäßigen Rezipienten- und Donorenverhältnis ($n_{silc} = n_{hbs}$) auf die Datenfusionsperformance auswirkt. Dementsprechend wird die oben beschriebene MC-Simulation zweimal durchgeführt, jedoch unter Verwendung verschiedener Stichprobengrößen n_1 und n_2. Im Rahmen der ersten MC-Simulation werden für jeden der $k = 1000$ Simulationszüge $n_1 = 600$ Beobachtungseinheiten aus der Datengrundlage gezogen, wobei gleich viele, nämlich jeweils 300 Beobachtungen dem EU-SILC- und dem HBS-Datenfile zugeordnet werden. Ein geringerer Stichprobenumfang ist nicht möglich, da die `regsubsets()`-Funktion aus dem `leaps`-Package von Lumley und Miller (2017), die Eurostat für die Backward Selection verwendet, aufgrund der durch Kategorisierung bedingten Vielzahl an Kovariaten einen wesentlich geringeren Stichprobenumfang als $n_{hbs} = 300$ nicht handhaben kann[11]. Daher gilt für die erste MC-Simulation $n_1 = n_{1_{silc}} + n_{1_{hbs}} = 300 + 300 = 600$. Die zweite MC-Simulation betrachtet ebenso $n_{1_{silc}} = n_{2_{silc}} = 300$ Beobachtungseinheiten in EU-SILC, wobei diesen 300 Rezipienten nun deutlich mehr, nämlich $n_{2_{hbs}} = 2700$ potentielle Donoren gegenüberstehen. Dementsprechend besteht der zweite Stichpobenumfang aus $n_2 = 5 \cdot n_1 = n_{1_{silc}} + 9 \cdot n_{1_{hbs}} = n_{2_{silc}} + n_{2_{hbs}} = 300 + 2700 = 3000$ Beobachtungen.

[11] Beachte, dass die Regressionen von $\mathbf{Z}$ auf $\mathbf{X}$ anhand des Donorendatenfiles, also entlang des HBS berechnet werden.

4.4 Programmgrundlage und R-Packages

Beide MC-Simulationen werden, wie alle datenbezogenen Aufbereitungs- und Analyseschritte, wie bereits deutlich wurde, mit dem Statistikprogramm R (Version 3.5.2) durchgeführt, wobei als Oberfläche RStudio (Version 1.1.463) genutzt wird. In diesem Abschnitt werden kurz die R-Packages dargelegt, die neben der Basisinstallation für die vorliegende Arbeit relevant sind.

Für das Random Hot-Deck-Verfahren von Eurostat werden folgende R-Packages verwendet: Das StatMatch-Package von D'Orazio (2017) zum Zwecke der Generierung eines fusionierten Datenfiles; das sqldf-Package von Grothendieck (2017) sowie das dplyr-Package von Wickham et al. (2019) zum Zweck des Findens von Matches entlang der ausgewählten Variablen $X_1, \ldots, X_a$; das bereits erwähnte leaps-Package von Lumley und Miller (2017), mithilfe dessen regubsets()-Funktion die Backward Selection für die Auswahl der in den Fusionsprozess einfließenden a gemeinsamen Variablen $X_1, \ldots, X_a$ erfolgt; das sampling-Package von Tillé und Matei (2016) zum Durchführen der Zufallszuweisungen zwischen Rezipient und Donor sowie das stringr-Package von Wickham (2019) zur Stringmanipulation im Rahmen der Variablenmodifikation. Auch mithilfe des dplyr-Packages von Wickham et al. (2019) wird, über die Funktion mutate(), die Variablenumkodierung vorgenommen. Entlang dieser genannten R-Packages hat Eurostat eine manuelle, durchaus aufwendige Programmkodierung des Random Hot-Deck-Verfahrens vorgenommen.

Für die Datenfusion mittels Predictive Mean Matching kann hingegen das BaBooN-Package von Meinfelder und Schnapp (2015) verwendet werden, das mit BBPMM.row() bereits eine passgenaue Funktion beinhaltet, die (unter anderem) speziell für die Durchführung von Datenfusionen über Predictive Mean Matching programmiert wurde und einen fusionierten Datenfile generiert (Koller-Meinfelder 2009: 93-96). Hier erfolgt die selektive Auswahl der a Fusionsvariablen $X_1, \ldots, X_a$ zwar nicht über regusbsets(), sondern über die Funktion stepAIC() aus dem MASS-Package[12] von Ripley et al. (2018), wobei BBPMM.row() per Default eine Backward Selection anwendet (Meinfelder und Schnapp 2015). Sofern PMM also gemäß der unterstellten Arbeitshypothese aus Kapitel 3.4 ein besseres Fusionsergebnis produzieren würde, ginge damit für das Fusionsvorhaben der amtlichen Statistik der Vorteil einher, dass für das PMM-Verfahren bereits eine einfache und unkomplizierte Umsetzung der Datenfusion über das BaBooN-Package möglich ist. Darüber hinaus muss bei der Eurostat-Kodierung für jede zu fusionierende **Z**-Variable die Fusionsprogrammierung an-

[12] Das MASS-Package von Ripley et al. (2018) wird aber nicht direkt benötigt.

gepasst und erweitert werden, was den Arbeitsaufwand noch weiter erhöht, während die Funktion `BBPMM.row()` automatisch eine Fusionierung *aller* **Z**-Variablen vornimmt, was für die Praxis eine erhebliche Arbeitsersparnis darstellt.

Neben den für die Fusionsverfahren verwendeten `R`-Packages wird für die grafische Ergebnisdarstellung auf `ggplot2` von Wickham et al. (2018), `ggthemes` von Talbot et al. (2019), `gridExtra` von Auguie und Antonov (2017), `gridGraphics` von Murrell und Wen (2018) sowie auf `latex2exp` von Meschiari (2015) zurückgegriffen. Für die Bewertung und Diskussion der anstehenden Simulationsstudie ist kurz Pearson's η^2 relevant, dessen Berechnungen mit dem `sjstats`-Package von Lüdecke (2019) durchgeführt werden.

Abschließend konnten nun aus der vorhandenen Datengrundlage Variablen für $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ ausgewählt werden, welche die von Eurostat angewandte Datenfusion von EU-SILC und HBS, die eine Analyse der gemeinsamen Verteilung von Einkommen und Konsumausgaben der privaten Haushalte gewährleisten soll, simulieren kann. Dabei wird die Simulationsstudie mit zwei Stichprobenumfängen durchgeführt, einmal mit $n_1 = n_{1_{silc}} + n_{1_{hbs}} = 300 + 300 = 600$ und einmal mit $n_2 = n_{2_{silc}} + n_{2_{hbs}} = 300 + 2700 = 3000$, wobei zu Beginn mittels `set.seed()` ein Seed (hier: 1234) gesetzt wird, um die Ergebnisse reproduzieren zu können. Die Performance beider Verfahren ist besonders anhand von $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$, also den geschätzten Korrelationen zwischen $\mathbf{Y}$ und $\widetilde{\mathbf{Z}}$, aber auch anhand von $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$, also den geschätzten Korrelationen zwischen $\mathbf{X}$ und $\widetilde{\mathbf{Z}}$ zu beurteilen. Beides erfolgt anhand der Ergebnisse der MC-Simulationsstudie, die im folgenden Kapitel dargestellt, bewertet und diskutiert werden sowie die Formulierung konkreter Handlungsperspektiven für die amtliche Statistik gewährleisten sollen.

5 Ergebnisse der Monte-Carlo-Simulation

5.1 Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$

Insbesondere ist bei Datenfusionen von Interesse, die Korrelationsstrukturen zwischen den nicht gemeinsam beobachteten Variablenblöcken $\mathbf{Y}$ und $\mathbf{Z}$ im fusionierten Datensatz zu erhalten, um in den späteren, wissenschaftlich inhaltlichen Analysen valide Schlussfolgerungen ziehen zu können. Gemäß der zugrundeliegenden Arbeitshypothese aus Kapitel 3.4 wird diesbezüglich erwartet, dass für Predictive Mean Matching aufgrund der präziseren Distanzmessung, die besonders der vorgenommenen Gewichtung der a ausgewählten, in den Fusionsprozess einfließenden Variablen $X_1, \ldots, X_a$ geschuldet sein dürfte, ein besseres Fusi-

onsergebnis resultiert, als für das Random Hot-Deck-Verfahren von Eurostat. Demzufolge sollte PMM, im Vergleich zur Eurostat-Fusionsmethode, die unbeobachteten Korrelationen zwischen $\mathbf{Y}$ und $\mathbf{Z}$ präziser reproduzieren können.

Um diese theoretische Erwartung zu beurteilen, sind in Tabelle 3 die aus der $N = 25857$ Beobachtungen umfassenden, künstlichen Grundgesamtheit resultierenden Korrelationen zwischen den spezifischen Variablen $\mathbf{Y} = (Y_1, Y_2)$ und $\mathbf{Z} = (Z_1, Z_2)$ als Benchmarks beziehungsweise wahre Werte abgetragen. Ersichtlich ist dabei, dass zwischen Y_1 und Z_1 sowie zwischen Y_1 und Z_2 relativ hohe Korrelationen (0.79 beziehungsweise 0.86), für Y_2 und Z_1 sowie für Y_2 und Z_2 hingegen eher mittelstarke (positiv lineare) Zusammenhänge vorliegen (0.25 beziehungsweise 0.29).

Wahre Korrelationen zwischen Y und Z

$\text{corr}(Y_1, Z_1)$	$\text{corr}(Y_1, Z_2)$	$\text{corr}(Y_2, Z_1)$	$\text{corr}(Y_2, Z_2)$
0.7948	0.8633	0.2543	0.2866

Quelle: EU-SILC 2013 PUF: DE, FR, NL.

Tabelle 3: Wahre Werte für $\rho_{\mathbf{YZ}}$

Wie gut können nun die beiden relevanten Datenfusionsalgorithmen, Random Hot-Deck und PMM, diese Korrelationen reproduzieren? Zur Klärung dessen werden im Folgenden aus Gründen der Übersichtlichkeit und Veranschaulichung besonders grafische Abbildungen diskutiert. Zunächst werden für beide Stichprobenumfänge n_1 und n_2 die Verteilungen der geschätzten Korrelationen $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ über alle $k = 1000$ Monte-Carlo-Simulationszüge deskriptiv betrachtet, bevor anschließend mit den Monte-Carlo-Varianzen, dem Bias sowie dem Mean Squared Error (MSE) auf weiterführende Diagnostiken einzugehen ist.

5.1.1 Monte-Carlo-Verteilungen

Die Monte-Carlo-Verteilungen spiegeln die Korrelationen zwischen den spezifischen Variablen der (simulierten) Rezipientenstichprobe EU-SILC ($\mathbf{Y}$) und den darin mittels Random Hot-Deck und PMM imputierten Variablenausprägungen der (simulierten) Donorenstichprobe HBS ($\widetilde{\mathbf{Z}}$) für alle $k = 1000$ Simulationsrunden unter den Stichprobenumfängen n_1 und n_2 wider, wobei die spezifischen HBS-Variablen nun eine künstliche Verteilung im Fusionsdatenfile aufweisen. Alle in diesem Abschnitt referierten, exakten Werte sind im Anhang B in den Tabellen 6 und 7 zu finden.

Die Kerndichten in Abbildung 3 zeigen für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ unter n_1, also einem gleichmäßigen Rezipienten- und Donorenverhältnis mit $n_{1_{silc}} = n_{1_{hbs}} = 300$, dass PMM in aggregierter Betrachtungsweise den wahren Korrelationen (0.79 beziehungsweise 0.86) deutlich näher kommt, als das Random Hot-Deck-Verfahren von Eurostat. Der Eurostat-Ansatz deckt über alle $k = 1000$ Simulationsrunden für n_1 nie, nichtmal durch Ausreißer, den unmittelbaren Bereich um die wahren Korrelationen von 0.79 beziehungsweise 0.86 ab – das jeweilige Maximum liegt bei 0.75 für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ beziehungsweise bei 0.74 für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$. Ebenso ist die breite Masse der Eurostat-Korrelationen weit von den wahren Parameterwerten entfernt, was auch entlang der Boxplots für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ in Abbildung 5 gut erkennbar ist: Beide Eurostat-Boxen, dessen Unter- und Obergrenze das 25%- und 75%-Quantil darstellt und dessen mittlere Markierung dem Median, also dem 50%-Quantil entspricht, sind deutlich unter den tatsächlichen Korrelationen verortet. Konkret liegen für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ die mittleren 50 % der bei Eurostat resultierenden Zusammenhänge zwischen 0.51 (25%-Quantil) und 0.61 (75%-Quantil), für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ zwischen 0.50 und 0.61. Somit nehmen bereits 75 % der durch das Random Hot-Deck errechneten Parameter höchstens einen Wert von 0.61 an, womit dreiviertel der Eurostat-Korrelationen mindestens um eine Differenz von $0.79 - 0.61 = 0.18$ beziehungsweise $0.86 - 0.61 = 0.25$ kleiner sind als die wahren Korrelationen zwischen Y_1 und Z_1 beziehungsweise zwischen Y_1 und Z_2.

Die unter PMM errechneten Zusammenhänge sind hingegen den hohen Originalkorrelationen bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ in ihrer Vielzahl zumindest deutlich näher und können diese häufiger annähernd abdecken, was im Kerndichteplot in Abbildung 3 gut erkennbar ist. Dabei ist jedoch auch zu sehen, dass die PMM-Performance bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ etwas besser ist, als das PMM-Fusionsergebnis bei einer noch stärkeren, wahren Korrelation von 0.86 bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ – hier decken die errechneten Zusammenhänge zu einem geringeren Anteil den unmittelbaren Bereich um den tatsächlichen Parameterwert von 0.86 ab, wobei PMM aber auch bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ im Vergleich zu Eurostat eine deutlich bessere Performance zuzuschreiben ist. So ist auch in den Boxplots in Abbildung 5 für PMM unter n_1 zu erkennen, dass die mittleren 50 % der k errechneten Korrelationen bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ zwischen 0.72 und 0.80, bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ hingegen zwischen 0.74 und 0.82 liegen. Besonders bei Eurostat ist darüber hinaus zu beobachten, dass geringfügige Änderungen in den starken Originalzusammenhängen, wie sie zwischen den wahren Parameterwerten 0.79 und 0.86 vorliegen, ein kaum verändertes Fusionsergebnis zur Folge hat. Bei PMM unterscheidet sich die Verteilung der Korrelationsschätzer zwischen $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ zwar leicht voneinander, so fallen die Zusammenhänge bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ aggregiert betrachtet etwas höher aus. Jedoch

steigen sie keineswegs in selber Intensität an, wie die wahre Korrelation zwischen Y_1 und Z_2 (0.86) im Vergleich zur Korrelation zwischen Y_1 und Z_1 (0.79), die sich um eine Differenz von $0.86 - 0.79 = 0.07$ voneinander unterscheiden.

Für die hohen Originalkorrelationen zwischen Y_1 und Z_1 sowie zwischen Y_1 und Z_2 ist mit Blick auf den Stichprobenumfang n_2, bei dem den Rezipienten im EU-SILC-Datenfile ($n_{2_{silc}} = 300$) deutlich mehr Donoren aus dem HBS ($n_{2_{hbs}} = 2700$) gegenüberstehen und dessen Kerndichten und Boxplots in den Abbildungen 4 und 6 dargestellt sind, festzuhalten, dass sich beim Random Hot-Deck-Verfahren für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ kaum Unterschiede im Vergleich zur Fusionsperformance bei n_1 einstellen. Die Reproduktion der hohen Originalkorrelationen scheint sich in deskriptiver Betrachtung für PMM unter n_2 geringfügig zu verbessern: Wie die Dichten in Abbildung 4 nahelegen, konzentriert sich unter n_2 bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ eine höhere Verteilungsmasse um die wahre Korrelation von 0.79 beziehungsweise 0.86, als im Vergleich zu den jeweiligen PMM-Dichten unter n_1 in Abbildung 3. Eine sehr gute Abdeckung des tatsächlichen Parameterwertes erreicht PMM nun bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$, was auch entlang des entsprechenden Boxplots in Abbildung 6 zu erkennen ist: Die mittleren 50 % der Korrelationen liegen bei PMM unter n_2 im Bereich von 0.74 und 0.82 und steigen somit im Vergleich zu n_1 bei PMM jeweils um 0.02 an, wobei der Median, also das 50%-Quantil, mit einem Wert von 0.78 für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ nun nahezu der wahren Korrelation von 0.79 entspricht. Anzumerken ist jedoch, dass dies nur marginale Verbesserungen darstellen. Gerade mit Blick auf Zufallsschwankungen verbinden sich damit keine allzu stichhaltigen Indizien für eine verbesserte PMM-Performance bei einem übermäßigen Donorenverhältnis. Das etwas veränderte PMM-Fusionsergebnis unter n_2 ist daher lediglich als Tendenz und dezenter Hinweis dahingehend zu sehen, dass PMM etwas Sensitivität mit Blick auf eine hohe Donorendominanz aufweisen könnte. Eurostat scheint hingegen kaum sensitiv gegenüber gleich- und übermäßigen Donorenverhältnissen zu sein. Insgesamt ist für hohe Originalzusammenhänge, wie sie bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ vorliegen, festzuhalten, dass PMM, sowohl unter n_1, also auch unter n_2, den wahren Korrelationen deutlich näher kommt, als das Random Hot-Deck-Verfahren von Eurostat. Es zeigte sich eine geringe Tendenz dahingehend, dass PMM bei hohen Originalkorrelationen von einer sehr hohen Donorenüberzahl etwas profitieren könnte.

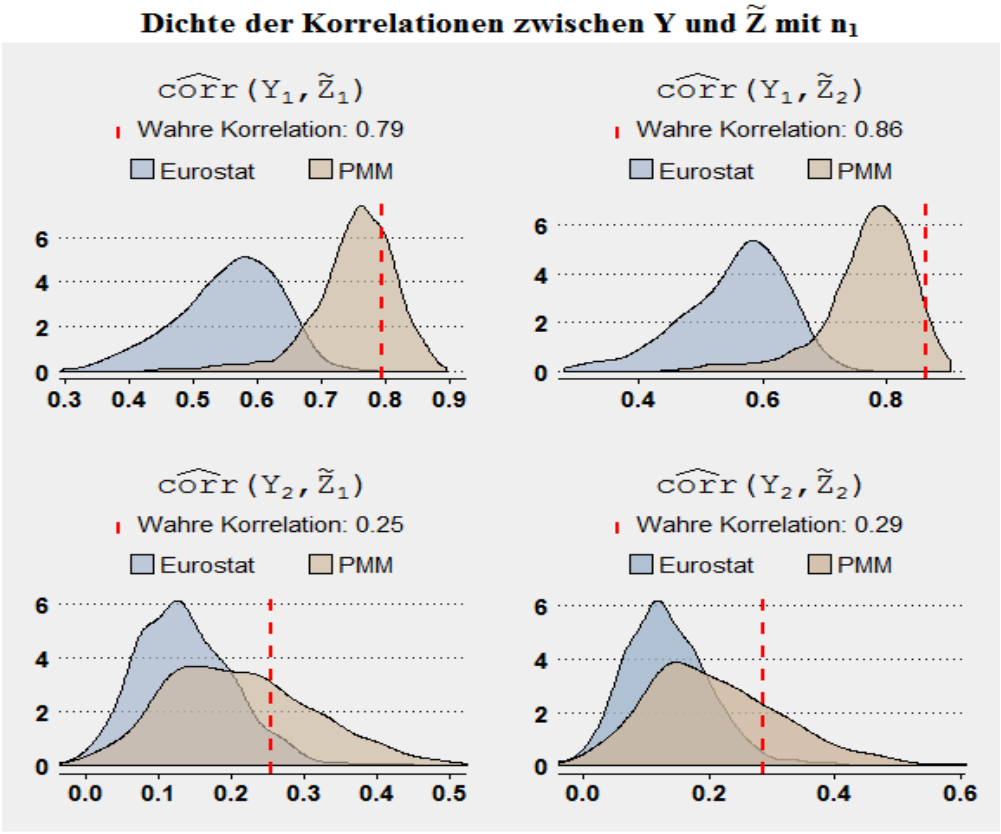

Abbildung 3: Kerndichteplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

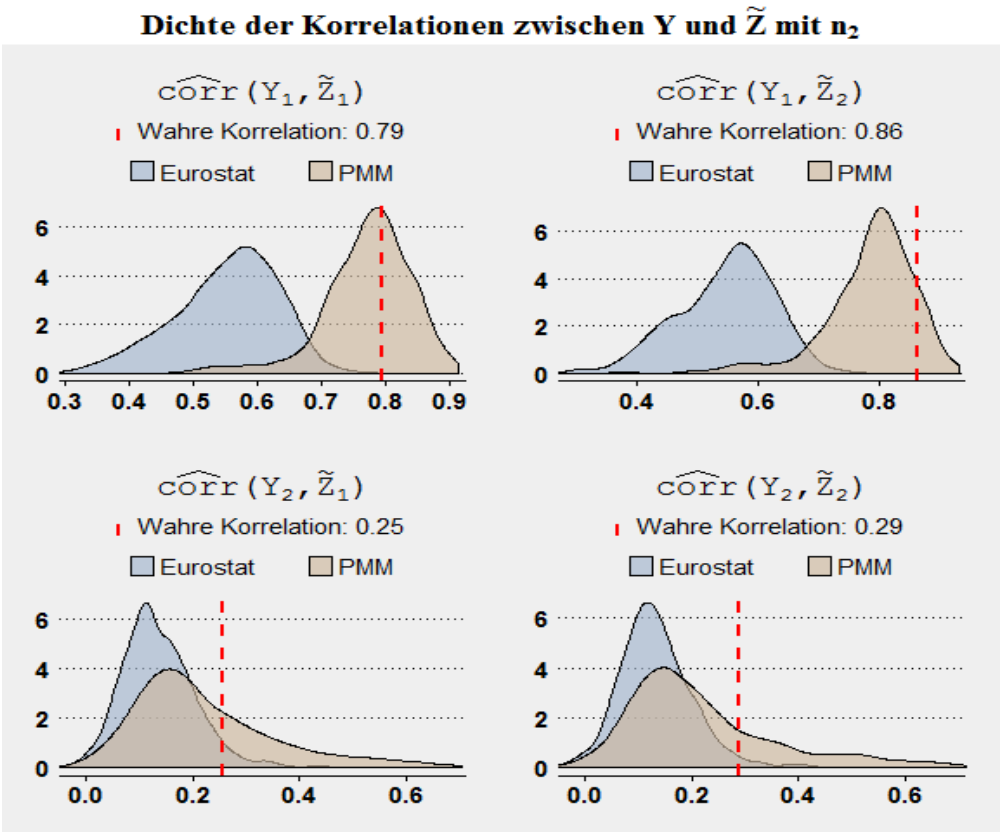

Abbildung 4: Kerndichteplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

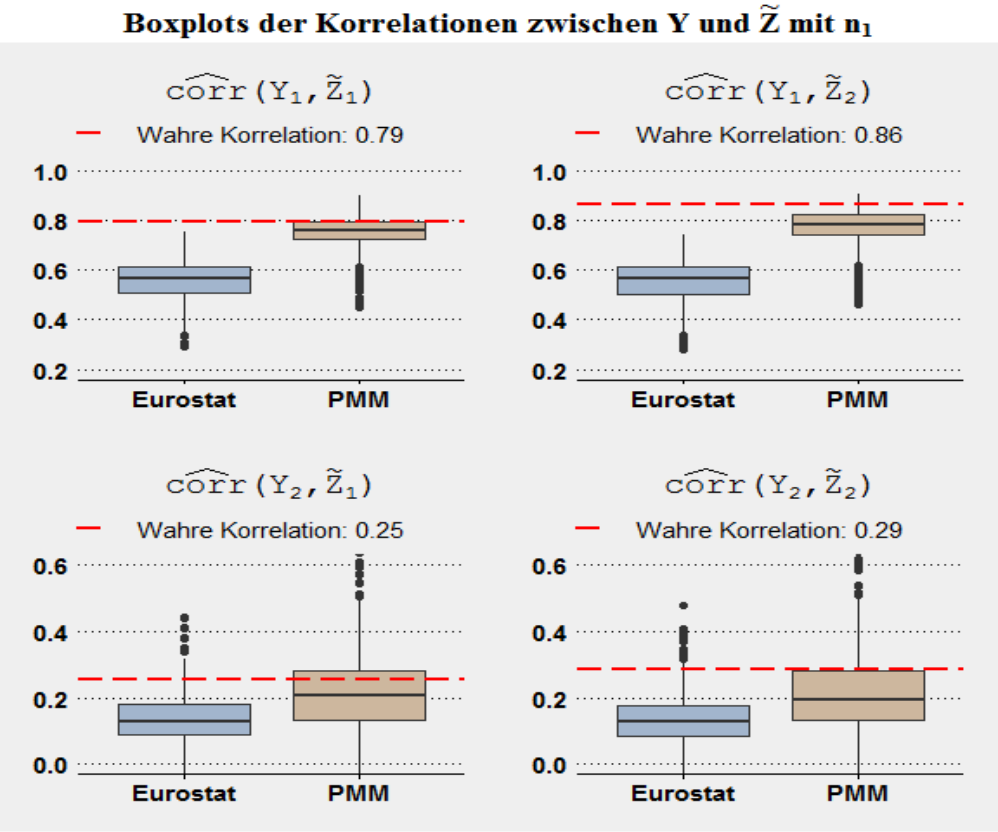

Abbildung 5: Boxplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

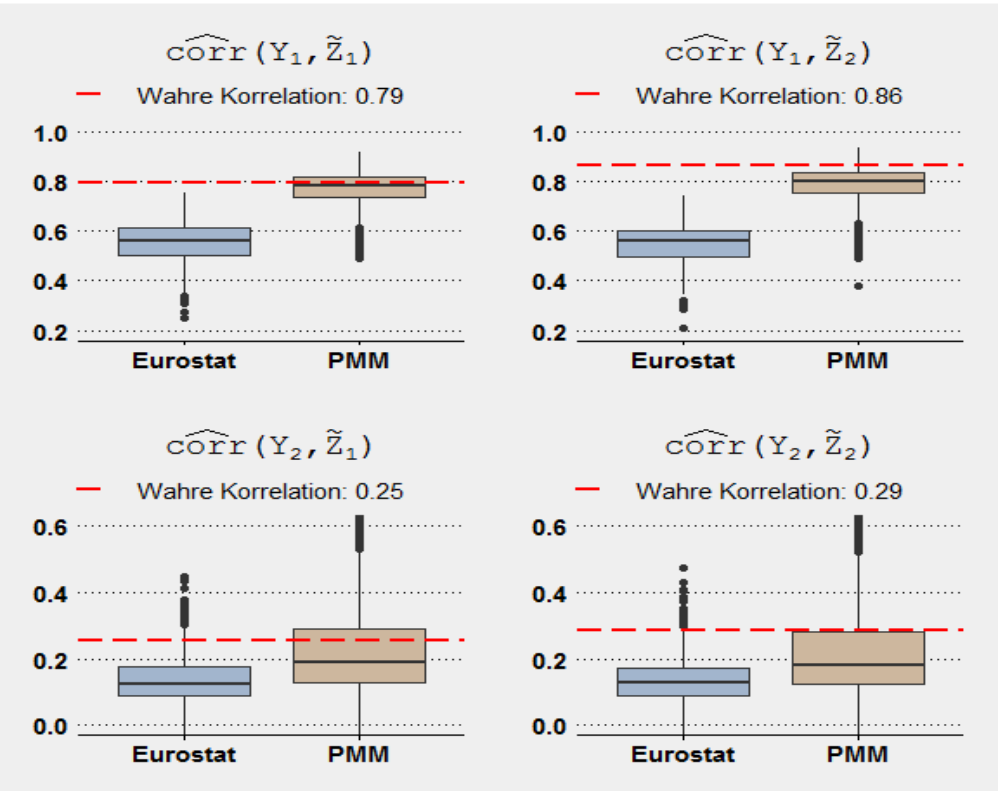

Abbildung 6: Boxplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

Für die mittleren Originalzusammenhänge zwischen Y_2 und Z_1 sowie zwischen Y_2 und Z_2 unterscheidet sich die Performance von Random Hot-Deck und PMM hingegen etwas weniger. Die Dichten für den Stichprobenumfang n_1 in Abbildung 3 zeigen, dass Eurostat nun zumindest vereinzelt die wahren Korrelationen von 0.25 beziehungsweise 0.29 abdecken kann, die überwiegende Verteilungsmasse aber dennoch in einem Bereich liegt, der geringer als die tatsächlichen Zusammenhänge ist. So ist etwa in den Boxplots in Abbildung 5 für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ ersichtlich, dass die mittleren 50 % der beim Random Hot-Deck-Verfahren resultierenden Zusammenhänge zwischen 0.09 und 0.18 liegen. 75 % der Eurostat-Korrelationsschätzer nehmen höchstens einen Parameterwert von 0.18 an, während die wahre Korrelation bei 0.25 beziehungsweise 0.29 liegt, wodurch Eurostat für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ eine bessere Performance gewährleistet, als für die etwas höhere, tatsächliche Korrelation von 0.29 bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$. Die Dichten für PMM in Abbildung 3 zeigen hingegen für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$, dass, im Vergleich zu Eurostat, eine höhere Verteilungsmasse zumindest in der Nähe der tatsächlichen Korrelationen liegt, wobei PMM sich dem wahren Zusammenhang zwischen Y_2 und Z_1 etwas häufiger annähert, als jenem zwischen Y_2 und Z_2. Dies ist auch in den Boxplots in Abbildung 5 erkennbar: Die mittleren 50 % der bei PMM resultierenden Korrelationen liegen sowohl für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$, als auch für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ zwischen 0.13 und 0.28. Insofern scheint PMM für geringe Schwankungen in mittleren Originalkorrelationen ein relativ ähnliches Fusionsergebnis zu produzieren, was auch bei Eurostat zu beobachten ist – die Verteilungen der Eurostat-Korrelationen sind für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ relativ identisch, was für Eurostat bereits bei hohen Originalzusammenhängen zu beobachten war.

Mit Blick auf den Stichprobenumfang n_2, der eine deutlich erhöhte Donorenanzahl aufweist, ist bezüglich der mittleren Originalzusammenhänge bei Eurostat erneut keine substanzielle Veränderung im Vergleich zu n_1 zu erkennen. Die Kerndichten von PMM unter n_2 zeigen hingegen in Abbildung 4 auf, dass für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ eine etwas geringere Masse der $k = 1000$ Korrelationsschätzer den Bereich der wahren Zusammenhänge von 0.25 und 0.29 abdeckt, als dies noch unter einem gleichmäßigen Rezipienten- und Donorenverhältnis in n_1, ersichtlich in Abbildung 3, für PMM der Fall war. Andererseits sind wiederum die Mittelwerte der Korrelationen, die für n_1 und n_2 in den Abbildungen 7 und 8 grafisch veranschaulicht sind, bei PMM für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ unter n_2 geringfügig höher, als noch unter n_1. Erstens sind dabei aber nur marginale Veränderungen zu beobachten, die mit Blick auf Zufallsschwankungen mit Vorsicht zu genießen sind. Zweitens induzieren die PMM-Dichten für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ unter n_2 (siehe Abb. 4) bereits, dass

PMM, im Vergleich zu n_1, für mittlere Originalkorrelationen eine höhere Ausreißerquote Richtung 1 aufweist, was auf eine etwas stärkere Streuung schließen lässt. Insofern deutet dies bereits darauf hin, dass die deskriptive Betrachtungsweise etwas präziser entlang konkreter Diagnostiken diskutiert werden sollte, weshalb auf die diesbezüglichen Variationen im Stichprobenumfang in Kapitel 5.1.2 noch näher einzugehen ist.

Mit Blick auf die Charakterisierung der MC-Verteilungen veranschaulichen die Kerndichteplots für n_1 und n_2 in den Abbildungen 3 und 4 zudem, dass diese bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, also bei hohen Originalzusammenhängen (0.79 beziehungsweise 0.86) über alle $k = 1000$ Simulationszüge für beide Fusionsverfahren relativ symmetrisch sind. Allenfalls kann den Verteilungen beider Fusionsalgorithmen äußerst leichte Linksschiefe konstatiert werden. PMM ist etwas spitzgipfliger, Eurostat hat hingegen geringfügig stärkere Ränder sowie eine flachgipfligere Kurve, also eine etwas höhere Varianz. Ein umgekehrtes Bild ergibt sich bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$, denen mittlere Originalkorrelationen zugrunde liegen (0.25 beziehungsweise 0.29). Die Verteilung der aus dem Random Hot-Deck-Verfahren von Eurostat resultierenden Korrelationen ist spitzgipfliger, aber noch weitgehend symmetrisch – allenfalls mit äußerst leichter Tendenz zur Rechtsschiefe. PMM hingegen hat hier deutlich stärkere Ränder sowie eine flachgipfligere Kurve und weist demzufolge eine spürbar höhere Varianz in den $k = 1000$ Korrelationen für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ auf. Ebenso scheint bei mittelstarken Originalkorrelationen für PMM eine Tendenz zur Rechtsschiefe vorzuliegen. Übersetzt bedeutet dies, dass Korrelationen, die unter den entsprechenden Mittelwerten liegen, in der MC-Simulation für PMM etwas häufiger vorkommen als Parameterwerte, die größer als das arithmetische Mittel sind.

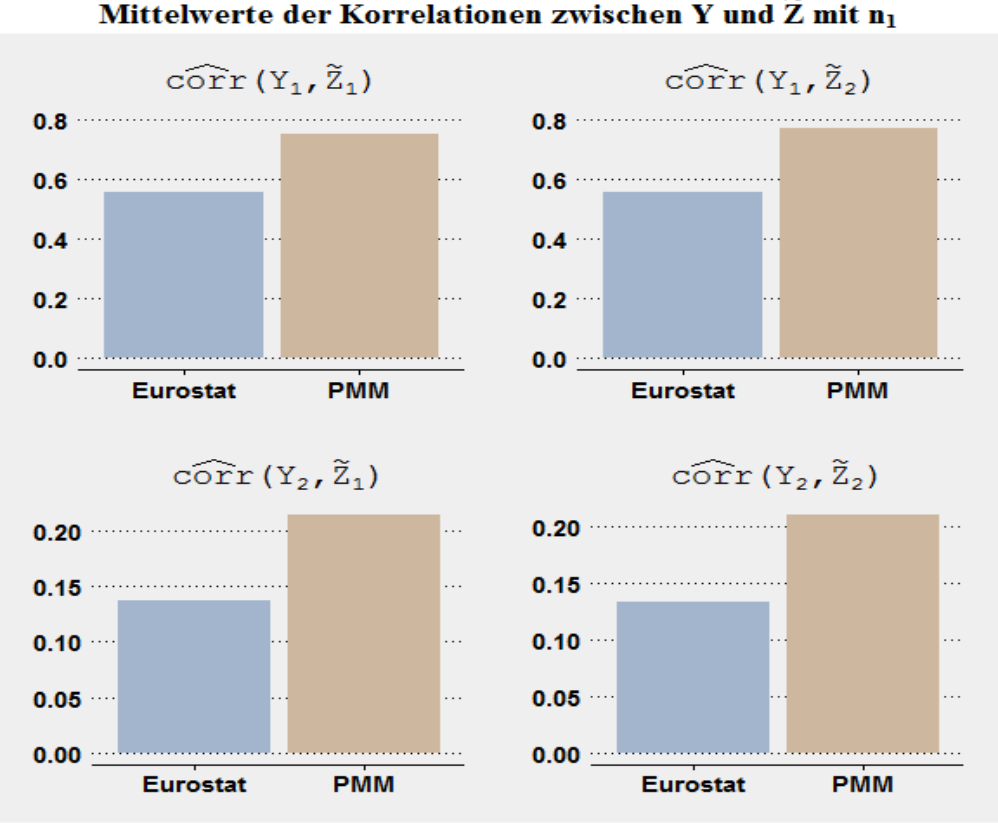

Abbildung 7: Barplots – Mittelwerte für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

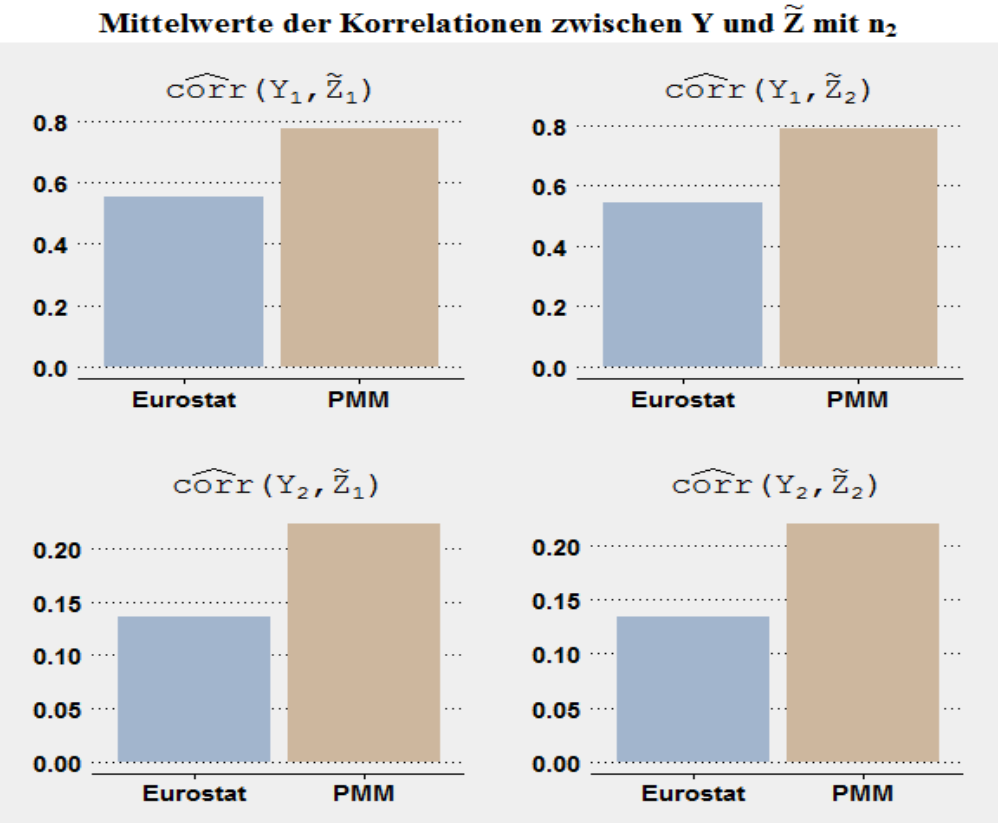

Abbildung 8: Barplots – Mittelwerte für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

Hinsichtlich der damit bereits angesprochenen Mittelwerte, die für n_1 und n_2 in den Abbildungen 7 und 8 veranschaulicht sind, lassen sich nochmals die bisher deskriptiv diskutierten, über alle $k = 1000$ Simulationsrunden aggregierten Implikationen der MC-Studie für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ zusammenfassen: Hohe Originalkorrelationen können von PMM offenbar deutlich besser reproduziert werden, als mit dem Random Hot-Deck von Eurostat. Die Mittelwerte der Korrelationsschätzer beziffern sich bei PMM unter n_1 für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ auf 0.75, für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ auf 0.77 und liegen damit im Vergleich zu Eurostat um eine Differenz von $0.75 - 0.55 = 0.20$ beziehungsweise $0.77 - 0.56 = 0.21$ näher an den wahren Parameterwerten. Unter n_2 beträgt die Differenz $0.77 - 0.55 = 0.22$ für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ beziehungsweise $0.79 - 0.55 = 0.24$ für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$. Bei mittleren Originalkorrelationen zeigt sich hingegen ein geringerer Unterschied zwischen dem Random Hot-Deck von Eurostat und PMM. Dennoch liefert PMM unter n_1 für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ im Mittel Korrelationsschätzer von jeweils 0.21, unter n_2 von jeweils 0.22. Damit liegt PMM mit Blick auf n_1 um eine Differenz von $0.21 - 0.14 = 0.07$ für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $0.21 - 0.13 = 0.08$ für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$, bezüglich n_2 um eine Differenz von $0.22 - 0.14 = 0.08$ für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $0.22 - 0.13 = 0.09$ für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ näher an den wahren Parameterwerten, als das Random Hot-Deck von Eurostat. Insofern können diese, zunächst deskriptiv betrachteten Ergebnisse der MC-Studie die zugrundeliegende Arbeitshypothese, die unterstellt, dass PMM ein besseres Fusionsergebnis produziert als Eurostat, stützen.

5.1.2 Monte-Carlo-Varianzen, Bias, MSE

Neben diesen deskriptiven Befunden legen aber bereits die diskutierten Kerndichteplots nahe, dass die präzisere Charakterisierung der MC-Verteilungen für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ noch tiefere Schlussfolgerungen und Implikationen über die Performance von Eurostat und PMM zulassen. Daher werden hier mit der Monte-Carlo-Varianz, dem Bias sowie dem Mean Squared Error (MSE) drei konkrete Evaluationskriterien diskutiert, anhand derer die deskriptiven Befunde noch genauer beleuchtet und beurteilt werden können. Die exakten Werte der in diesem Abschnitt referierten Diagnostiken sind in den Tabellen 8 bis 10 im Anhang B zu finden.

Im Rahmen der Charakterisierung der MC-Verteilungen mithilfe der Kerndichteplots wurde bereits thematisiert, dass die Korrelationsschätzer beider Datenfusionsverfahren mit Blick auf unterschiedliche Originalkorrelationen verschiedene Varianzen aufweisen. Dies ist in den Abbildungen 9 und 10 für die Stichprobenumfänge n_1 und n_2 konkret erkennbar, worin für

beide Fusionsmethoden die spezifischen Monte-Carlo-Varianzen[13], also die durchschnittlichen quadratischen Abweichungen vom Mittelwert über alle $k = 1000$ MC-Simulationen abgebildet sind. Die MC-Varianzen beim Random Hot-Deck-Verfahren von Eurostat sind für die hohen Originalkorrelationen bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ etwas höher als für die mittleren, wahren Zusammenhänge bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$, jedoch mit Blick auf ein variierendes Rezipienten- und Donorenverhältnis relativ konstant – zwischen den Werten für n_1 und n_2 zeigen sich bei Eurostat jeweils kaum Unterschiede. Für PMM hingegen ist bei beiden Stichprobenumfängen – n_1 und n_2 – zu beobachten, dass mittlere Originalkorrelationen bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ offenbar eine spürbar höhere Varianz induzieren, als hohe, wahre Zusammenhänge zwischen den spezifischen Variablen $\mathbf{Y}$ und $\mathbf{Z}$. Unter einer übermäßigen Donorenanzahl bei n_2 verstärkt sich für PMM dieser Effekt gegenüber n_1 dahingehend, als dass die geschätzten Korrelationen für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ nun geringfügig stärker, die Korrelationsschätzer für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ hingegen spürbar stärker streuen. Insofern liegen die Eurostat-Varianzen bei den hohen Originalzusammenhängen für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ noch über den beiden PMM-Varianzen, wobei unter n_2, also unter einer übermäßigen Donorendominanz, der Unterschied aufgrund höherer Streuung bei PMM geringer wird. Mit Blick auf mittlere, tatsächliche Korrelationen für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ ist hingegen festzuhalten, dass PMM hier deutlich höhere Varianzen aufweist, als das Random Hot-Deck-Verfahren – unter n_2 ist der Unterschied zugunsten des Eurostat-Ansatzes noch höher, da die PMM-Varianzen weiter ansteigen.

[13] $\text{V}_{\text{MC}}(\widehat{\rho}) = \frac{1}{k-1} \sum_{i=1}^{k} (\widehat{\rho}_i - \text{E}(\widehat{\rho}))^2$ mit $\text{E}(\widehat{\rho}) = \frac{1}{k} \sum_{i=1}^{k} \widehat{\rho}_i$

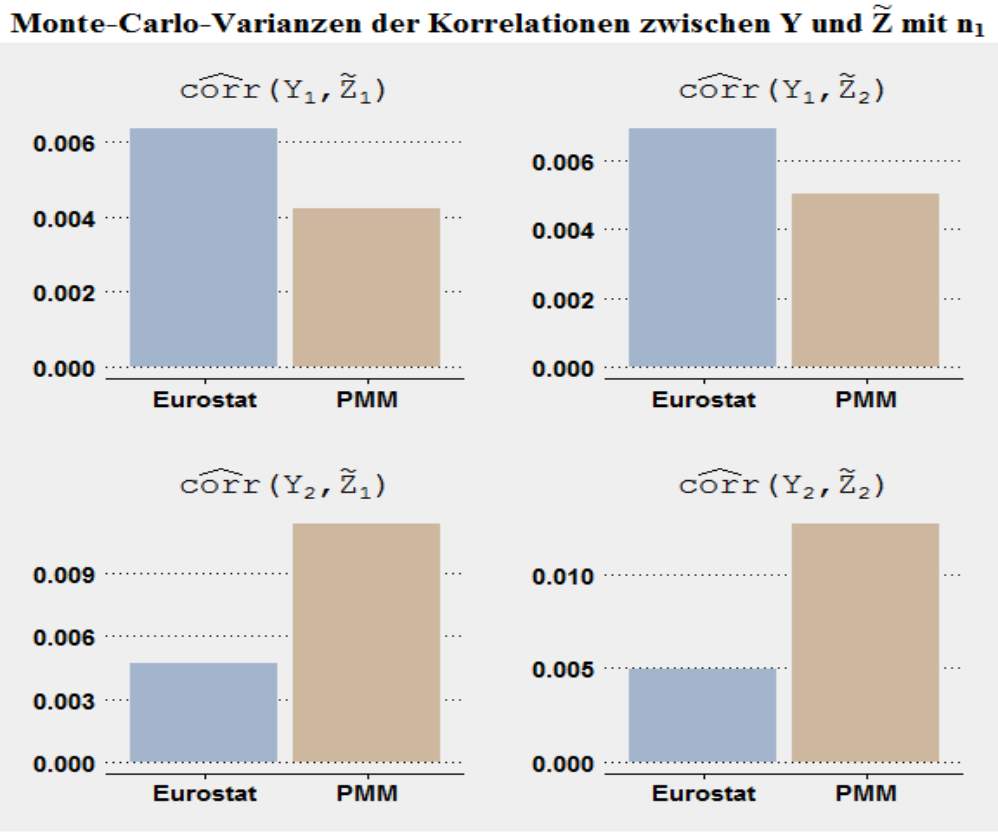

Abbildung 9: Barplots – MC-Varianzen für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

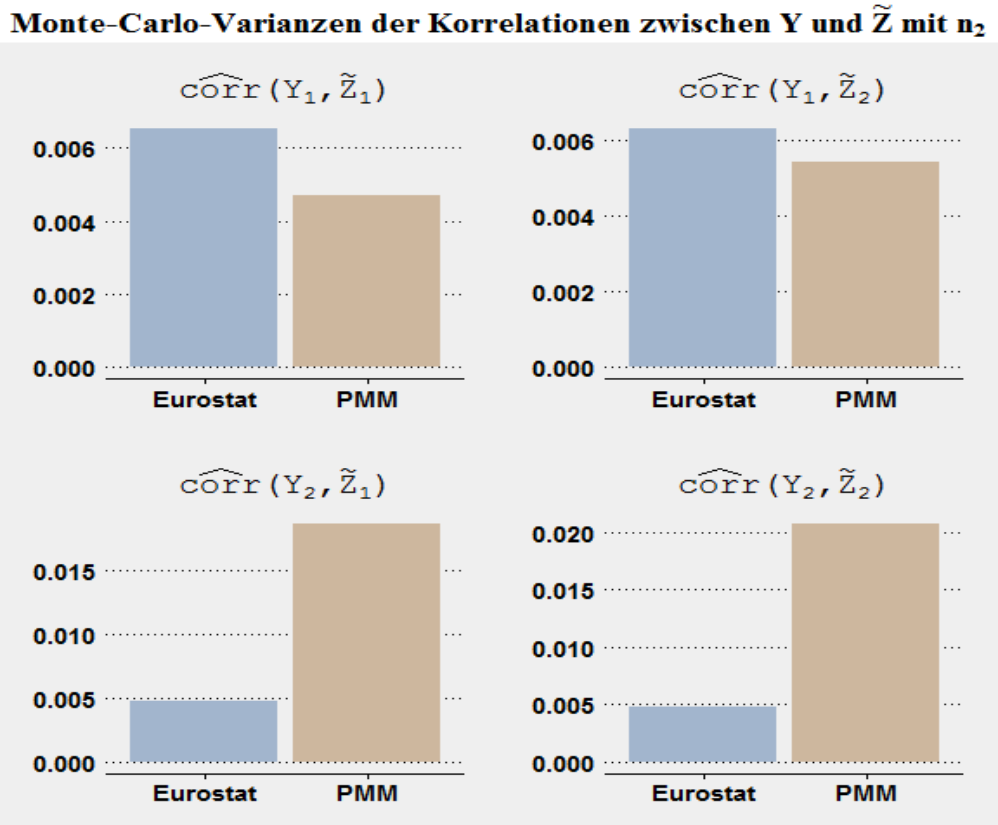

Abbildung 10: Barplots – MC-Varianzen für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

Eine hohe MC-Varianz ist nicht unbedingt wünschenswert, da sich damit eine höhere Unsicherheit verbindet. Über die konkrete Abdeckung der wahren Korrelationen sagt die MC-Varianz hingegen nichts aus, weshalb hierzu der Bias[14] miteinbezogen werden sollte. Dieser stellt die gemittelten, einfachen Abweichungen der $k = 1000$ geschätzten Korrelationen vom wahren Parameterwert dar, welche in den Abbildungen 11 und 12 grafisch veranschaulicht sind. Der Bias liegt bei PMM stets unter 0.1, sowohl für n_1, als auch für n_2. Dies bedeutet, dass die Korrelationsschätzer von PMM durchschnittlich höchstens um eine Differenz von 0.1 von den wahren Parameterwerten abweichen, was unter diesem Aspekt auf eine relativ ordentliche Fusionsperformance hindeutet. Eurostat hingegen verfehlt die wahren Zusammenhänge stärker. Ein deutlicher Unterschied ergibt sich für hohe Originalkorrelationen, also bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, wo Eurostat unter n_1 die wahren Zusammenhänge durchschnittlich um eine Differenz von 0.24 beziehungsweise 0.31 verfehlt – $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ ist unter n_2 bei Eurostat um 0.32 verzerrt. Aber auch bei mittleren, wahren Korrelationen der Grundgesamtheit ist für Eurostat unter n_1 und n_2 eine Verzerrung von 0.12 bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und 0.15 bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ zu beobachten, wodurch sich die Fusionsperformance im Vergleich zu hohen Korrelationen etwas verbessert, aber dennoch weiterhin PMM unterlegen bleibt. Die Parameterwerte von Eurostat sind für alle vier betrachteten Korrelationen unter n_1 und n_2 oft um ein vielfaches, jedoch immer mindestens doppelt so stark verzerrt wie die Korrelationsschätzer von PMM.

Damit ergibt sich hinsichtlich der MC-Varianzen und des Bias' ein teilweise konträres Bild: Während PMM immer deutlich unverzerrtere Korrelationsschätzer liefert, ist die Varianz für die mittleren Originalkorrelationen bei PMM höher, als bei Eurostat – unter einer starken Donorenüberzahl bei n_2 steigt für PMM die Varianz weiter an. Somit erscheint es sinnvoll, ein Performancekriterium zu betrachten, welches den Bias und die MC-Varianzen gleichermaßen inkludiert. Dies gewährleistet der Mean Squared Error (MSE)[15], der in den Abbildungen 13 und 14 dargestellt ist. Der MSE bringt die eben diskutierten Diagnostiken dahingehend zusammen, als dass er die Summe der (einfachen) Varianz[16] und des quadrierten Bias betrachtet. Inhaltlich spiegelt der MSE die über alle $k = 1000$ MC-Simulationsrunden gemittelten, quadratischen Abweichungen der geschätzten Korrelationen $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ von den wahren Korrelationen $\rho_{\mathbf{YZ}}$ wider.

[14] $\text{Bias}(\widehat{\rho}) = \text{E}(\widehat{\rho} - \rho) = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\rho}_i - \rho)$

[15] $\text{MSE}(\widehat{\rho}) = \text{E}\left[(\widehat{\rho} - \rho)^2\right] = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\rho}_i - \rho)^2 = \text{V}(\widehat{\rho}) + [\text{Bias}(\widehat{\rho})]^2$

[16] Hier wurde aus Präzisionsgründen die Monte-Carlo-Varianz verwendet, welche die quadrierten Abweichungen vom Mittelwert mit der Stichprobenkorrektur beziehungsweise der Monte-Carlo-Korrektur $\frac{1}{k-1}$ multipliziert, statt, wie bei der einfachen Varianz, mit $\frac{1}{k}$. Daher ist es möglich, dass die MSE-Werte geringfügig von der Summe zwischen MC-Varianzen und quadriertem Bias abweichen.

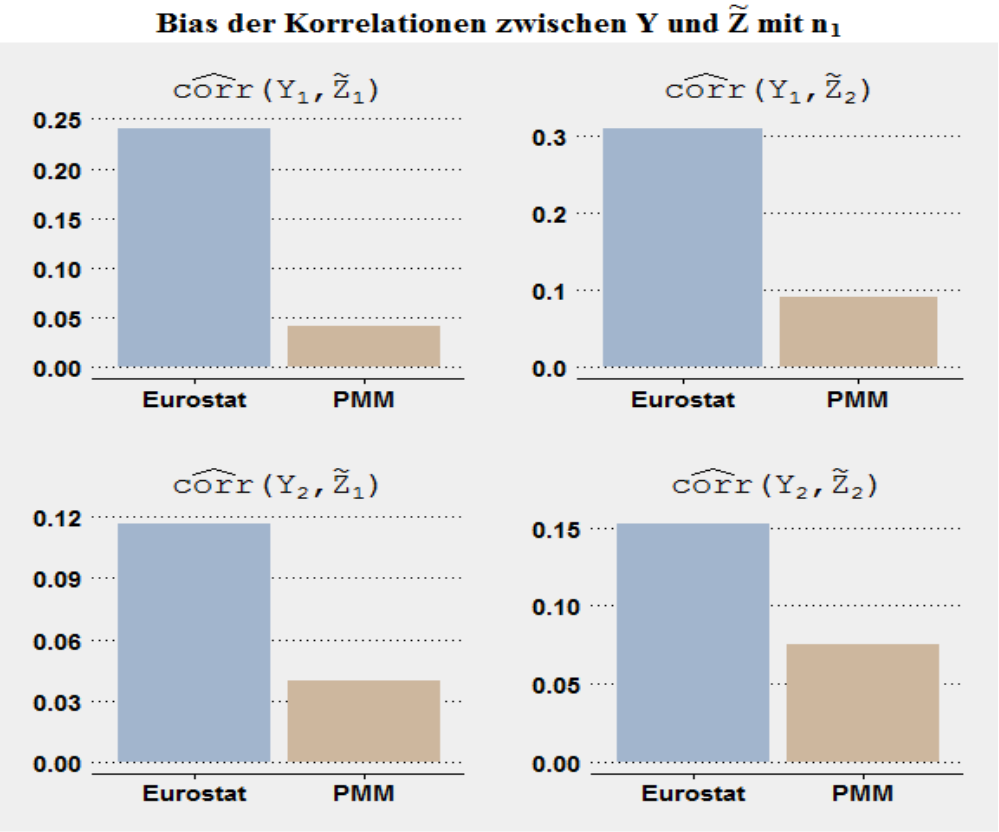

Abbildung 11: Barplots – Bias für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

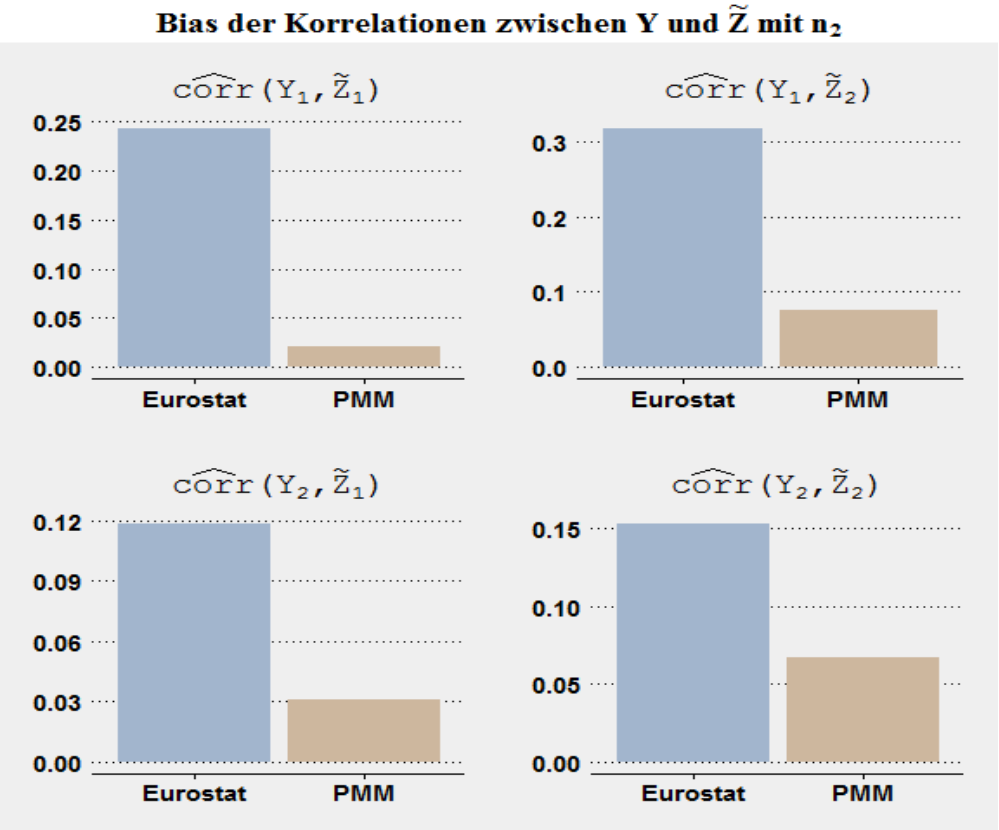

Abbildung 12: Barplots – Bias für $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

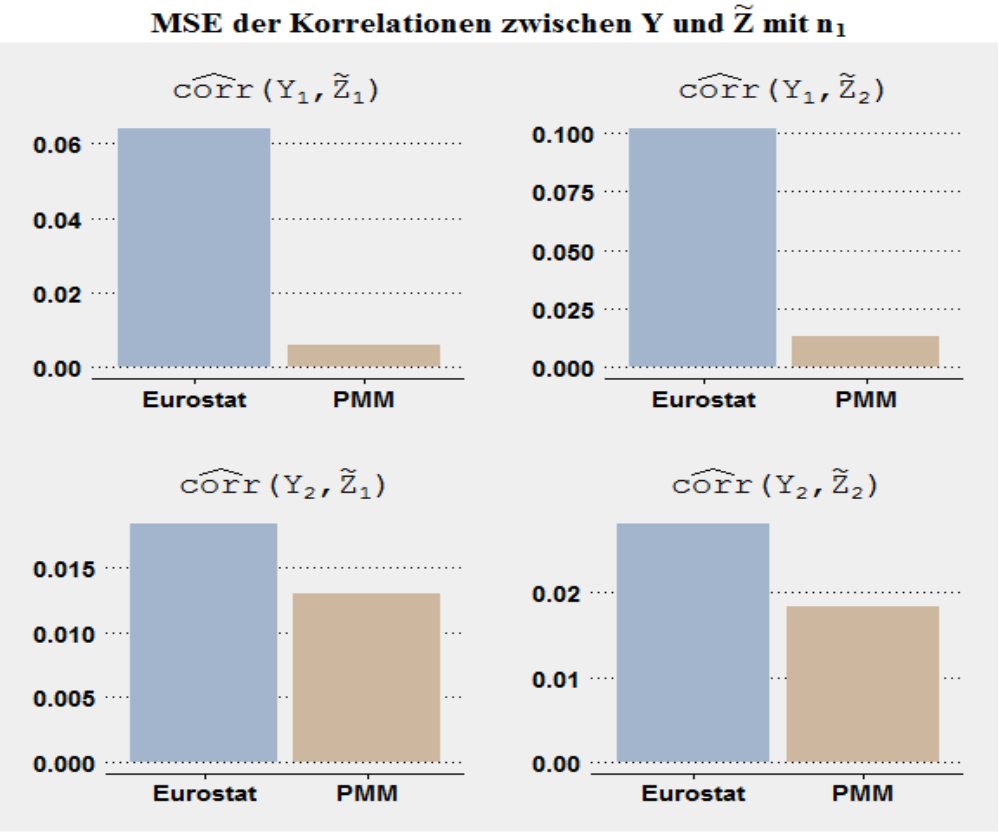

Abbildung 13: Barplots – MSE für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

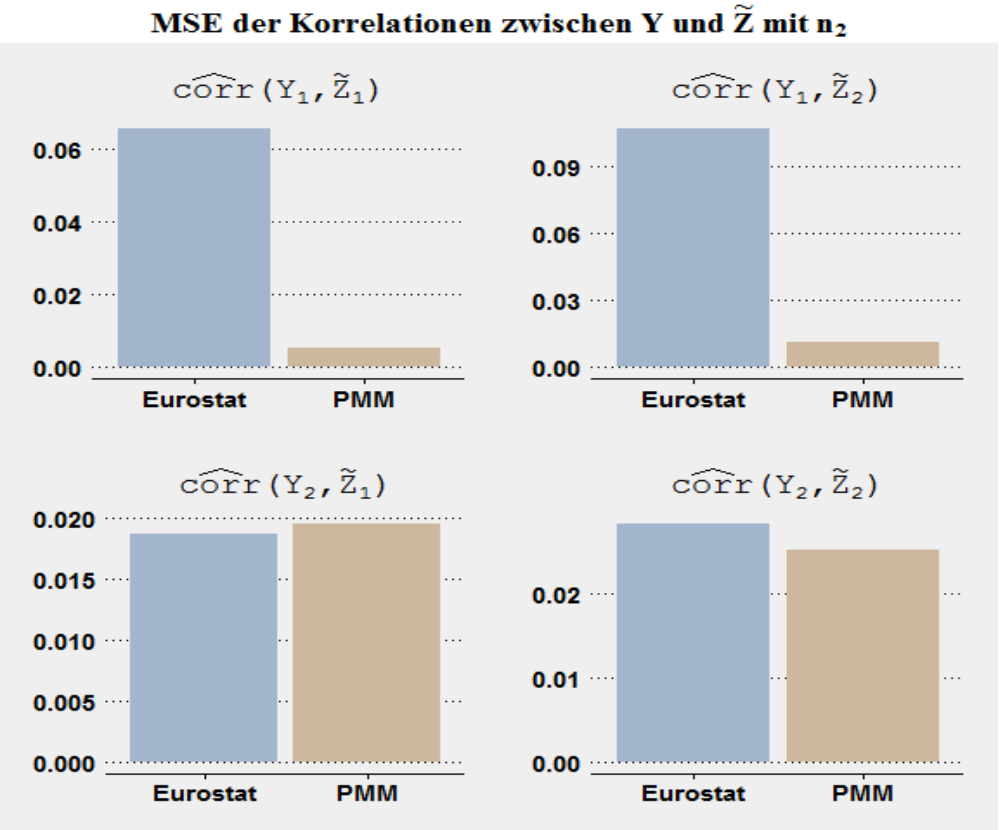

Abbildung 14: Barplots – MSE für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

Unter einem gleichmäßigen Rezipienten- und Donorenverhältnis (n_1) ist hierbei für die hohen Originalkorrelationen bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ festzustellen, dass der durch PMM induzierte mittlere quadratische Fehler deutlich geringer ist, als beim Random Hot-Deck von Eurostat, was entlang der vorigen Befunde, die sowohl einen niedrigeren Bias, als auch eine niedrigere MC-Varianz für PMM ergaben, wenig überraschend ist. Mit Blick auf eine übermäßig hohe Donorendominanz in n_2 wurde in Kapitel 5.1.1 bezüglich der hohen Originalkorrelationen für $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ angeschnitten, dass in deskriptiver Betrachtungsweise PMM von einem übermäßigen Donorenverhältnis geringfügig profitieren könnte. Die Ergebnisse dieser weiterführenden Diagnostiken zeigen zwar, dass der PMM-bezogene Bias unter n_2 bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ etwas zurückgeht, sich jedoch der zunächst sichtbare Performancevorteil durch eine geringfügig höhere MC-Varianz wieder etwas neutralisiert. Demzufolge verbessert sich der MSE bei PMM bezüglich hoher Originalkorrelationen unter n_2 (0.0051 beziehungsweise 0.0110) nur sehr marginal gegenüber n_1 (0.0059 beziehungsweise 0.0132) – stichhaltige Indizien für eine verbesserte PMM-Performance unter einer hohen Donorenüberzahl können daraus nicht abgeleitet werden. Mit Blick auf den konkreten Vergleich zwischen dem Random Hot-Deck von Eurostat und PMM ist aber unter hohen Originalzusammenhängen relativ eindeutig, dass PMM, auch entlang des MSE und somit sowohl bezogen auf die MC-Varianz, als auch hinsichtlich des Bias', ein besseres Fusionsergebnis produziert als Eurostat.

Für die mittleren Originalkorrelationen bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ liegen geringere Unterschiede zwischen Eurostat und PMM vor, wie die MC-Varianzen und der Bias sowie die Kerndichteplots bereits nahelegen. Unter n_2, also einer übermäßigen Donorenanzahl erhöht sich für PMM der MSE bei $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ im Vergleich zu selbigen, PMM-bezogenen Werten bei n_1, wobei für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ unter n_2 der MSE nun leicht über dem mittleren quadratischen Fehler von Eurostat liegt. Dies ist besonders das Resultat der bei PMM unter n_2 vorliegenden, höheren MC-Varianzen, die bereits in den Kerndichten in Abbildung 4 für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ beispielsweise dahingehend ersichtlich sind, als dass hier beide PMM-Kurven eine höhere Ausreißerquote nach oben, also Richtung 1, implizieren, als die selben PMM-Dichten in Abbildung 3 unter n_1. Bezüglich des Random Hot-Deck-Verfahrens von Eurostat ist wiederum erneut zu konstatieren, dass auch entlang des MSE die Fusionsperformance unter n_1 und n_2 nahezu konstant bleibt, also eine übermäßige Donorenanzahl bei Eurostat weder zu einer spürbaren Verbesserung, noch zu einer Verschlechterung führt. PMM hingegen muss mit Blick auf den MSE unter einer starken Donorenüberzahl in n_2 ein Performanceverlust aufgrund deutlich höherer MC-Varianzen,

trotz leichter Verbesserungen beim Bias, hinnehmen, weshalb die MSE-Werte bezüglich $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ für Eurostat und PMM unter n_2 relativ ähnlich sind. Insofern ist bei der Implementierung von PMM zu beachten, dass für mittlere Originalkorrelationen, die mit einer starken Donorenüberzahl einhergeht, die Fusionsperformance unter einer erhöhten Varianz und Unsicherheit leidet. Unter einem gleichmäßigen Rezipienten- und Donorenverhältnis kann PMM hingegen auch mit Blick auf den MSE, der die teilweise gegenläufigen Kriterien der MC-Varianz und des Bias' vereint, das Eurostat-Verfahren durch die bessere Reproduktion der Originalkorrelationen optimieren.

Zusammenfassend konnten somit die weiterführenden Diagnosekriterien – MC-Varianz, Bias und MSE – die deskriptiven Verteilungsergebnisse aus Kapitel 5.1.1 zusätzlich vertiefen: Hohe Originalkorrelationen kann PMM nicht nur deutlich besser reproduzieren, vielmehr induziert PMM, im Vergleich zum Random Hot-Deck von Eurostat, auch geringere MC-Varianzen, eine deutlich niedrigere Verzerrung und, somit, einen optimaleren, weil geringeren MSE. Unter n_2 steigen jedoch die Varianzen für PMM etwas an, der Bias hingegen reduziert sich geringfügig, weshalb der MSE für PMM bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ relativ ähnlich zu n_1 ist. Bei mittleren Originalkorrelationen weist PMM eine höhere MC-Varianz, aber dennoch einen geringeren Bias als das Eurostat-Verfahren auf. Hier optimiert PMM unter n_1, also einem gleichmäßigen Rezipienten- und Donorenverhältnis, das derzeitige Eurostat-Fusionsverfahren auch entlang des MSE, welcher MC-Varianz und Bias zusammenbringt. Für eine übermäßig hohe Anzahl an Donoren ist hingegen ebenfalls ein geringerer Bias bei PMM zu verorten, der jedoch mit einer spürbar höheren MC-Varianz einhergeht, was durch den MSE dahingehend bestraft wird, als dass die MSE-Werte für $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ unter n_2 ähnlich denen zu Eurostat sind. In dieser Hinsicht stützen also auch die hier diskutierten Diagnosekriterien weitgehend die in Kapitel 3.4 formulierte Arbeitshypothese, wobei zu beachten ist, dass PMM bei mittleren Originalkorrelationen unter einer erhöhten Varianz, insbesondere bei einer im Vergleich zur Rezipientenstichprobe übermäßig großen Donorenstichprobe, leidet.

Somit konnten in diesem Kapitel nun die Ergebnisse der MC-Simulation mit Blick auf die Reproduktion der wahren Korrelationen zwischen $\mathbf{Y}$ und $\mathbf{Z}$ veranschaulicht und entlang der deskriptiven Monte-Carlo-Verteilungen sowie konkreter Diagnosekriterien dargelegt werden. Dies dient der Evaluierung der Simulationsergebnisse entlang der dritten Validitätsstufe nach Rässler (2002: 30-31), die sich auf den Erhalt der Korrelationsstruktur der nicht gemeinsam beobachteten, spezifischen Variablen $\mathbf{Y}$ und $\mathbf{Z}$ bezieht. Neben der dritten Vali-

ditätsstufe ist die vierte Validitätsstufe nach Rässler (2002: 30-32), die den Erhalt der bereits im Donorendatenfile beobachteten Verteilungen betrifft, eine Art Mindestanforderung an eine Datenfusion. Inwiefern Random Hot-Deck und PMM dazu in der Lage sind, dieser Mindestanforderung nachzukommen, wird im folgenden Abschnitt thematisiert.

5.2 Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$

Um den Erhalt der bereits im Donorendatenfile beobachteten, gemeinsamen Verteilung von **X** und **Z** zu beurteilen, ist für empirische und statistisch-methodische Analysen insbesondere von Interesse, möglichst exakt die Zusammenhänge $\rho_{\mathbf{XZ}}$ im Fusionsdatenfile zu reproduzieren, da der Erhalt der Korrelationsstruktur elementar für inferenzstatistische Rückschlüsse ist. Daher soll in diesem Abschnitt zusätzlich dargestellt werden, inwiefern die beiden Fusionsverfahren, Random Hot-Deck und PMM, dazu imstande sind, die tatsächlichen Korrelationen zwischen den metrischen **X**-Variablen (X_2, X_7) und den spezifischen **Z**-Variablen zu erhalten. In Tabelle 4 sind die wahren Korrelationen $\rho_{\mathbf{XZ}}$ der aus $N = 25857$ Individuen bestehenden Ersatzpopulation tabelliert.

Wahre Korrelationen zwischen X und Z

$\text{corr}(X_2, Z_1)$	$\text{corr}(X_2, Z_2)$	$\text{corr}(X_7, Z_1)$	$\text{corr}(X_7, Z_2)$
-0.0136	-0.0085	0.9333	0.9547

Quelle: EU-SILC 2013 PUF: DE, FR, NL.

Tabelle 4: Wahre Werte für $\rho_{\mathbf{XZ}}$

Dabei ist zu erkennen, dass lediglich das Einkommen (X_7) mit den **Z**-Variablen hoch korreliert ist. Das Alter (X_2) weist wiederum einen Zusammenhang nahe 0 auf, was dem in Kapitel 4.2.1 erwähnten Umstand geschuldet ist, dass die Informationen des Personendatenfiles zufällig dem Haushaltsdatenfile zugewiesen wurden. Tatsächliche Korrelationen nahe 0 eignen sich jedoch für keinen Performancevergleich, da, wie in den Abbildungen 15 bis 18 ersichtlich ist, die Korrelationsschätzung eher einem Zufallsereignis gleicht, weshalb hier Random Hot-Deck und PMM aggregiert auch relativ ähnliche Ergebnisse erzielen. Insofern ist eine Bewertung und Betrachtung der geschätzten Korrelationen zwischen X_2 und $\widetilde{Z}_1$ sowie zwischen X_2 und $\widetilde{Z}_2$ wenig vielversprechend. Daher liegt das Augenmerk auf der Beurteilung von $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$, also den geschätzten Korrelationen zwischen X_7 und den beiden in EU-SILC hineinfusionierten, simulierten HBS-Variablen $\widetilde{Z}_1$ und $\widetilde{Z}_2$. Hierfür

werden, analog zu Kapitel 5.1, der Anschaulichkeit wegen besonders grafische Abbildungen diskutiert, wobei der Vollständigkeit halber dennoch die geschätzten Korrelationen für $\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$ enthalten sind. Alle entsprechenden, exakten Werte, sowohl bezüglich der MC-Verteilungen, als auch mit Blick auf die MC-Varianz, den Bias sowie den MSE sind im Anhang C in den Tabellen 11 bis 15 zu finden.

Abbildungen 15 und 16 zeigen die Verteilung der jeweiligen Korrelationsschätzer für alle $k = 1000$ MC-Simulationsrunden unter n_1, also einem gleichmäßigen Rezipienten- und Donorenverhältnis mit $n_{1_{silc}} = n_{1_{hbs}} = 300$, sowie unter n_2, in der den $n_{2_{silc}} = 300$ Rezipienten im simulierten EU-SILC Datenfile $n_{2_{hbs}} = 2700$ Donoren aus dem simulierten HBS-Datensatz gegenüberstehen. Unter einer gleichgroßen Rezipienten- und Donorenstichprobe bei n_1 zeigt sich, dass die geschätzten Korrelationen von PMM deutlich besser die wahren Zusammenhänge repräsentieren. Die überwiegende Mehrheit der Korrelationsschätzer von PMM befindet sich im näheren Bereich der wahren Korrelation, wie in den PMM-Dichten unter n_1 in Abbildung 15 gut zu erkennen ist. Eurostat liegt hingegen immer unter den tatsächlichen Zusammenhängen von 0.93 beziehungsweise 0.95, wobei das Maximum der beim Random Hot-Deck-Verfahren resultierenden Korrelationen für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ 0.84 und für $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ 0.87 beträgt. Damit sind selbst Ausreißer um eine Differenz von $0.93 - 0.84 = 0.09$ beziehungsweise $0.95 - 0.87 = 0.08$ vom wahren Parameterwert entfernt. Wie in den Boxplots in Abbildung 17 zu erkennen ist, liegen unter n_1 bei Eurostat die mittleren 50 % der geschätzten Korrelationen für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ zwischen 0.62 und 0.75, bei PMM hingegen für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ zwischen 0.90 und 0.94 sowie für $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ zwischen 0.92 und 0.97. Damit decken bei PMM die mittleren 50 % der Korrelationen den unmittelbaren Bereich der wahren Parameterwerte ab. Auch die Durchschnittswerte, die in Abbildung 19 dargestellt sind, betragen bei PMM unter n_1 für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ 0.91 und für $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ 0.93, was die wahren Korrelationen jeweils um eine Differenz von $0.93 - 0.91 = 0.95 - 0.93 = 0.02$ abbildet. Eurostat liefert hingegen unter n_1 sowohl für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$, als auch für $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ durchschnittliche Korrelationsschätzer von 0.68 und verfehlt damit die wahre Korrelation im Mittel um $0.93 - 0.68 = 0.25$ beziehungsweise $0.95 - 0.68 = 0.27$. Die Differenzwerte spiegeln im Übrigen die Verzerrung, also den Bias wider, der sich im Anhang C in den Abbildungen 23 und 24 wiederfindet.

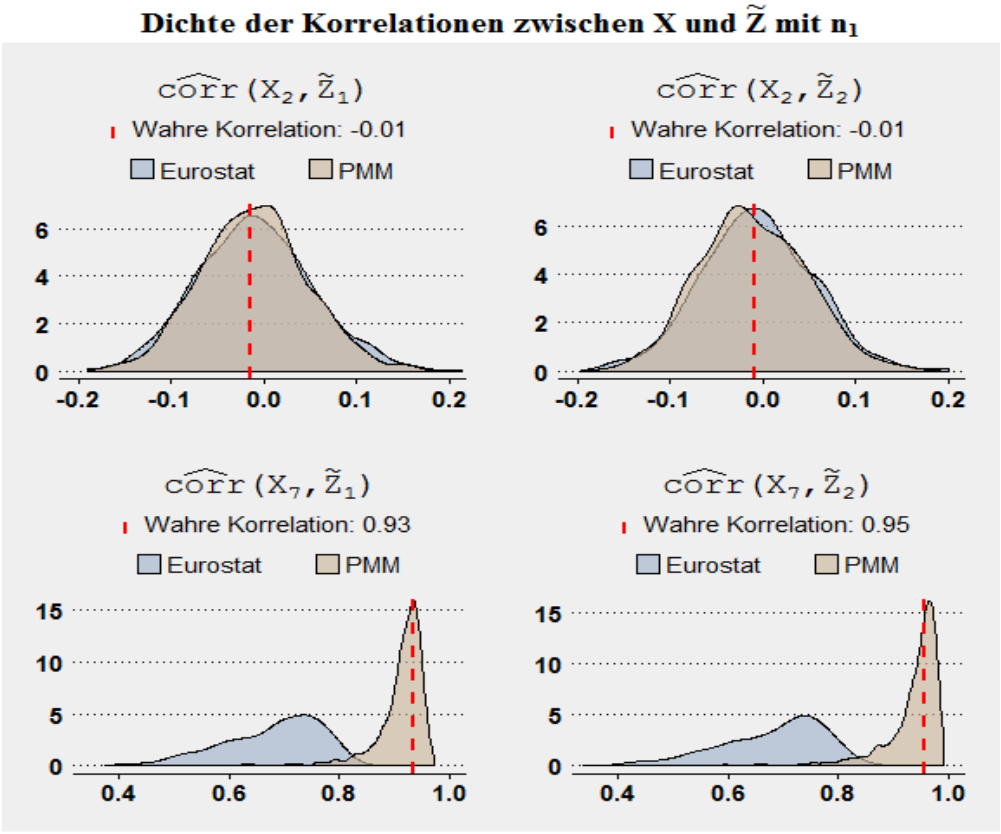

Abbildung 15: Kerndichteplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

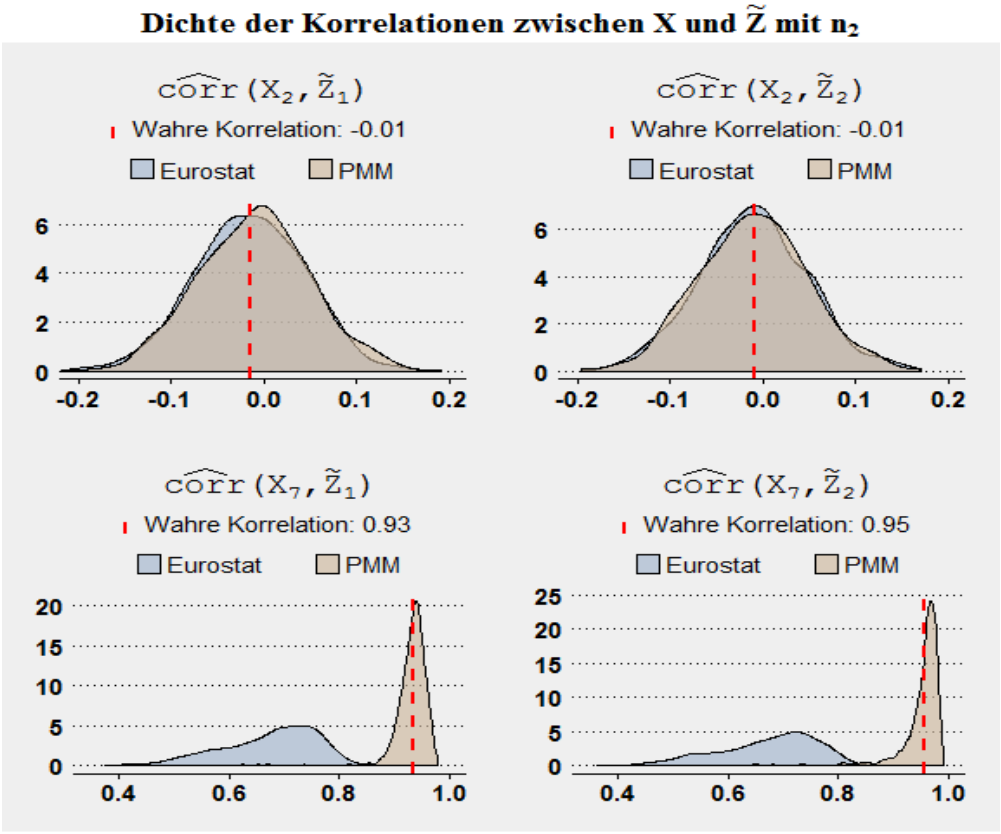

Abbildung 16: Kerndichteplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

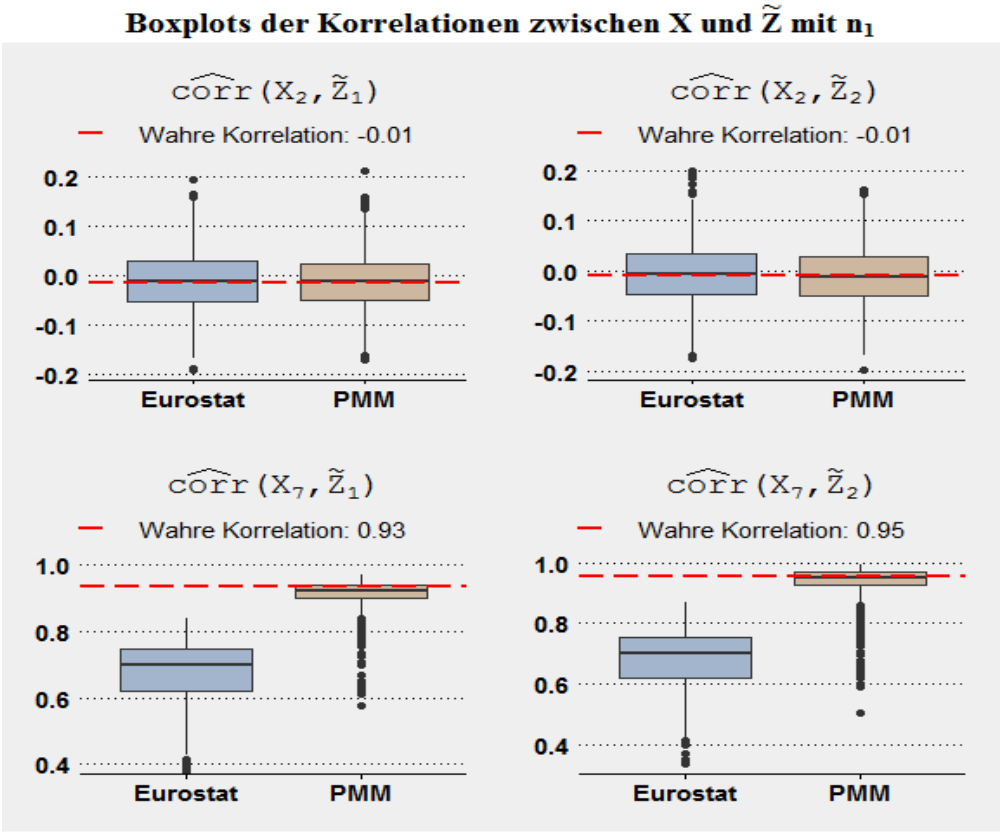

Abbildung 17: Boxplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

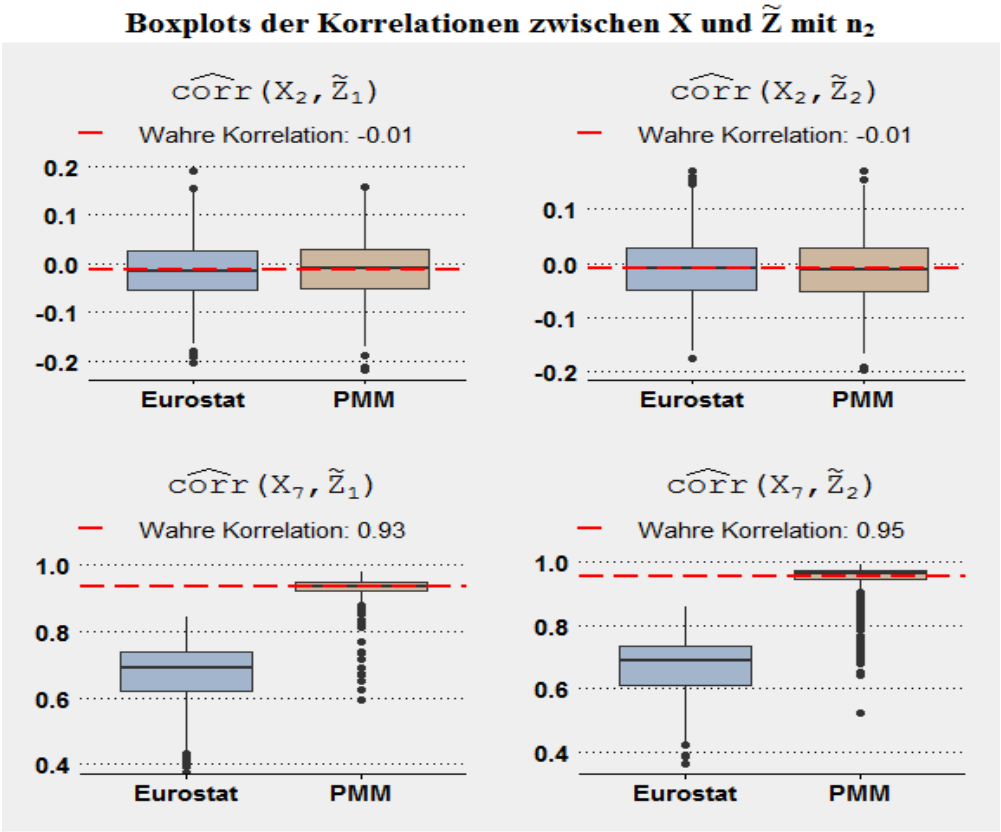

Abbildung 18: Boxplots – MC-Verteilungen für $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

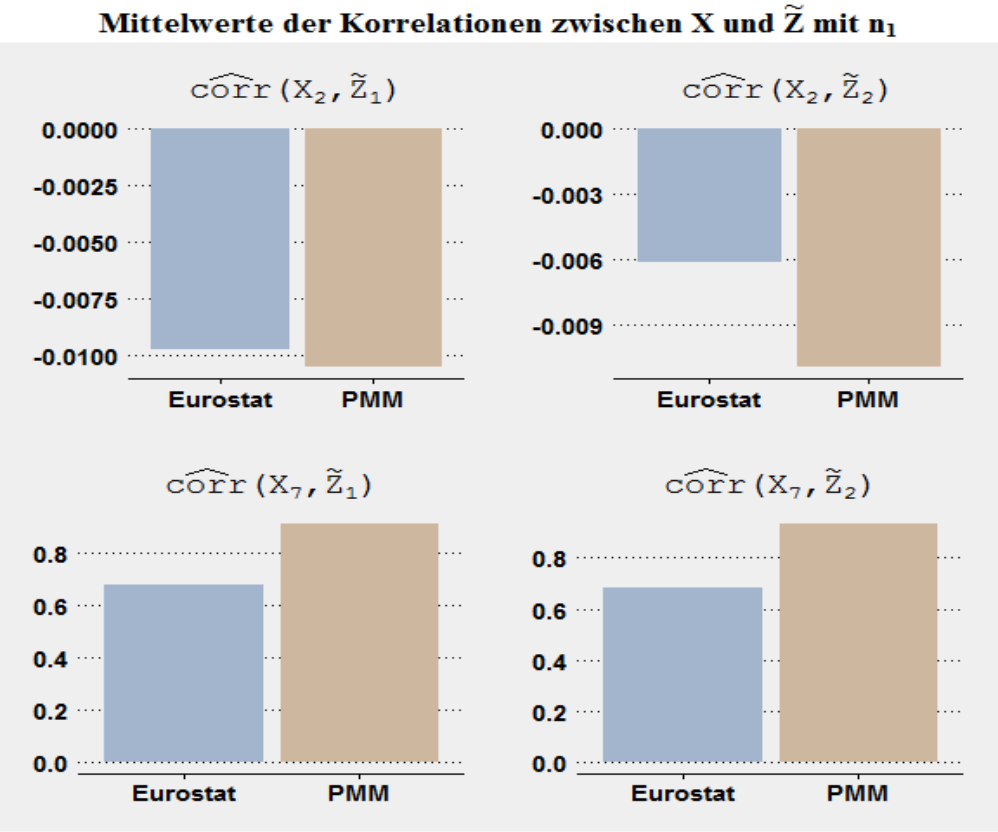

Abbildung 19: Barplots – Mittelwerte für $\widehat{\rho}_{\mathbf{X\tilde{Z}}}$ mit Stichprobenumfang n_1

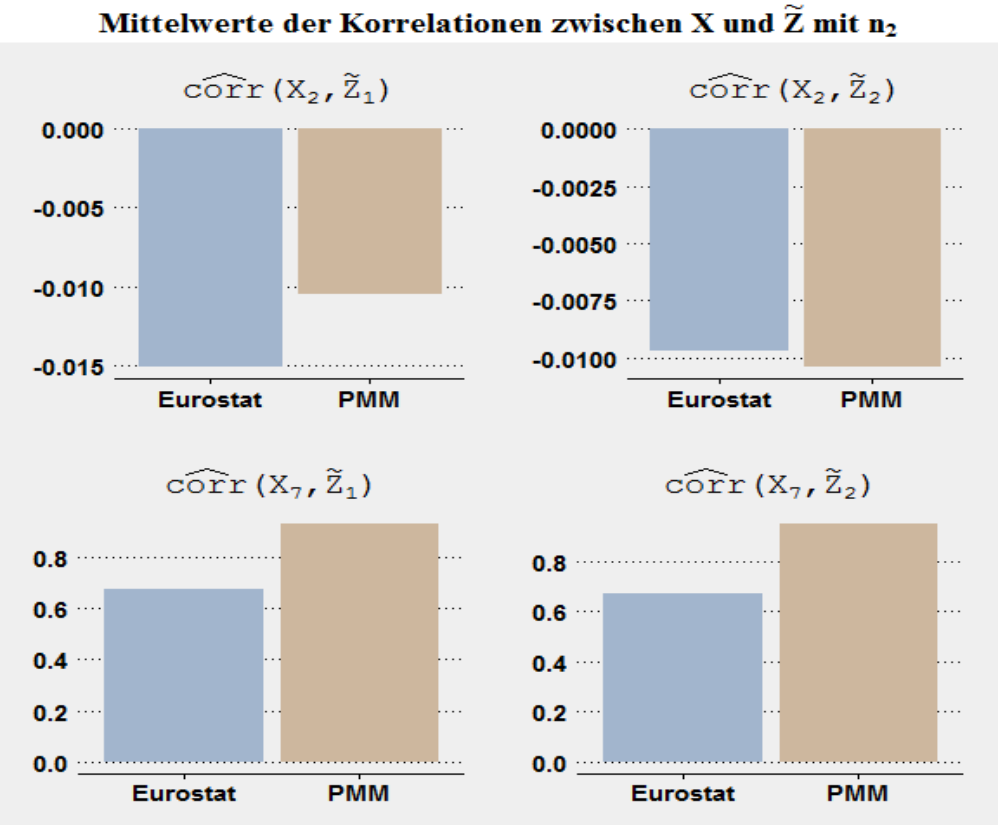

Abbildung 20: Barplots – Mittelwerte für $\widehat{\rho}_{\mathbf{X\tilde{Z}}}$ mit Stichprobenumfang n_2

Unter n_2, also einer im Vergleich zur Rezipientenstichprobe deutlich höheren Fallzahl bei der Donorenstichprobe, zeigen sich relativ ähnliche Ergebnisse, wie in den Abbildungen 16 und 18 ersichtlich ist. Bei PMM verschieben sich die Dichten (siehe Abb. 16) für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ etwas in Richtung höherer Korrelationen, wodurch allenfalls PMM konstatiert werden kann, dass die Zusammenhänge aggregiert betrachtet unter n_2 etwas höher sind, als unter einem gleichmäßigen Rezipienten- und Donorenverhältnis bei n_1. Auch die Boxplots in Abbildung 18 zeigen, dass unter n_2 bei $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ etwa der Median, also die Korrelation, die von 50 % der $k = 1000$ geschätzten Zusammenhänge nicht überschritten wird, nun bei $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ nahezu dem tatsächlichen Parameterwert von 0.93 entspricht. Bei $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ liegt der Median wiederum geringfügig über der wahren Korrelation von 0.95. Auch die Mittelwerte unter n_2, die in Abbildung 20 dargestellt sind, steigen im Vergleich zu n_1 von 0.91 auf 0.93 beziehungsweise von 0.93 auf 0.95 an, also um eine Differenz von 0.02.

Hinsichtlich der Charakterisierung der MC-Verteilungen für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ zeigen die Kurven unter n_1 und n_2 in den Abbildungen 15 und 16, dass die Eurostat-Werte bezüglich der geschätzten Korrelationen stärker streuen, als bei PMM und sie zudem eine Tendenz zur Linksschiefe aufweisen, weshalb bei Eurostat mehr Parameterschätzer über den entsprechenden, Eurostat-spezifischen Mittelwerten liegen, als darunter. Für PMM ergibt sich eine deutlich spitzere Kurve und somit eine geringere Streuung, aber ebenso mit leichter Tendenz zur Linksschiefe, wodurch auch bei PMM Korrelationsschätzer, die über den entsprechenden PMM-Mittelwerten liegen, etwas häufiger vorkommen, als Korrelationen unter dem arithmetischen Mittel.

Bezüglich der weiteren Diagnostiken der MC-Varianz, des Bias' und des MSE ergibt sich für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$, also für hohe Originalkorrelationen zwischen $\mathbf{X}$ und $\mathbf{Z}$ ein weitgehend einheitliches Bild. Daher werden die entsprechenden grafischen Darstellungen hier nur angeschnitten und sind konkret im Anhang C in den Abbildungen 21 bis 26 ebenso zu finden, wie die numerischen Werte, die sich in den Tabellen 13 bis 15 widerspiegeln. Dabei ist festzuhalten, dass für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ die PMM-Korrelationen, im Vergleich zu den mittels Random Hot-Deck geschätzten Zusammenhängen, sowohl deutlich niedrigere MC-Varianzen aufweisen, im Mittel einen viel geringeren Bias erzeugen und auch der mittlere quadratische Fehler (MSE) für PMM erheblich niedriger ist, als für Eurostat. Bei PMM ist überdies zu beobachten, dass sich alle drei Kriterien unter einem übermäßigen Donorenverhältnis verbessern. So ergeben sich bei PMM für $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ und $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ unter n_2 im Vergleich zu n_1 jeweils geringere MC-Varianzen, niedrigere Verzerrungen sowie gerin-

gere MSE-Werte. Allerdings ist hierbei zu beachten, dass diese PMM-Verbesserung unter n_2 nur für hohe Originalkorrelationen nachgewiesen werden konnte. Analog zu den mittelstarken Korrelationen zwischen $\mathbf{Y}$ und $\widetilde{\mathbf{Z}}$ könnte PMM jedoch auch für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ bei mittleren Originalzusammenhängen eine spürbar höhere Varianz aufweisen, was die Performance entlang des MSE entsprechend beeinträchtigen würde, sofern die Verbesserungen beim quadrierten Bias die Varianzeinbußen nicht ausgleichen könnten.

Zusammenfassend ergeben sich für hohe Originalkorrelationen zwischen den bereits im Donorendatenfile beobachteten Variablen $\mathbf{X}$ und $\mathbf{Z}$ relativ eindeutige Befunde dahingehend, dass PMM im fusionierten Datenfile die Korrelationen zwischen $\mathbf{X}$ und $\mathbf{Z}$ deutlich besser reproduzieren kann, als das Random Hot-Deck von Eurostat. Damit verbindet sich der Hinweis, dass PMM auch entlang der vierten Validitätsstufe einer Datenfusion, die eine Art Mindestanforderung darstellt, die Fusionsperformance von Eurostat optimiert und ein besseres Fusionsergebnis produziert, was auch mit Blick auf die Korrelation zwischen $\mathbf{X}$ und $\mathbf{Z}$ die zugrundeliegende Arbeitshypothese unterstützt. Allerdings ist dabei zu beachten, dass dies nur für hohe Originalkorrelationen (hier 0.93 beziehungsweise 0.95) nachgewiesen werden konnte. Wie sich der Performancevergleich für schwache oder mittlere Originalzusammenhänge zwischen $\mathbf{X}$ und $\mathbf{Z}$ darstellt, kann aufgrund der Datensituation hier nicht geklärt werden. Denkbar wäre, dass PMM, analog zu den Korrelationen zwischen den nicht gemeinsam beobachteten Variablen $\mathbf{Y}$ und $\mathbf{Z}$, auch bei mittleren Originalzusammenhängen zwischen $\mathbf{X}$ und $\mathbf{Z}$ eine höhere Varianz und Unsicherheit aufweist.

Abschließend bleibt festzuhalten, dass somit die Ergebnisse der MC-Simulation, sowohl bezüglich der Reproduktion der jeweils nicht gemeinsam beobachteten Variablen $\mathbf{Y}$ und $\mathbf{Z}$, als auch hinsichtlich des Erhalts der bereits im Donorendatenfile beobachteten Korrelationen zwischen $\mathbf{X}$ und $\mathbf{Z}$, dargeboten werden konnten, was einer Evaluierung entlang der dritten und vierten Validitätsstufe nach Rässler (2002: 29-32) entspricht. Um für das Datenfusionsvorhaben der amtlichen Statistik bereits eine Abschätzung dahingehend abzugeben, wie sensitiv beide Verfahren auf ein deutlich erhöhtes Rezipienten- und Donorenverhältnis zugunsten der Donorenstichprobe im Vergleich zu einem gleichgroßen Stichprobenumfang des Rezipienten- und Donorendatenfiles reagieren, wurde die Simulationsstudie unter beiden Bedingungen durchgeführt.

5.3 Bewertung und Diskussion

Doch wie sind die dargelegten Ergebnisse nun inhaltlich und mit Blick auf das konkrete Vorhaben der amtlichen Statistik, EU-SILC und HBS zu fusionieren, um die detaillierten Einkommensangaben aus EU-SILC sowie die umfangreichen Konsuminformationen aus dem HBS gemeinsam zu betrachten, zu bewerten? Dieser Abschnitt diskutiert die Implikationen der durchgeführten Simulationsstudie.

Zunächst ist entlang der theoretischen Ausführungen aus Kapitel 3.4 festzuhalten, dass von PMM eine präzisere Distanzmessung erwartet wird, da PMM keine Kategorisierung metrischer **X**-Variablen benötigt und damit einen Informationsverlust vermeidet sowie, im Besonderen, eine Abstufung zwischen gemeinsamen und für den Fusionsprozess relevanten Variablen $X_1, \ldots, X_a$ entlang ihres Einflusses auf die Erklärung der zu fusionierenden Variable Z_r vornimmt. Insbesondere aufgrund der Abstufung in den Einflüssen der Fusionsvariablen $X_1, \ldots, X_a$ sollte PMM eine bessere Fusionsperformance als das Random Hot-Deck-Verfahren von Eurostat aufweisen. Hierbei ist ersichtlich, dass PMM wohl nur dann ein besseres Fusionsergebnis produziert, wenn auch tatsächlich Unterschiede in den Zusammenhängen zwischen den gemeinsamen **X**-Variablen und den spezifischen, zu fusionierenden **Z**-Variablen vorliegen. Aus theoretischer Perspektive wäre daher zu erwarten, dass je unterschiedlicher die Zusammenhänge $\rho_{\mathbf{XZ}}$ sind, desto eher sollte PMM die Datenfusion von EU-SILC und HBS optimieren können. Im Prinzip stellt dies eine weitere Arbeitshypothese dar. Diese kann jedoch in vorliegender Arbeit nicht konsistent überprüft werden, da die Datensituation der EU-SILC PUFs von 2013 bereits durch die ausgewählten Variablen nahezu ausgeschöpft ist.

Dennoch kann über eine Betrachtung der Zusammenhänge zwischen den hier verwendeten gemeinsamen Variablen $\mathbf{X} = (X_1, \ldots, X_7)$ und den spezifischen HBS-Variablen $\mathbf{Z} = (Z_1, Z_2)$ das vorliegende Simulationsergebnis bewertet werden. Die interessierende Frage ist, wie stark sich die Einzelzusammenhänge zwischen **X** und **Z** voneinander unterscheiden. Eine grobe Einschätzung bietet bereits ein Blick auf die numerischen Korrelationen, die hier auch für die kategorialen **X**-Variablen $(X_1, X_3, X_4, X_5, X_6)$ errechnet wurden: Lediglich das Einkommen (X_7) ist, wie bereits deutlich wurde, mit Z_1 in einer Höhe von $\text{corr}(X_7, Z_1) = 0.93$, mit Z_2 in einer Intensität von $\text{corr}(X_7, Z_2) = 0.95$ hoch korreliert (EU-SILC 2013 PUF: DE, FR, NL). Alle anderen gemeinsamen Variablen $X_1, \ldots, X_6$ weisen mit beiden **Z**-Variablen eine Korrelation von nahezu 0 auf (EU-SILC 2013 PUF: DE, FR, NL). Da für Zusammenhänge zwischen metrischen und kategorialen Variablen, wozu das nominale so-

wie das ordinale Skalenniveau zählt, der Korrelationskoeffizient keine geeignete Maßzahl darstellt, wurde aus Validitätsgründen für den Zusammenhang zwischen den kategorialen **X**-Variablen $(X_1, X_3, X_4, X_5, X_6)$ und den **Z**-Variablen (Z_1, Z_2) auch Pearson's η^2 betrachtet, was ein Zusammenhangsmaß für nominale und metrische Merkmale darstellt. Auch hier sind die Zusammenhänge zwischen allen kategorialen **X**-Variablen sowie den spezifischen HBS-Variablen nahezu 0 (EU-SILC 2013 PUF: DE, FR, NL). Inhaltlich erscheint es zwar unlogisch, dass etwa das Alter (X_2) nicht mit den hier ausgewählten **Z**-Merkmalen, welche spezifische Einkommensvariablen darstellen, korrelieren. Dies ist jedoch dem in Kapitel 4.2.1 beschriebenen Umstand geschuldet, dass die Informationen aus dem Personendatenfile zufällig dem Haushaltsdatenfile zugewiesen wurden, weshalb eine Korrelation von nahezu 0 hier schlüssig und hinzunehmen ist.

Damit ist nun für den Fusionsprozess der Simulationsstudie festzuhalten, dass lediglich ein gemeinsames **X**-Merkmal, nämlich X_7, mit beiden **Z**-Variablen hoch korreliert. Für alle weiteren gemeinsamen Merkmale $X_1, \ldots, X_6$ ist der Zusammenhang mit Z_1 und Z_2 nahe 0. Von PMM wird aber gerade dann eine bessere Fusionsperformance erwartet, wenn es *Unterschiede* in den Einflüssen der **X**-Variablen auf die zu fusionierenden **Z**-Variablen gibt. Diese Unterschiede treten hier jedoch nur dahingehend auf, als dass sich lediglich *eine* gemeinsame Variable (X_7) von den Einflüssen der weiteren **X**-Variablen unterscheidet. Für sechs von sieben **X**-Variablen ergeben sich (nahezu) *keine Unterschiede* in den Einflüssen auf die zu fusionierenden **Z**-Variablen. Dieser Sachverhalt dürfte der Fusionsperformance des Random Hot-Deck-Verfahrens von Eurostat eher zugute kommen, als dem PMM-Algorithmus, der die auftretenden Zusammenhangsunterschiede zwischen **X** und **Z** berücksichtigt. Sofern diesbezüglich jedoch für sechs **X**-Variablen kaum Unterschiede in den Korrelationen mit **Z** festzustellen sind, wird der mit der Abstufung verbundene Vorteil von PMM etwas marginalisiert. Das Random Hot-Deck von Eurostat hingegen nimmt keine Abstufung der **X**-Variablen bezüglich ihres Einflusses auf die zu fusionierenden **Z**-Variablen vor, was aber für sechs **X**-Merkmale $(X_1, \ldots, X_6)$ auch kaum nötig erscheint. Damit sollte dem Random Hot-Deck von Eurostat durch die Beschaffenheit der hier verwendeten Daten aus theoretischer Perspektive ein Performancevorteil zukommen. Die Tatsache, dass beide Verfahren, Random Hot-Deck und PMM, vor dem eigentlichen Fusionsprozess mithilfe einer Backward Selection relevante **X**-Variablen auswählen, vermindert diesen Datenvorteil für Random Hot-Deck etwas. Innerhalb der a ausgewählten Variablen $X_1, \ldots, X_a$ ist dann aber dennoch, sofern X_7 darin enthalten und $a > 2$ ist, was in der durchgeführten Simulationsstudie durchaus, besonders unter n_2, vorkommt, nur X_7 mit höherem Gewicht in den Fusionsprozess miteinzubeziehen

– für alle anderen $a-1$ ausgewählten Variablen ist eine Abstufung kaum relevant. Insofern dürfte die Datengrundlage, sofern diese überhaupt einen Performancevorteil für eines der beiden Verfahren induziert, tendenziell eher die Random Hot-Deck-Methode von Eurostat bevorzugen, als PMM.

Dementsprechend wäre mit Blick auf die zugrundeliegende Arbeitshypothese, dass PMM ein besseres Fusionsergebnis aufgrund einer präziseren Distanzmessung produziert (siehe Kap. 3.4) von einem eher konservativen Hypothesentest im Rahmen der vorgestellten Simulationsstudie, welche PMM eher etwas benachteiligt, auszugehen. Dies dürfte die vorliegenden Simulationsergebnisse zugunsten von PMM zusätzlich validieren, zumal es in der Realität wahrscheinlich ist, dass im Rahmen der tatsächlichen Datenfusion von EU-SILC und HBS die **X**-Variablen sehr wohl deutlich variierendere Zusammenhänge mit den spezifischen HBS-Variablen aufweisen müssten. Dass lediglich ein **X**-Merkmal, wie es in der zugrundeliegenden Datensituation der Fall ist, mit **Z** hoch korreliert, während alle weiteren Zusammenhänge nahe 0 sind, ist jedenfalls für die mit reellen Datenquellen durchzuführende Datenfusion unwahrscheinlich. Je mehr Variation in den Zusammenhängen vorliegt, desto besser müsste das Fusionsergebnis von PMM, zumindest aus theoretischer Perspektive, ausfallen. Dies zu testen wäre aber beispielsweise ein Auftrag für die weitere Forschung.

Ferner konnte in der Simulationsstudie mit Blick auf die Zusammenhänge zwischen den jeweils nicht gemeinsam beobachteten Variablen **Y** und **Z** gezeigt werden, dass PMM hohe Originalkorrelationen deutlich besser reproduzieren kann, als Eurostat. Auch den mittleren Originalzusammenhängen zwischen **Y** und **Z** kommt PMM aggregiert betrachtet näher, als das Random Hot-Deck-Verfahren, wobei dies bei PMM, besonders unter einer übermäßigen Donorenanzahl, mit einer höheren Varianz und Unsicherheit verbunden ist. Ersichtlich ist jedoch, dass weder Eurostat, noch PMM im Mittel die tatsächlichen Korrelationen zwischen den jeweils nicht gemeinsam beobachteten Variablenmengen **Y** und **Z** erreichen können, was in den Abbildungen 7 und 8 sowie in Tabelle 7 im Anhang B gut zu erkennen ist – PMM kommt durchschnittlich lediglich bei $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ unter n_2 sehr nah an die tatsächliche Korrelation von 0.79 heran. Dies könnte dem in Kapitel 2.2 beschriebenen Dilemma geschuldet sein, dass herkömmlichen Datenfusionsalgorithmen, auch Random Hot-Deck und PMM, mit der Annahme der Conditional Independence Assumption (CIA) eine problematische weil häufig unrealistische und verletzte Annahme zugrunde liegt. In Kapitel 2.2 wurde erläutert, dass die CIA auch mit Blick auf die Datenfusionssituation von EU-SILC und HBS unzutreffend sein dürfte, zumal diese den MAR-Datenausfall beinhaltet, obwohl ein ignorierbarer

Datenausfallmechanismus unrealistisch erscheint. Insofern ist zu beachten, dass PMM zwar ein besseres, jedoch auch kein perfektes Fusionsergebnis erzeugt. Um die durch die (wahrscheinliche) Verletzung der CIA induzierte Fehlerquelle einzugrenzen, sei es ratsam, sofern möglich, Hilfsinformationen in den Fusionsprozess miteinzubeziehen, wie in Kapitel 2.2 kurz angeschnitten.

Neben der durch die CIA induzierten Fehlerquelle ist es denkbar, dass Datenfusionen zusätzlich darunter leiden, entlang der ausgewählten Variablen $X_1, \ldots, X_a$ (mit sehr hoher Wahrscheinlichkeit) keine perfekten Übereinstimmungen unter den Beobachtungseinheiten der Rezipienten- und Donorenstichprobe zu erhalten (Rodgers 1984). Dies könnte, neben der CIA, eine weitere Fehlerquelle für den Fusionsprozess darstellen (Rodgers 1984). Da Random Hot-Deck und PMM der Fehlerquelle der CIA unterliegen, PMM aber insgesamt ein weniger verzerrtes Fusionsergebnis produziert, ist es denkbar, dass PMM den Fehler, der durch die Abwesenheit perfekter Matches resultiert, besser handhaben kann und diesen minimiert. Die konkreten Fehlerquellen einer Datenfusion und deren Beziehung zueinander sind jedoch noch nicht näher erforscht, weshalb die Ausführungen lediglich auf theoretischen Überlegungen beruhen und daher mit Vorsicht zu genießen sind. Für die weitere Forschung verbindet sich damit aber der Auftrag, etwa die zugrundeliegenden Fehlerquellen einer Datenfusion genauer zu untersuchen und zu vergleichen.

Grundsätzlich ist darüber hinaus für Simulationsstudien zu konstatieren, dass deren generelle Schwäche der geringe Verallgemeinerungsgrad ist. Die Ergebnisse sind nur auf die vorliegende Datengrundlage, die aus den EU-SILC PUFs für Deutschland, Frankreich und die Niederlande aus dem Jahre 2013 besteht, anwendbar. Dementsprechend muss die Annahme getroffen werden, dass die diskutierte und dargelegte Performance des Random Hot-Deck-Verfahrens sowie des PMM-Fusionsalgorithmus' auch auf andere Datensituationen übertragbar ist. Gerade in dieser Hinsicht erscheint es den Ergebnissen der vorliegenden Arbeit zuträglich, dass die zugrundeliegende Datensituation aufgrund der Beschaffenheit der Public Usefiles, wenn überhaupt, eher das Random Hot-Deck-Verfahren von Eurostat bevorzugt, als die PMM-Methode. Sofern die amtliche Statistik die Ergebnisse dieser Arbeit dennoch aus Validitätsgründen mit reelleren Daten, etwa mit Scientific Usefiles, replizieren möchte, ist dies mit dem mitgelieferten `R`-Code der vorliegenden Arbeit jederzeit und unkompliziert möglich.

5.4 Handlungsperspektiven für die amtliche Statistik

Welche Handlungsperspektiven könnte nun die amtliche Statistik aus den dargelegten und eben diskutierten Ergebnissen der vorliegenden MC-Simulationsstudie ableiten, um im Rahmen ihres Datenfusionsvorhabens der gemeinsamen Analyse der umfangreichen Einkommensvariablen aus EU-SILC sowie den detailliert erfassten Konsumvariablen aus dem HBS möglichst präzise nachzukommen? Aus theoretischer Perspektive wurde umfassend erläutert, dass PMM ein besseres Fusionsergebnis als die gegenwärtige, von Eurostat verwendete Random Hot-Deck-Methode gewährleisten sollte. Entlang der Ergebnisse der Simulationsstudie konnte dies für hohe Originalkorrelationen zwischen den jeweils nicht gemeinsam beobachteten, simulierten Einkommens- ($\mathbf{Y}$) und Konsumvariablen ($\mathbf{Z}$) relativ eindeutig bestätigt werden. Für mittlere Originalzusammenhänge zwischen $\mathbf{Y}$ und $\mathbf{Z}$ ist der Performanceunterschied geringer, wobei PMM dennoch eine präzisere Reproduktion der Korrelationen $\rho_{\mathbf{YZ}}$ liefert, als Eurostat, die bei PMM allerdings eine erhöhte Varianz und Unsicherheit, besonders unter einer starken Donorenüberzahl, zur Folge hat. Da PMM jedoch aggregiert die tatsächlichen Korrelationen zwischen $\mathbf{Y}$ und $\mathbf{Z}$ besser repräsentiert, als das gegenwärtige Eurostat-Verfahren, ist der amtlichen Statistik grundsätzlich eine Implementation und Anwendung von Predictive Mean Matching anzuraten.

Bei einer Implementation von PMM sollte sich die amtliche Statistik bei mittleren Originalkorrelationen zwischen $\mathbf{Y}$ und $\mathbf{Z}$ allerdings der erhöhten Varianz und Unsicherheit bewusst sein. Nun ist es natürlich so, dass die tatsächlichen Korrelationen zwischen den Einkommensvariablen ($\mathbf{Y}$) und den Konsuminformationen ($\mathbf{Z}$) unbekannt sind. Sofern der amtlichen Statistik Hilfsinformationen vorliegen, die grobe Hinweise über die gemeinsamen Verteilungen von Einkommen und Konsum liefern könnten, wäre es für eine erste Einschätzung der Varianzen ratsam, die entsprechenden Hilfsdaten zu betrachten. Je nachdem, welche Hinweise die Hilfsinformationen zur Höhe der Korrelationen zwischen Einkommen und Konsum liefern, wäre dann bei einer Implementation von PMM eine Abschätzung dahingehend möglich, wie unsicher das Fusionsergebnis ausfallen könnte. Es scheint, dass je stärkere Originalkorrelationen vorliegen, desto weniger Unsicherheit verbindet sich mit dem PMM-Fusionsergebnis. In dieser Hinsicht wäre für eine dem Fusionsproblem angemessene Varianzschätzung auch die Verknüpfung von PMM mit Multipler Imputation, wie in Rubin (1987) ausführlich und detailliert beschrieben, eine durchaus sinnvolle, zusätzliche Alternative, zumal sich damit insbesondere auch gehaltvollere, inferenzstatistische Rückschlüsse verbinden (Rubin 1987).

Da hohe Originalzusammenhänge, etwa von 0.8 oder 0.9, in empirischen Daten eher unwahrscheinlich sind und selten vorkommen, erscheint es bei einer Implementation von PMM im Zweifelsfall sinnvoll, auf ein eher gleichmäßiges Rezipienten- und Donorenverhältnis zu achten, da damit das offenbar vorhandene Risiko einer zu hohen Varianz sowie eines zu großen, mittleren quadratischen Fehlers (MSE) minimiert wird. Sofern die unbekannten Korrelationen dennoch unerwartet hoch sein sollten, wird bei PMM dadurch zwar die potentielle, durch ein übermäßiges Donorenverhältnis induzierte Performanceverbesserung unterdrückt, was aber im Zweifelsfall aufgrund der eher geringen PMM-Performanceunterschiede zwischen n_1 und n_2 bei hohen Originalzusammenhängen in Kauf genommen werden kann. Unklar bleibt hingegen, wie sich die Performance beider Verfahren mit Blick auf geringe, wahre Korrelationen der Grundgesamtheit, etwa von 0.1 oder 0.15, verhält. Die Ergebnisse der MC-Studie liefern den Hinweis, dass sich bei geringen Originalkorrelationen die Unterschiede zwischen Random Hot-Deck und PMM gegenüber starken und mittleren Originalzusammenhängen noch weiter verringern könnten.

Neben dem zentralen Interesse, die Korrelationsstruktur zwischen den jeweils nicht gemeinsam beobachteten Einkommensvariablen $\mathbf{Y}$ und den Konsummerkmalen $\mathbf{Z}$ möglichst exakt zu reproduzieren, stellt der Erhalt der bereits im Donorendatenfile beobachteten Zusammenhänge zwischen $\mathbf{X}$ und $\mathbf{Z}$, was der vierten Validitätsstufe nach Rässler (2002: 30-32) entspricht, eine Art Mindestanforderung an eine Datenfusion dar. Dieser kann, wie die Ergebnisse der MC-Simulation darlegen, PMM besser nachkommen, als das Random Hot-Deck von Eurostat, wobei hier zu beachten ist, dass eine Evaluation dessen lediglich für hohe Originalkorrelationen zwischen $\mathbf{X}$ und $\mathbf{Z}$ möglich war. Unabhängig davon, welches Datenfusionsverfahren die amtliche Statistik nun implementiert, ist es ratsam, bei der reell durchgeführten Datenfusion von EU-SILC und HBS zumindest diese vierte Validitätsstufe, die sich mit empirischen Daten überprüfen lässt, zu evaluieren, was immerhin eine grobe Einschätzung der Performance der angewandten Fusionsmethode ermöglichen könnte. Dabei ist jedoch zu beachten, dass stichhaltige Indizien über das primäre Ziel einer Datenfusion, die gemeinsame Verteilung der jeweils nicht gemeinsam beobachteten Variablen möglichst präzise zu reproduzieren, daraus nicht zu gewinnen sind.

Zur möglichst validen Schätzung der gemeinsamen Verteilung der Einkommensvariablen aus EU-SILC und der Konsuminformationen aus dem HBS wäre die Implementation von PMM ein sinnvoller, erster Schritt, der jedenfalls den gegenwärtigen Fusionsalgorithmus von Eurostat entlang der vorliegenden Simulationsergebnisse optimieren kann. Jedoch produziert

auch PMM, wie bereits diskutiert, im Mittel kein perfektes Fusionsergebnis, was wohl besonders der (wahrscheinlich) verletzten Annahme der bedingten Unabhängigkeit (CIA) geschuldet ist. Dementsprechend ist der amtlichen Statistik in einem weiteren Schritt zusätzlich nahezulegen, Hilfsinformationen, sofern vorhanden, in den konkreten Datenfusionsprozess miteinzubeziehen, wie etwa in D'Orazio et al. (2006: 65-95) beschrieben. Der *glue*-Ansatz von Fosdick et al. (2015) stellt hier ein neueres und die CIA abschwächendes Verfahren dar. Inwiefern derartige Fusionsszenarien zu einer weiteren Optimierung des Datenfusionsvorhabens der amtlichen Statistik beitragen, wäre beispielsweise ein Untersuchungsauftrag für künftige Forschungsstudien.

Unter programmtechnischen Aspekten kann PMM etwa über das `BaBooN`-Package von Meinfelder und Schnapp (2015) in `R` relativ unkompliziert umgesetzt werden, da das Package mit `BBPMM.row()` eine für Datenfusionen passgenaue Funktion beinhaltet. Auch ist hierbei eine simultane Datenfusion im multivariaten Fall, also mit mehr als einer spezifischen Konsumvariable aus dem HBS, möglich, da dies `BBPMM.row()` automatisch vornimmt. Die Programmierung vom Random Hot-Deck-Verfahren ist hingegen erstens relativ aufwendig und gleichzeitig wenig vielversprechend, da die Ergebnisse dieser Arbeit darauf schließen lassen, dass die Eurostat-Korrelationsschätzer selten auch nur den näheren Bereich der tatsächlichen Zusammenhänge abdecken können. Zweitens ist für jede zusätzliche, zu fusionierende Konsumvariable Z_r die Fusionsprogrammierung zu erweitern und anzupassen, was die Komplexität und Unübersichtlichkeit noch weiter erhöht. Im Rahmen dieser Arbeit konnte aber exemplarisch anhand des bivariaten Falls, also einer Ergänzung des EU-SILC-Datenfiles um *zwei* spezifische HBS-Variablen, gezeigt werden, dass eine Erweiterung der bisherigen, von Eurostat vorgenommenen Datenfusion, die lediglich den univariaten Fall, also die Ergänzung *einer* Konsumvariable betrachtet, um zusätzliche HBS-Konsumvariablen möglich ist. Während dies mit `BBPMM.row()` automatisiert erfolgen kann, muss beim Eurostat-Verfahren die Programmierung für jede zusätzliche, spezifische HBS-Variable erweitert und angepasst werden.

Damit konnten abschließend die Ergebnisse der durchgeführten Simulationsstudie nicht nur ausgiebig dargelegt, bewertet und diskutiert, sondern auch mit damit einhergehenden, konkreten Handlungsperspektiven für die amtliche Statistik verknüpft werden. Dadurch wurde eine umfassende Überprüfung der zugrundeliegenden Arbeitshypothese gewährleistet, die der konsistenten Beantwortung der Fragestellung dieser Arbeit dienen soll.

6 Zusammenfassung und Fazit

Der Untersuchungsauftrag der vorliegenden Arbeit spiegelte sich in der Frage wider, ob Predictive Mean Matching das von Eurostat verwendete Random Hot-Deck-Verfahren zur Datenfusion von EU-SILC und HBS optimieren kann. Die Relevanz dieses Untersuchungsauftrages begründet sich besonders in der Motivation der amtlichen Statistik in Europa, eine Analyse der gemeinsamen Verteilung von Einkommen und Konsumausgaben zu ermöglichen, um einerseits den sozialen und wirtschaftlichen Lebensstandard der privaten Haushalte in der EU, andererseits Armutsrisiken und Armutsindikatoren präziser messen und erfassen zu können. Für den Vergleich von Random Hot-Deck und Predictive Mean Matching wurde zunächst ein einleitender Überblick zu Datenfusionen, sowohl mit Blick auf definitorische Aspekte, als auch bezüglich der zentralen Annahme und der relevanten Validitätsstufen einer Datenfusion, gegeben. Anschließend erfolgte ein kurzer Abriss der bisher in der Forschung und bei Eurostat diskutierten Fusionsansätze, bevor Random Hot-Deck und Predictive Mean Matching detailliert erläutert und vergleichend diskutiert wurden, wobei es daraus abgeleitete Implikationen in eine Arbeitshypothese zu überführen galt. Sodann erfolgte eine Erläuterung der Datengrundlage sowie des Simulationsdesigns zur Überprüfung der Hypothese. Anschließend konnten die Ergebnisse der Monte-Carlo-Simulation, sowohl mit Blick auf die Korrelationen zwischen den nicht gemeinsam beobachteten Variablen **Y** und **Z**, als auch hinsichtlich der bereits im Donorendatenfile beobachteten Zusammenhänge zwischen **X** und **Z**, dargestellt werden. Diese wurden bewertet und diskutiert, woraus sich konkrete Handlungsperspektiven für die amtliche Statistik ableiten ließen.

Die Ergebnisse der MC-Simulation konnten darlegen, dass PMM die Korrelationsstruktur zwischen den substituierten Einkommensvariablen **Y** aus EU-SILC und den substituierten Konsumvariablen **Z** aus dem HBS besser reproduziert, als das Random Hot-Deck-Verfahren von Eurostat. Ein relativ hoher Performancevorteil zugunsten von PMM wurde für starke Originalzusammenhänge der Grundgesamtheit, hier in Höhe von 0.79 und 0.86, nachgewiesen. Die mittelstarken Korrelationen der Ersatzpopulation von 0.25 und 0.29 spiegeln sich bei PMM aggregiert betrachtet ebenfalls präziser im Fusionsdatenfile wider, wobei hier der Performanceunterschied zwischen PMM und Random Hot-Deck geringer ausfällt. Allerdings verbindet sich bei PMM damit der Makel, dass die Varianz für mittlere Zusammenhänge, im Vergleich zu hohen Originalkorrelationen, spürbar zunimmt, was eine erhöhte Unsicherheit zur Folge hat.

Für eine erste Sensitivitätseinschätzung wurde darüber hinaus die Simulationsstudie sowohl mit einem gleichmäßigen Rezipienten- und Donorenverhältnis, als auch mit einem deutlich erhöhten Verhältnis zugunsten der Donorenstichprobe durchgeführt. Während sich das Random Hot-Deck-Verfahren hinsichtlich derartig variierender Stichprobenverhältnisse als relativ robust erwies, scheint die Sensitivität bei PMM abhängig von den tatsächlichen, in der Grundgesamtheit vorhandenen Originalzusammenhängen zu sein: Bei hohen Korrelationen von 0.79 und 0.86 konnte bei PMM eine leichte, aber aufgrund von Zufallsschwankungen mit Vorsicht zu genießende Tendenz hin zu genaueren und damit weniger verzerrten Korrelationsschätzern beobachtet werden. Auch für mittelstarke Originalzusammenhänge von 0.25 und 0.29 scheint sich bei PMM die einfache Verzerrung, der Bias, bei einer starken Donorenüberzahl zu reduzieren, jedoch steigen die Varianzen dabei deutlich an und führen daher zu einem spürbar höheren mittleren quadratischen Fehler (MSE), als bei einem gleichmäßigen Rezipienten- und Donorenverhältnis. Um dieses Risiko zu verringern, wäre bei PMM ein eher ausgeglichenes Stichprobenverhältnis zwischen Rezipienten- und Donorendatenfile, sofern möglich, empfehlenswert, außer es lassen sehr zuverlässige Indizien, etwa aus Hilfsinformationen, darauf schließen, dass hohe Zusammenhänge zwischen den jeweils nicht gemeinsam beobachteten Variablen **Y** und **Z** vorliegen könnten. Aggregiert betrachtet sind die Korrelationsschätzer von PMM aber auch unter einem übermäßig hohen Rezipienten- und Donorenverhältnis zugunsten der Donorenstichprobe bei mittleren Originalzusammenhängen deutlich weniger verzerrt, als die mit dem Random Hot-Deck von Eurostat errechneten Korrelationen.

Hinsichtlich der Mindestanforderung an eine Datenfusion, gemäß derer sich die bereits im Donorendatenfile beobachtete Verteilung **X** und **Z** möglichst exakt im Fusionsdatenfile widerspiegeln sollte, zeigte sich für hohe Originalzusammenhänge zwischen **X** und **Z**, dass PMM diese deutlich besser reproduziert, als das Eurostat-Verfahren. Dementsprechend konnte sowohl mit Blick auf die Korrelationsstrukturen der jeweils nicht gemeinsam beobachteten Variablen **Y** und **Z**, als auch bezüglich der Reproduktion der Zusammenhänge zwischen den im Donorendatenfile beobachteten Variablen **X** und **Z** die zugrundeliegende Arbeitshypothese, die unterstellt, dass PMM aufgrund einer präziseren Distanzmessung ein besseres Fusionsergebnis produziert, weitestgehend bestätigt werden. Für mittlere Originalzusammenhänge zwischen den simulierten Einkommensvariablen **Y** und den simulierten Konsummerkmalen **Z** ist PMM jedoch eine Schwäche aus Varianz- und Unsicherheitsaspekten, besonders bei einer starken Donorenüberzahl, zu konstatieren.

Wie kann nun entlang der Ergebnisse der Simulationsstudie die Fragestellung beantwortet werden? Kann Predictive Mean Matching das von Eurostat verwendete Random Hot-Deck-Verfahren zur Datenfusion von EU-SILC und HBS optimieren? Eine Optimierung der Eurostat-Fusionsmethode kann durch Predictive Mean Matching dahingehend gewährleistet werden, als dass PMM die Zusammenhänge der gemeinsamen Verteilung von $\mathbf{Y}$ und $\mathbf{Z}$, also den Einkommens- und Konsumvariablen, besser im Fusionsdatenfile reproduzieren kann, was für inhaltliche und inferenzstatistische Analysen eine essentielle Voraussetzung darstellt, um etwa, gemäß des Vorhabens der amtlichen Statistik, der Erfassung des sozialen und wirtschaftlichen Lebensstandards der privaten Haushalte in der EU sowie der präziseren Messung von Armutsrisiken und Armutsindikatoren möglichst adäquat nachzukommen. Dementsprechend ist der amtlichen Statistik eine Implementation von PMM aus wissenschaftlicher Perspektive grundsätzlich zu empfehlen, da damit eine präzisere und unverzerrtere Datenfusion von EU-SILC und HBS einhergeht, als beim gegenwärtigen Random Hot-Deck-Verfahren von Eurostat, zumal mit dem `BaBooN`-Package von Meinfelder und Schnapp (2015) eine unkomplizierte Fusionsdurchführung möglich ist. Bei einer Anwendung von PMM gilt es jedoch, den PMM-spezifischen Makel, welcher sich in der erhöhten Varianz für mittelstarke, wahre Zusammenhänge der Grundgesamtheit begründet, in Betracht zu ziehen und diesen etwa mittels Hilfsinformationen im Voraus abzuschätzen. Auch die Anwendung von Multipler Imputation wäre diesbezüglich ein probates Mittel, zumal damit gehaltvollere, inferenzstatistische Rückschlüsse möglich sind. Da PMM zwar das Random Hot-Deck-Verfahren von Eurostat optimieren kann, jedoch auch kein perfektes Fusionsergebnis induziert, was wohl besonders den Problematiken mit der Annahme der bedingten Unabhängigkeit (CIA) geschuldet ist, stellt die Einbeziehung von Hilfsinformationen, sofern vorhanden, im Rahmen des konkreten Fusionsprozesses eine weitere, durchaus sinnvolle Handlungsoption für die amtliche Statistik dar.

Da das Random Hot-Deck von Eurostat ein nicht-parametrischer Fusionsansatz, Predictive Mean Matching hingegen ein semi-parametrisches Verfahren darstellt, konnte die vorliegende Arbeit auch nochmals die bisherigen, bei Eurostat diskutierten Fusionsmethoden aufgreifen und kompetitiv rezipieren. Die Studie von Webber und Tonkin (2013), welche nicht-parametrische, parametrische und semi-parametrische Verfahren vergleicht, kommt etwa zu dem Ergebnis, dass lediglich geringe Unterschiede zwischen den genannten Ansätzen vorliegen, jedoch den semi-parametrischen Verfahren ein leichter Performancevorteil zuzuschreiben ist (Webber und Tonkin 2013). Möglicherweise implementierte Eurostat mit dem Random Hot-Deck einen nicht-parametrischen Fusionsansatz gerade deshalb, weil entlang

der Resultate von Webber und Tonkin (2013) mit der Anwendung eines semi-parametrischen Verfahrens nur geringe Performanceverbesserungen einhergingen. Die Ergebnisse dieser Arbeit konnten jedoch für Predictive Mean Matching, einem semi-parametrischen Ansatz, durchaus auffallende Performancevorteile gegenüber dem nicht-parametrischen Random Hot-Deck-Verfahren nachweisen, wodurch eine Weiterverwendung der Random Hot-Deck-Methode entlang der Ergebnisse dieser Arbeit als wenig vielversprechend und kaum gerechtfertigt anzusehen ist.

Insofern konnte die vorliegende Arbeit auch die bei Eurostat geführte Diskussion rund um eine adäquate Datenfusion von EU-SILC und HBS, die eine Analyse der gemeinsamen Verteilung von Einkommen und Konsumausgaben der privaten Haushalte ermöglichen soll, befruchten. Hinsichtlich der Einordnung der Ergebnisse ist diesbezüglich jedoch festzuhalten, dass beide Verfahren, Random Hot-Deck und Predictive Mean Matching, der Annahme der bedingten Unabhängigkeit unterliegen, die in reellen Datensituationen selten erfüllt ist, weshalb eine exakte Reproduktion der gemeinsamen Verteilung der umfassenden Einkommensvariablen ($\mathbf{Y}$) aus EU-SILC und den detailliert erfassten Konsuminformationen ($\mathbf{Z}$) aus dem HBS nahezu unrealistisch erscheint. Dass PMM dennoch zumindest den nahen Bereich der tatsächlichen Zusammenhänge zwischen den jeweils nicht gemeinsam beobachteten Variablen $\mathbf{Y}$ und $\mathbf{Z}$ besser abdeckt, als Eurostat, konnte hier lediglich für hohe, tatsächliche Korrelationen von 0.79 und 0.86 sowie für mittelstarke Zusammenhänge von 0.25 und 0.29 gezeigt werden. Eine Bewertung der Reproduktion der Zusammenhänge zwischen $\mathbf{X}$ und $\mathbf{Z}$ war lediglich entlang von hohen Korrelationen, konkret von 0.93 und 0.95, möglich. Wie sich die Performance beider Verfahren für anderweitige Originalkorrelationen zwischen $\mathbf{Y}$ und $\mathbf{Z}$ sowie zwischen $\mathbf{X}$ und $\mathbf{Z}$ verhält, konnte im Rahmen dieser Arbeit nicht geklärt werden. Geschuldet ist dies insbesondere der Datensituation der Public Usefiles, die durch die verwendeten Variablen nahezu ausgeschöpft ist. Diesbezüglich wurde jedoch erläutert, dass, zumindest bei $a > 2$, die Datengrundlage aus theoretischen Erwägungen eher eine Fusionsperformance zugunsten des Random Hot-Deck-Verfahrens von Eurostat ermöglicht, als zugunsten von PMM, was darauf zurückzuführen ist, dass sechs von sieben $\mathbf{X}$-Variablen mit den simulierten HBS-Variablen einen Zusammenhang nahe 0 aufweisen. Dementsprechend kann zwar von einem eher konservativen Hypothesentest gesprochen werden, wobei dennoch bei Simulationsstudien allgemein zu beachten ist, dass mit den daraus gewonnenen Implikationen ein geringer Verallgemeinerungsgrad einhergeht. Somit bleibt empirisch ungeklärt, wie sich die Fusionsperformance von Random Hot-Deck und PMM unter anderweitigen Datensituationen darstellt. Daher unterliegen die Ergebnisse der durchgeführten Simulationsstudie

der Annahme, dass diese auch auf andere Datenkonstellationen übertragbar sind. Entlang des mitgelieferten R-Codes kann jedoch eine Replikation der Simulationsstudie mit anderweitigen und reelleren Daten, zum Beispiel mit Scientific Usefiles, unkompliziert erfolgen.

Mit einer solchen Replikation verbindet sich etwa der Auftrag für die künftige Forschung, die vorliegende Simulationsstudie mit einer realistischeren Datengrundlage durchzuführen, um abzuschätzen, wie sich eine höhere Variation in den Zusammenhängen zwischen $\mathbf{X}$ und $\mathbf{Z}$ auf die Fusionsperformance von PMM auswirkt. Hier wäre zu erwarten, dass je unterschiedlicher die Korrelationen $\rho_{\mathbf{XZ}}$ sind, desto besser dürfte das Fusionsergebnis von PMM aufgrund der Abstufungskomponente der ausgewählten Variablen $X_1, \ldots, X_a$ ausfallen. Neben einer solchen Replikation könnten künftige Forschungsbeiträge an diese Arbeit anknüpfen, indem etwa der Frage nachgegangen wird, wie sich die Problematiken mit der Conditional Independence Assumption, beispielsweise mit Blick auf die Einbeziehung von Hilfsinformationen, umgehen lassen. Weitere Simulationsstudien könnten so etwa Predictive Mean Matching gegen den in der aktuellen Forschung diskutierten *glue*-Ansatz von Fosdick et al. (2015) testen. Darüber hinaus bleibt in der bisherigen Forschung die Frage, welche konkreten Fehlerquellen den Datenfusionsalgorithmen, etwa Random Hot-Deck und PMM, zugrundeliegen, weitgehend unterbelichtet. Neben der Fehlerquelle der Verletzung der Conditional Independence Assumption dürfte auch die Abwesenheit perfekter Übereinstimmungen entlang der ausgewählten Variablen $X_1, \ldots, X_a$ einen Fehler mit Blick auf die Distanzmessung induzieren (Rodgers 1984). Hiermit verbindet sich für die künftige Forschung der Auftrag, die Fehlerquellen von Datenfusionsalgorithmen sowie deren Beziehung zueinander genauer zu untersuchen.

Somit bleibt die Implementation sowie die wissenschaftliche Beurteilung eines für die Datenfusion von EU-SILC und HBS geeigneten Verfahrens, um der Messung der gemeinsamen Verteilung von Einkommen und Konsumausgaben der privaten Haushalte in der EU möglichst präzise nachzukommen, auch künftig eine für die statistisch-methodische Forschung herausragende Aufgabe. Die vorliegende Arbeit lieferte zur Optimierung des bisherigen Datenfusionsverfahrens von Eurostat lediglich einen initialen Aufschlag, dem weitere Forschungsbeiträge folgen sollten.

Anhang

A Erläuterungen zu eigenhändigen Variablengenerierungen

Activity Status (X_1)

Als Grundlage für die Generierung dient die Variable PL031: „Self-defined current economic status“ (Eurostat 2013a: 277).

Ursprüngliche Ausprägungen gemäß PL031 (siehe Eurostat 2013a: 277):

1: Employee working full-time
2: Employee working part-time
3: Self-employed working full-time (including family worker)
4: Self-employed working part-time (including family worker)
5: Unemployed
6: Pupil, student, further training, unpaid work experience
7: In retirement or in early retirement or has given up business
8: Permanently disabled or/and unfit to work
9: In compulsory military community or service
10: Fulfilling domestic tasks and care responsibilities
11: Other inactive person

Modifizierungen für *Activity Status*:

Ursprüngliche Ausprägung	$\Longrightarrow$	*Activity Status* (inkl. Wertelabel gem. R-Code Eurostat)
1, 2, 3, 4	$\Longrightarrow$	1: Working
5	$\Longrightarrow$	2: Unemployed
7	$\Longrightarrow$	3: In retirement or early retirement
6	$\Longrightarrow$	4: Pupil, student, further training
10	$\Longrightarrow$	5: Fulfilling domestic tasks
8	$\Longrightarrow$	6: Permanently disabled
11	$\Longrightarrow$	7: Not applicable
9, NA	$\Longrightarrow$	9: Not specified (NA)

Anmerkung: Die ursprüngliche Ausprägung 9 (In compulsory military community or service) ist beim *Activity Status* gemäß Eurostat eine separate Kategorie. In den zugrundeliegenden Public Usefiles beträgt die absolute Häufigkeit dieser Ausprägung jedoch lediglich $N_9 = 11$ Beobachtungen, was zu Schwierigkeiten bei der Funktion `BBPMM.row()` führt (EU-SILC 2013 PUF: DE, FR, NL). Daher wird die ursprüngliche Ausprägung 9 für den *Activity Status* der vorliegenden Arbeit der zweithäufigsten Kategorie zugeordnet. Die zweithäufigste Kategorie bilden die fehlenden Werte (EU-SILC 2013 PUF: DE, FR, NL). Daher wird die ursprüngliche Ausprägung 9 (In compulsory military community or service) der NA-Kategorie (beim *Activity Status* ebenfalls als 9 kodiert) hinzugefügt.

Population Density Level (X_3)

Die entsprechend relevante und von Eurostat verwendete Variable ist die DB100: „Degree of urbanisation“ (Eurostat 2013a: 103).

Folgende Ausprägungen liegen der Variable zugrunde (siehe Eurostat 2013a: 103):
1: Densely populated area
2: Intermediate area
3: Thinly populated area

Allerdings ist die entsprechende Variable in den zugrundeliegenden Public Usefiles leer und wird daher zufällig generiert (EU-SILC 2013 PUF: DE, FR, NL). Mit dem `sample()`-Befehl in `R` wird unter Verwendung des Arguments `prob` für jedes Land (Deutschland, Frankreich, Niederlande) eine Ziehungswahrscheinlichkeit gemäß der tatsächlichen, relativen Häufigkeitsverteilungen der EU-SILC-Daten von 2013 gewährleistet, welche wiederum dem onlinebasierten Data Explorer von Eurostat entnommen werden können (Eurostat 2019: Data Explorer). In Tabelle 5 sind die entsprechenden relativen Häufigkeitswerte pro Land abgetragen:

	1: Densely populated area	2: Intermediate area	3: Thinly populated area
Deutschland	0.348	0.411	0.241
Frankreich	0.460	0.200	0.341
Niederlande	0.468	0.385	0.147

Werte entnommen aus Eurostat (2019: Data Explorer)

Tabelle 5: Relative Häufigkeiten – Degree of urbanisation

Anmerkung: Im Data Explorer sind die entsprechenden Labels als 1: „Cities“, 2: „Towns and suburbs“ und 3: „Rural Areas“ ausgewiesen (Eusotat 2019: Data Explorer).

Main Source of Income (X_6)

Die Variable *Main Source of Income* basiert auf einer von Eurostat eigens berechneten Hilfsvariable, die sechs Einkommensquellen beinhaltet. Dessen Generierung kann im `R`-Code nicht nachvollzogen werden. Lediglich die Wertelabels der Hilfsvariable sind darin enthalten:

1: Wages or salary
2: Income from self-employment
3: Property income
4: Pensions
5: Unemployment benefits
6: Other benefits
NA: Not specified

Entlang der Informationen und Ausführungen in Eurostat (2013a: 7, 306-309, 315-329) und Eurostat (2013b: 20, 27-28) wurde versucht, die oben angegebene Hilfsvariable entlang folgender Variablen des Personendatenfiles (p-file), die allesamt Einkommensquellen darstellen, zu generieren:

PY010G: „Employee cash or near cash income“ (Eurostat 2013a: 306)
PY020G: „Non-cash employee income“ (Eurostat 2013a: 308)
PY050G: „Cash benefits or losses from self-employment“ (Eurostat 2013a: 315)
PY080G: „Pension from individual private plans“ (Eurostat 2013a: 320)
PY090G: „Unemployment benefits“ (Eurostat 2013a: 322)
PY100G: „Old-age benefits“ (Eurostat 2013a: 322)
PY110G: „Survivor' benefits“ (Eurostat 2013a: 322)
PY120G: „Sickness benefits“ (Eurostat 2013a: 322)
PY130G: „Disability benefits“ (Eurostat 2013a: 322)
PY140G: „Education-related allowances“ (Eurostat 2013a: 322)

Diese Variablen (PY010G bis PY140G) sind allesamt metrisch mit einem Wertebereich gemäß der absoluten Höhe des Einkommensbestandteils. Um die oben genannte Hilfsvariable von Eurostat nachzubilden, wird für jedes Individuum ermittelt, welche der Variablen PY010G bis PY140G den höchsten Wert aufweist, also den höchsten Einkommensbestandteil darstellt. Anschließend erfolgt eine Zuteilung zu den Kategorien 1 bis 6 der Eurostat-Hilfsvariablen wie folgt:

Maximum bei

– PY010G oder PY020G $\Longrightarrow$ Kategorie 1
– PY050G $\Longrightarrow$ Kategorie 2
– PY080G $\Longrightarrow$ Kategorie 3
– PY100G $\Longrightarrow$ Kategorie 4
– PY090G $\Longrightarrow$ Kategorie 5
– PY110G, PY120G, PY130G oder PY140G $\Longrightarrow$ Kategorie 6

Sofern alle Variablen PY010G bis PY140G die Ausprägung 0 haben, wird dies als Kategorie 9 (NA) kodiert. Da laut Eurostat (2013a: 314, 320) „Pension from individual private plans“ (Variable PY080G) zu „Property Income“ zu zählen ist, wird dies als 3 kodiert.

Damit konnte die Hilfsvariable von Eurostat approximativ nachgebildet werden. Analog zu Eurostat erfolgte zur finalen Erstellung von *Main Source of Income* noch folgende Kategorisierung:

Ausprägung Hilfsvariable		*Main Source of Income*
1, 2, 3	$\Longrightarrow$	1
4, 5, 6	$\Longrightarrow$	2
9	$\Longrightarrow$	9 (NA)

B Relevante Tabellen zu $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$

Minimum, Maximum und Quantile der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

		$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	Minimum	0.2941	0.2777	−0.0522	−0.0620
	25%-Quantil	0.5066	0.5036	0.0878	0.0855
	50%-Quantil	0.5648	0.5687	0.1298	0.1274
	75%-Quantil	0.6134	0.6149	0.1796	0.1752
	Maximum	0.7537	0.7392	0.4437	0.4770
PMM	Minimum	0.4472	0.4595	−0.0622	−0.0301
	25%-Quantil	0.7233	0.7411	0.1338	0.1311
	50%-Quantil	0.7610	0.7835	0.2051	0.1956
	75%-Quantil	0.7956	0.8210	0.2795	0.2798
	Maximum	0.8969	0.9053	0.6358	0.6843

Minimum, Maximum und Quantile der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

		$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	Minimum	0.2485	0.2071	−0.0344	−0.0367
	25%-Quantil	0.5027	0.4958	0.0894	0.0896
	50%-Quantil	0.5623	0.5582	0.1254	0.1264
	75%-Quantil	0.6114	0.6018	0.1749	0.1721
	Maximum	0.7526	0.7432	0.4462	0.4741
PMM	Minimum	0.4878	0.3764	−0.0495	−0.0724
	25%-Quantil	0.7379	0.7541	0.1295	0.1217
	50%-Quantil	0.7816	0.7991	0.1911	0.1824
	75%-Quantil	0.8182	0.8364	0.2892	0.2820
	Maximum	0.9138	0.9328	0.7036	0.7351

Tabelle 6: Minimum, Maximum und Quantile für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

Mittelwerte der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	0.5546	0.5551	0.1375	0.1341
PMM	0.7530	0.7728	0.2145	0.2110

Mittelwerte der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	0.5519	0.5460	0.1359	0.1334
PMM	0.7737	0.7882	0.2232	0.2199

Tabelle 7: Mittelwerte für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

Monte-Carlo-Varianzen der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	0.0063	0.0069	0.0047	0.0049
PMM	0.0042	0.0050	0.0114	0.0127

Monte-Carlo-Varianzen der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	0.0065	0.0063	0.0047	0.0048
PMM	0.0047	0.0054	0.0186	0.0208

Tabelle 8: MC-Varianzen für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

Bias der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	0.2402	0.3082	0.1168	0.1525
PMM	0.0417	0.0905	0.0398	0.0756

Bias der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
Eurostat	0.2428	0.3173	0.1184	0.1532
PMM	0.0210	0.0750	0.0311	0.0668

Tabelle 9: Bias für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

MSE der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$
Eurostat	0.0640	0.1019	0.0184	0.0282
PMM	0.0059	0.0132	0.0130	0.0184

MSE der Korrelationen zwischen Y und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$
Eurostat	0.0655	0.1070	0.0187	0.0283
PMM	0.0051	0.0110	0.0195	0.0252

Tabelle 10: MSE für $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

C Relevante Tabellen und Grafiken zu $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$

Minimum, Maximum und Quantile der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

		$\widehat{\text{corr}}(X_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2,\widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_2)$
Eurostat	Minimum	−0.1902	−0.1736	0.3753	0.3363
	25%-Quantil	−0.0528	−0.0457	0.6189	0.6177
	50%-Quantil	−0.0115	−0.0067	0.6969	0.7027
	75%-Quantil	0.0310	0.0336	0.7468	0.7534
	Maximum	0.1966	0.2005	0.8399	0.8678
PMM	Minimum	−0.1702	−0.1960	0.5746	0.5036
	25%-Quantil	−0.0484	−0.0504	0.8996	0.9248
	50%-Quantil	−0.0108	−0.0131	0.9221	0.9521
	75%-Quantil	0.0248	0.0300	0.9383	0.9674
	Maximum	0.2140	0.1611	0.9707	0.9921

Minimum, Maximum und Quantile der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

		$\widehat{\text{corr}}(X_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2,\widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_2)$
Eurostat	Minimum	−0.2033	−0.1745	0.3732	0.3631
	25%-Quantil	−0.0552	−0.0484	0.6176	0.6129
	50%-Quantil	−0.0155	−0.0094	0.6889	0.6880
	75%-Quantil	0.0264	0.0283	0.7401	0.7367
	Maximum	0.1917	0.1714	0.8449	0.8574
PMM	Minimum	−0.2179	−0.1960	0.5936	0.5258
	25%-Quantil	−0.0517	−0.0514	0.9198	0.9452
	50%-Quantil	−0.0087	−0.0105	0.9354	0.9620
	75%-Quantil	0.0297	0.0290	0.9473	0.9720
	Maximum	0.1564	0.1717	0.9790	0.9901

Tabelle 11: Minimum, Maximum und Quantile für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

Mittelwerte der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	-0.0098	-0.0061	0.6788	0.6812
PMM	-0.0106	-0.0109	0.9105	0.9342

Mittelwerte der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	-0.0151	-0.0097	0.6728	0.6700
PMM	-0.0105	-0.0104	0.9294	0.9493

Tabelle 12: Mittelwerte für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

Monte-Carlo-Varianzen der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	0.0038	0.0036	0.0080	0.0091
PMM	0.0033	0.0033	0.0022	0.0033

Monte-Carlo-Varianzen der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	0.0035	0.0034	0.0074	0.0082
PMM	0.0036	0.0034	0.0013	0.0021

Tabelle 13: MC-Varianzen für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

Bias der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	0.0038	0.0024	0.2545	0.2735
PMM	0.0030	0.0024	0.0228	0.0205

Bias der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	0.0015	0.0012	0.2605	0.2847
PMM	0.0031	0.0019	0.0039	0.0054

Tabelle 14: Bias für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

MSE der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_1}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	0.0038	0.0036	0.0727	0.0839
PMM	0.0033	0.0033	0.0027	0.0037

MSE der Korrelationen zwischen X und $\widetilde{\mathbf{Z}}$ mit $\mathbf{n_2}$

	$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
Eurostat	0.0035	0.0034	0.0752	0.0892
PMM	0.0036	0.0034	0.0014	0.0021

Tabelle 15: MSE für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1 und n_2

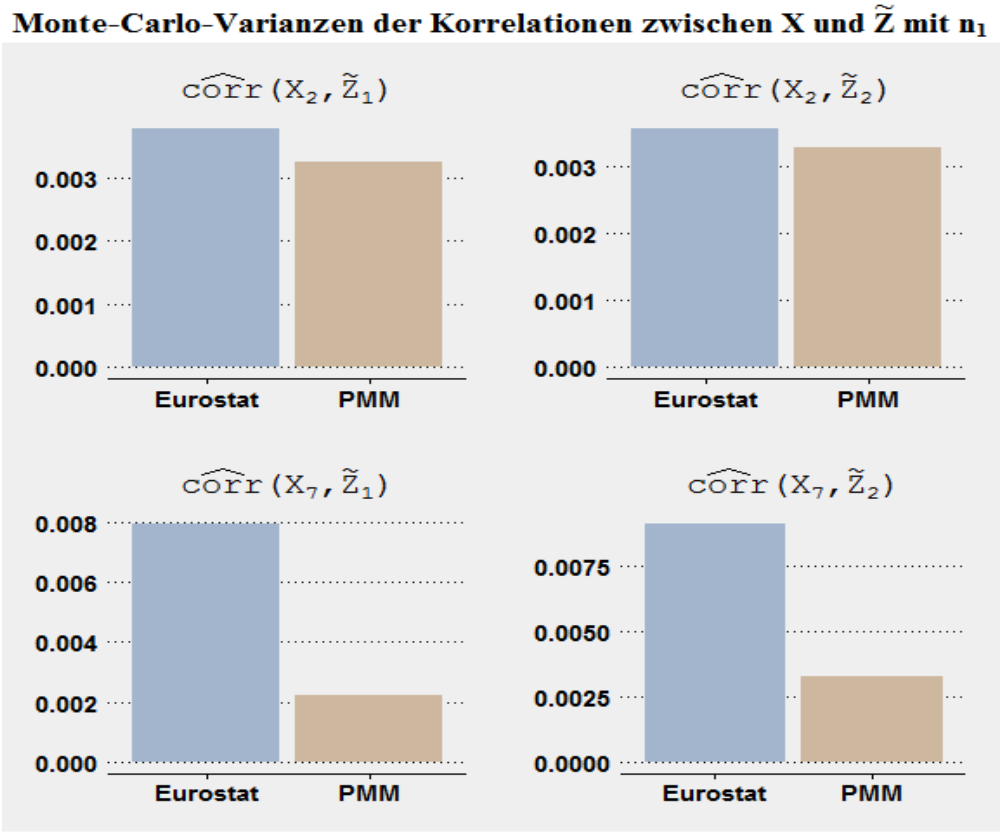

Abbildung 21: Barplots – MC-Varianzen für $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

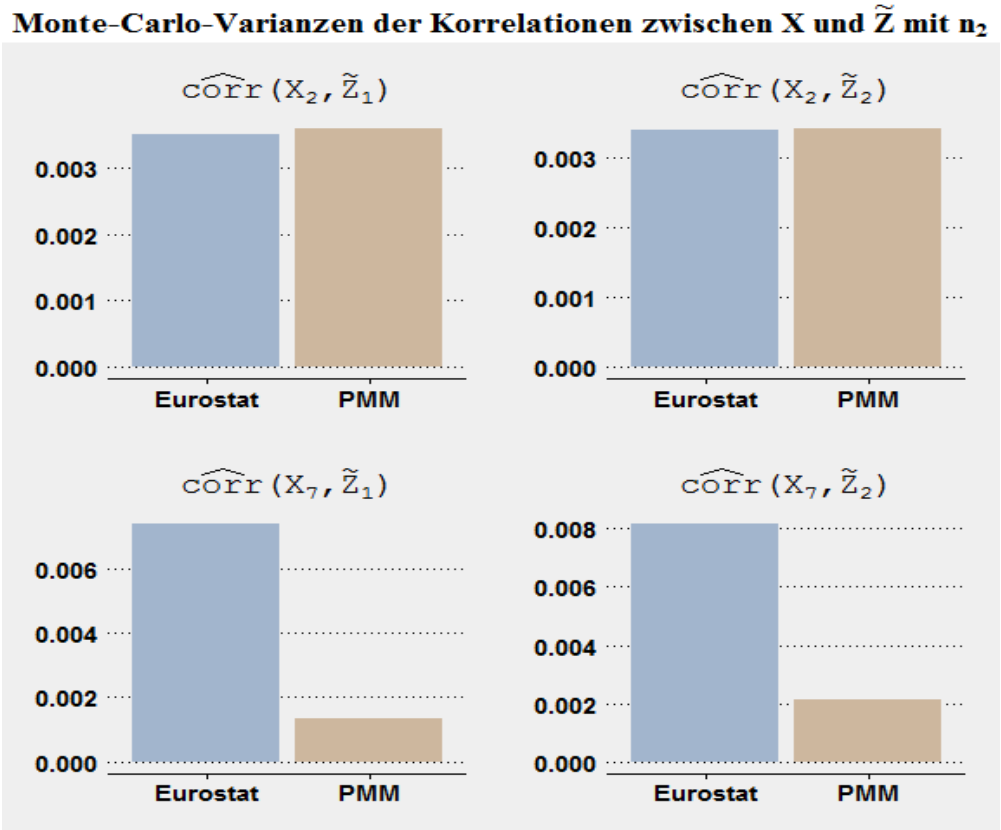

Abbildung 22: Barplots – MC-Varianzen für $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

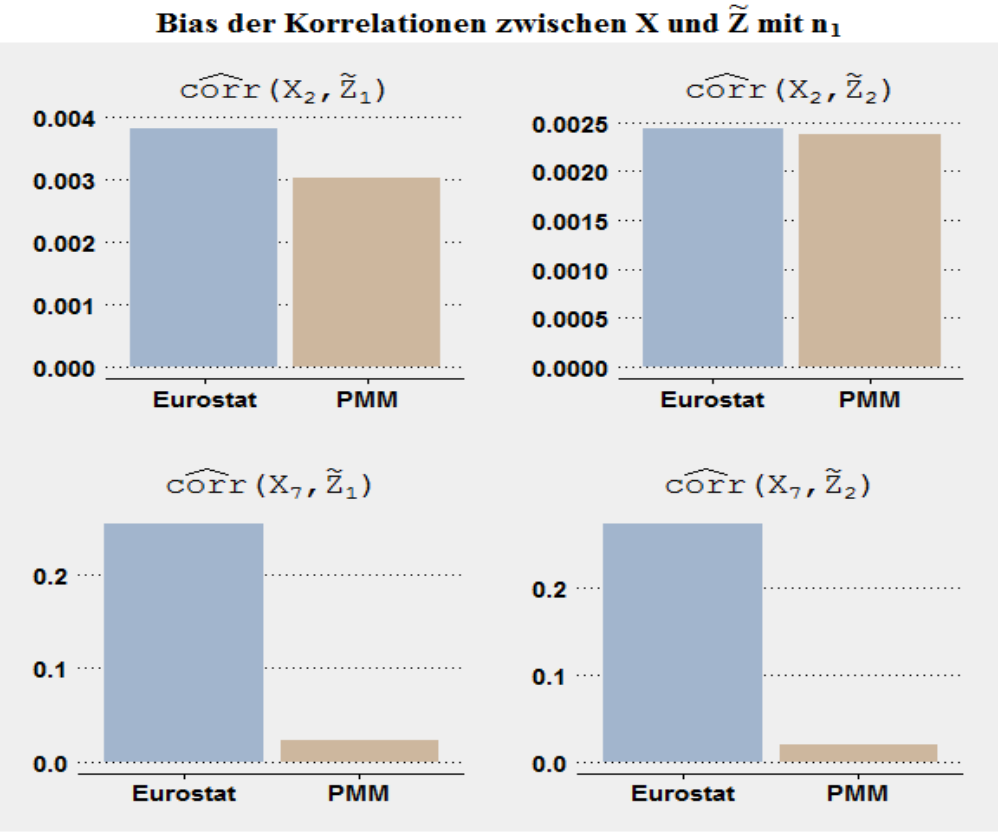

Abbildung 23: Barplots – Bias für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

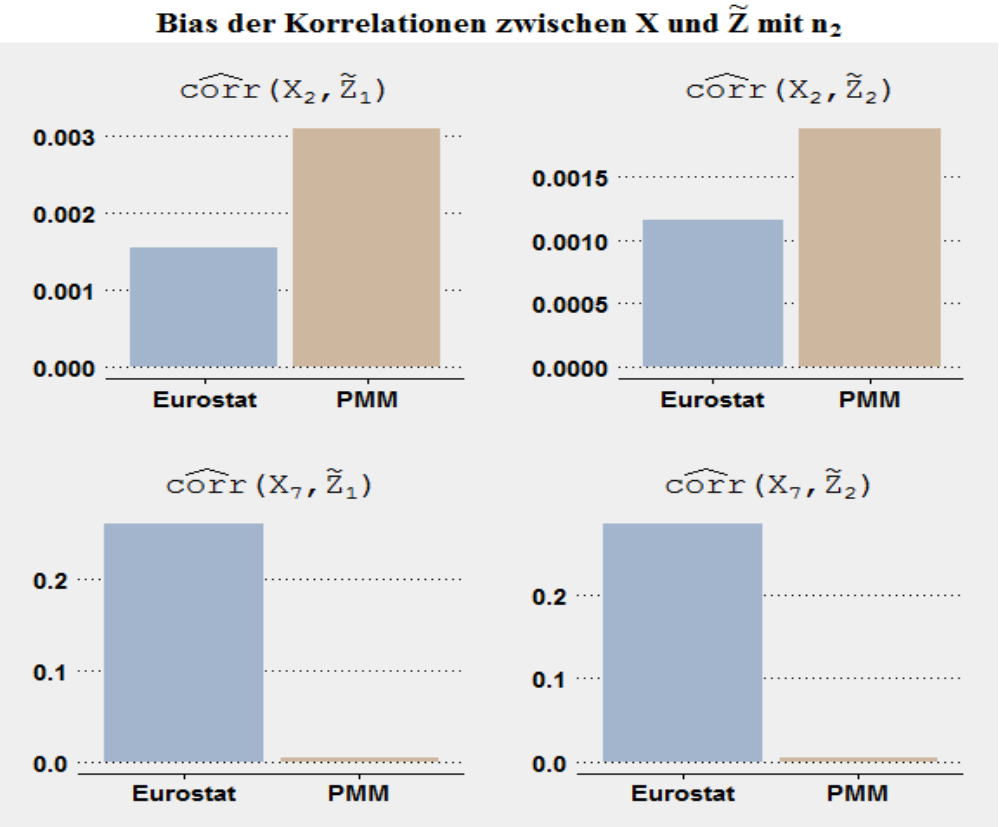

Abbildung 24: Barplots – Bias für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

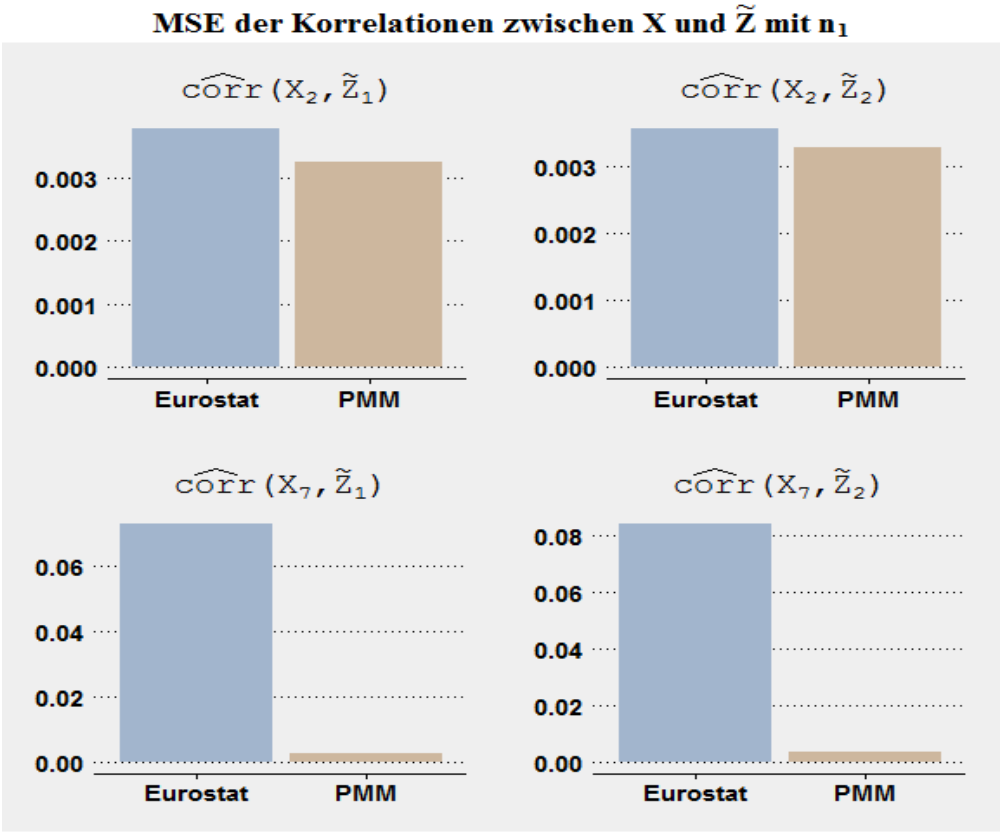

Abbildung 25: Barplots – MSE für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_1

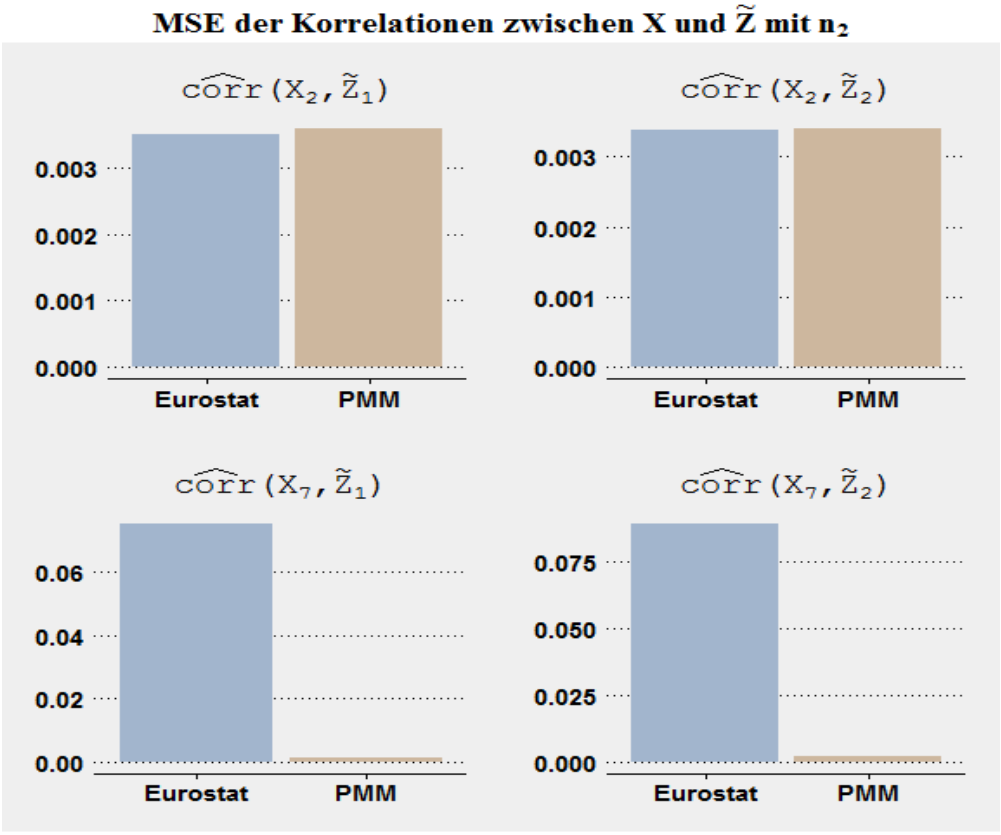

Abbildung 26: Barplots – MSE für $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ mit Stichprobenumfang n_2

Literaturverzeichnis

Auguie, B. / Antonov, A. (2017): `gridExtra`: Miscellaneous Functions for "Grid" Graphics. `R`-Package. Version 2.3. Link: https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf (aufgerufen am 07.03.2019).

Bacher, J. (2002): Statistisches Matching: Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS. In: ZA-Information / Zentralarchiv für Empirische Sozialforschung, 51, S. 38-66.

Balestra, C. (2018): Insights from the Eurostat-OECD Expert Group on Micro Statistics on Household Income, Consumption and Wealth. Slides. Expert Meeting on Measuring Poverty and Inequality. Link: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.15/2018/mtg1/Presentation_EmerigingI._OECD_NEW.pdf (aufgerufen am 06.02.2019).

Cielebak, J. / Rässler, S. (2014): Data Fusion, Record Linkage und Data Mining. In: Baur, N. / Blasuis, J. (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, S. 367-382.

D'Orazio, M. / Di Zio, M. / Scanu, M. (2006): Statistical Matching. Theory and Practice. Chichester: John Wiley & Sons, Ltd.

D'Orazio, M. (2017): `StatMatch`: Statistical Matching. `R`-Package. Version 1.2.5. Link: https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf
(aufgerufen am 24.01.2019).

Eurostat (2013a): Description of Target Variables: Cross-sectional and Longitudinal. EU-SILC 065, 2013 operation (Version May 2013). Link: https://circabc.europa.eu/sd/a/d7e88330-3502-44fa-96ea-eab5579b4d1e/SILC065%20operation%202013%20VERSION%20MAY%202013.pdf (aufgerufen am 19.01.2019).

Eurostat (2013b): European household income by groups of household. 2013 edition. Luxembourg: Publications Office of the European Union (Mothodologies & Working papers). Link: https://ec.europa.eu/eurostat/documents/3888793/5858173/KS-RA-13-023-EN.PDF/7e1dcfb2-2735-4334-9b0a-5a95af934b1d (aufgerufen am 15.03.2019).

Eurostat (2019): Eurostat – Data Explorer (Online-Website). Link: http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do (aufgerufen am 10.03.2019).

Fosdick, B. K. / DeYoreo, M. / Reiter, J. P. (2015): Categorical Data Fusion Using Auxiliary Information. June 22, 2015. Link: https://arxiv.org/pdf/1506.05886.pdf (aufgerufen am 09.02.2019).

Gilula, Z. / McCulloch, R. E. / Rossi, P. E. (2006): A Direct Approach to Data Fusion. In: Journal of Marketing Research, Vol. 43, No. 1 (Feb., 2006), S. 73-83.

Grothendieck, G. (2017): `sqldf`: Manipulate R Data Frames Using SQL. `R`-Package. Version 0.4-11. Link: https://cran.r-project.org/package=sqldf/sqldf.pdf (aufgerufen am 27.02.2019).

Kiesl, H. / Rässler, S. (2005): Techniken und Einsatzgebiete von Datenintegration und Datenfusion. In: König, C. / Stahl, M. / Wiegand, E. (Hrsg): Datenfusion und Datenintegration, 6. wissenschaftliche Tagung, S. 17-32.

Kiesl, H. / Rässler, S. (2006): How Valid Can Data Fusion Be? In: IAB Discussion Paper, No. 15/2006.

Koller-Meinfelder, F. (2009): Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching. Dissertation, Otto-Friedrich-Universität Bamberg. Link: https://opus4.kobv.de/opus4-bamberg/files/200/thesisKollerMeinfelderAop.pdf (aufgerufen am 01.02.2019).

Koschnick, W. J. (1995): Standard-Lexikon für Markt- und Konsumforschung. Band 1, A – K. München: K.G. Saur Verlag GmbH.

Lamarche, P. (2017): Measuring Income, Consumption and Wealth jointly at the micro-level. Luxembourg: Eurostat (preliminary version, June 20, 2017). Link: https://ec.europa.eu/eurostat/documents/7894008/8074103/income_methodological_note.pdf (aufgerufen am 24.01.2019).

Leulescu, A. / Agafitei, M. (2013): Statistical matching: a model based approach for data integration. 2013 edition. Luxembourg: Publications Office of the European Union (Methodologies & Working papers). Link: http://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF/477dd541-92ee-4259-95d4-1c42fcf2ef34?version=1.0 (aufgerufen am 18.02.2019).

Little, R. J. A. (1988): Missing-Data Adjustments in Large Surveys. In: Journal of Business & Economic Statistics, Vol. 6, No. 3 (Jul., 1988), S. 287-296.

Little, R. J. A. / Rubin, D. B. (1991): Statistical Analysis With Missing Data. Book Review. In: Journal of Educational Statistics, Vol. 16, No. 2 (Summer, 1991), S. 150-155.

Little, R. J. A. / Rubin, D. B. (2002): Statistical Analysis with Missing Data. Second Edition. Hoboken: John Wiley & Sons, Inc.

Lüdecke, D. (2019): `sjstats`: Collection of Convenient Functions for Common Statistical Computations. `R`-Package. Version 0.17.4. Link: https://cran.r-project.org/web/packages/sjstats/sjstats.pdf (aufgerufen am 19.03.2019).

Lumley, T. / (based on Fortran code by) Miller, A. (2017): `leaps`: Regression Subset Selection. `R`-Package. Version 3.0. Link: https://cran.r-project.org/web/packages/leaps/leaps.pdf (aufgerufen am 26.01.2019).

Meinfelder, F. (2013): Datenfusion: Theoretische Implikationen und praktische Umsetzung. In: Riede, T. / Schmidt, T. / Eisele, M. / Schimpl-Neimanns, B. / Meinfelder, F. / Münnich, R. / Burgard, J. P. / Zimmermann, T. / Ott, N. / Brechthold, S. (Hrsg.): Weiterentwicklung der amtlichen Haushaltsstatistiken, 1. aufl. edn, Berlin: Scivero, S. 83-98.

Meinfelder, F. / Schnapp, T. (2015): BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and Single Imputation for Discrete Data. R-Package. Version 0.2-0. Link: https://cran.r-project.org/web/packages/BaBooN/BaBooN.pdf (aufgerufen am 24.01.2019).

Meschiari, S. (2015): latex2exp: Use LaTeX Expressions in Plots. R-Package. Version 0.4.0. Link: https://cran.r-project.org/web/packages/latex2exp/latex2exp.pdf (aufgerufen am 07.03.2019).

Morris, T. P. / White, I. R. / Crowther, M. J. (2019): Using simulation studies to evaluate statistical methods. In: Statistics in Medicine, Early View (16 January 2019), S. 1-29.

Murrell, P. / Wen, Z. (2018): gridGraphics: Redraw Base Graphics Using 'grid' Graphics. R-Package. Version 0.3-0. Link: https://cran.r-project.org/web/packages/gridGraphics/gridGraphics.pdf (aufgerufen am 14.03.2019).

Okner, B. A. (1972): Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File. In: Annals of Economic and Social Measurement, Vol. 1, No. 3, S. 325-342.

Raghunathan, T. (2016): Missing Data Analysis in Practice. Boca Raton: CRC Press.

Rässler, S. (2002): Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer-Verlag, Inc.

Ripley, B. / Venables, B. / Bates, D. M. / Hornik, K. / Gebhardt, A. / Firth, D. (2018): MASS: Support Functions and Datasets for Venables and Ripley's MASS. R-Package. Version 7.3-51.1. Link: https://cran.r-project.org/web/packages/MASS/MASS.pdf (aufgerufen am 01.03.2019).

Rodgers, W. L. (1984): An Evaluation of Statistical Matching. In: Journal of Business & Economic Statistics, Vol. 2, No. 1 (Jan., 1984), S. 91-102.

Rodgers, J. L. (1999): The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy. In: Multivariate Behavioral Research, Vol. 34, No. 4, S. 441-456.

Rubin, D. B. (1976): Inference and missing data. In: Biometrika, Vol. 63, No. 3 (Dec., 1976), S. 581-592.

Rubin, D. B. (1986): Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. In: Journal of Business & Economic Statistics, Vol. 4, No. 1 (Jan., 1986), S. 87-94.

Rubin, D. B. (1987): Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Serafino, P. / Tonkin, R. (2017): Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey. 2017 edition. Luxembourg: Publications Office of the European Union (Statistical working papers, theme 3, 2017 edition).

Sims, C. A. (1972): Comments (zu Okner 1972). In: Annals of Economic and Social Measurement, Vol. 1, No. 3, S. 343-354.

Singh, A. C. / Mantel, H. J. / Kinack M. D. / Rowe, G. (1993): Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. In: Survey Methodology, Vol. 19, No. 1 (June 1993), S. 59-79.

Spieß, M. (2008): Missing-Data Techniken. Analyse von Daten mit fehlenden Werten. Hamburg: Lit Verlag.

Statistisches Bundesamt (2016a): Gemeinschaftsstatistik über Einkommen und Lebensbedingungen. Leben in Europa 2013. Qualitätsbericht (erschienen am 27. Januar 2016). Link: https://www.forschungsdatenzentrum.de/sites/default/files/eu-silc_2013_qb.pdf (aufgerufen am 11.02.2019).

Statistisches Bundesamt (2016b): Einkommens- und Verbrauchsstichprobe. EVS 2013. Qualitätsbericht (erschienen am 5. Oktober 2016). Link: https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/EinkommenKonsumLebensbedingungen/WirtschaftsrechnEVS13.pdf?__blob=publicationFile (aufgerufen am 11.02.2019).

Stiglitz, E. / Sen, A. / Fitoussi, J. P. (2009): Report by the Commission on the Measurement of Economic Performance and Social Progress. Link: https://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report (aufgerufen am 04.02.2019).

Talbot, J. / Arnold, J. B. / Auguie, B. (2019): `ggthemes`: Extra Themes, Scales and Geoms for 'ggplot2'. `R`-Package. Version 4.1.0. Link: https://cran.r-project.org/web/packages/ggthemes/ggthemes.pdf (aufgerufen am 02.03.2019).

Tillé, Y. / Matei, A (2016): `sampling`: Survey Sampling. `R`-Package. Version 2.8. Link: https://cran.r-project.org/web/packages/sampling/sampling.pdf (aufgerufen am 27.02.2019).

Van Buuren, S. (2018): `mice`: Multivariate Imputation by Chained Equations. `R`-Package. Version 3.3.0. Link: https://cran.r-project.org/web/packages/mice/mice.pdf (aufgerufen am 06.02.2019).

Van der Putten, P. / Kok, J. N. / Gupta, A. (2002): Data Fusion Through Statistical Matching. In: MIT Sloan School of Management, Working Paper 4342-02 (Jan., 2002).

Webber, D. / Tonkin, R. (2013): Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. 2013 edition. Luxembourg: Publications Office of the European Union (Methodologies & Working papers). Link: https://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF/37d4ffcc-e9fc-42bc-8d4f-fc89c65ff6b1 (aufgerufen am 02.03.2019).

Wickham, H. / Chang, W. / Henry, L. (2018): `ggplot2`: Create Elegant Data Visualisations Using the Grammar of Graphics. `R`-Package. Version 3.1.0. Link: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf (aufgerufen am 02.03.2019).

Wickham, H. / François, R. / Henry, L. / Müller, L. (2019): dplyr: A Grammar of Data Manipulation. R-Package. Version 0.8.0.1. Link: https://cran.r-project.org/web/packages/dplyr/dplyr.pdf (aufgerufen am 27.02.2019).

Wickham, H. (2019): stringr: Simple, Consistent Wrappers for Common String Operations. R-Package. Version 1.4.0. Link: https://cran.r-project.org/web/packages/stringr/stringr.pdf (aufgerufen am 28.02.2019).

Quellenverzeichnis

EU-SILC 2013 PUF DE: European Union Statistics on Income and Living Conditions, 2013. Pubilc Usefile Germany.

EU-SILC 2013 PUF FR: European Union Statistics on Income and Living Conditions, 2013. Pubilc Usefile France.

EU-SILC 2013 PUF NL: European Union Statistics on Income and Living Conditions, 2013. Pubilc Usefile Netherlands.

Disclaimer: The responsibility for all conclusions drawn from the data lies entirely with the author.

Die Daten wurden in dieser Arbeit mit der Statistik-Software R (Version 3.5.2) unter Verwendung von RStudio (Version 1.1.463) bearbeitet und können als ZIP-Dateien unter folgendem Link runtergeladen werden: https://ec.europa.eu/eurostat/web/microdata/statistics-on-income-and-living-conditions (aufgerufen am 07.01.2019).

Erklärung

Ich erkläre hiermit gem. § 5 Abs. 3 PuStO, dass ich die vorstehende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

(Datum) (Unterschrift)