Schaller, Jannik

Article

Datenfusion von EU-SILC und Household Budget Survey – ein Vergleich zweier Fusionsmethoden

WISTA – Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Schaller, Jannik (2021) : Datenfusion von EU-SILC und Household Budget Survey – ein Vergleich zweier Fusionsmethoden, WISTA – Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 73, Iss. 4, pp. 76-86

This Version is available at:
https://hdl.handle.net/10419/237402

Jannik Schaller

ist seit seinem Masterabschluss in „Survey-Statistik“ Wissenschaftlicher Mitarbeiter im Referat „Forschungsdatenzentrum, Methoden der Datenanalyse“ des Statistischen Bundesamtes. Er betreut das Thema Mikrosimulation und erforscht Verfahren der Datenfusion/Statistical Matching zur Erstellung einer synthetischen Mikrosimulationsdatengrundlage.

Für seine an der Otto-Friedrich-Universität Bamberg bei Professor Dr. Martin Messingschlager verfasste Masterarbeit zum Thema „Datenfusion von EU-SILC und HBS: Vergleich zwischen Random Hot-Deck und Predictive Mean Matching im Rahmen einer Simulationsstudie” wurde er mit dem Gerhard-Fürst-Preis 2020 des Statistischen Bundesamtes ausgezeichnet. Mit dem folgenden Beitrag stellt er diese Masterarbeit vor.

DATENFUSION VON EU-SILC UND HOUSEHOLD BUDGET SURVEY – EIN VERGLEICH ZWEIER FUSIONSMETHODEN

Jannik Schaller

↘ **Schlüsselwörter:** Statistisches Matching — fehlende Daten — Imputationsverfahren — Missing-By-Design-Pattern — Predictive Mean Matching

ZUSAMMENFASSUNG

Zur Beurteilung des sozialen und wirtschaftlichen Lebensstandards in der Europäischen Union (EU) strebt die amtliche Statistik an, Einkommen und Konsumausgaben privater Haushalte gemeinsam zu betrachten. Hierfür existiert derzeit keine einheitliche Datengrundlage. Die Einkommensangaben sind detailliert in der europaweiten EU-SILC-Erhebung, die Konsuminformationen wiederum im Household Budget Survey erfasst. Beide Datenbestände sollen fusioniert werden, um eine gemeinsame Analyse von Einkommen und Konsumausgaben privater Haushalte zu ermöglichen. Der vorliegende Beitrag vergleicht den von Eurostat vorgeschlagenen Random Hot-Deck-Ansatz mit einer alternativen Datenfusionsmethode, dem Predictive Mean Matching. Er evaluiert die Performance beider Verfahren in einer Simulationsstudie.

↘ ***Keywords:*** *statistical matching — missing data — imputation techniques — missing-by-design pattern — predictive mean matching*

ABSTRACT

In order to assess the social and economic living standards in the European Union (EU), official statistics seek to provide joint information on income and consumption expenditures of private households. Currently, no uniform data basis exists for this purpose. While income is measured across the EU in great detail by EU-SILC, information on consumption is reported in the household budget survey. Both stocks of data are to be fused in order to enable a joint analysis of household income and consumption expenditure. This article compares the random hot deck approach proposed by Eurostat with predictive mean matching, an alternative data fusion method. It evaluates the performance of both methods in a simulation study.

1

Einleitung

Zur umfassenderen Beurteilung des sozialen und wirtschaftlichen Lebensstandards privater Haushalte in der Europäischen Union (EU) strebt die amtliche Statistik seit einigen Jahren an, die gemeinsame Verteilung von Einkommen, Konsumausgaben und Vermögen privater Haushalte zu messen. Diese Bestrebungen basieren auf Empfehlungen der Stiglitz-Sen-Fitoussi-Kommission von 2009 (Stiglitz und andere, 2009). Momentan existiert in der amtlichen Statistik jedoch keine befriedigende Datenquelle, die alle drei Komponenten gemeinsam erfasst. Daher sollen entsprechende Datenquellen, die die jeweiligen Determinanten umfassend widerspiegeln, fusioniert werden. Dabei stehen drei Datenquellen zur Verfügung: für Einkommen die Erhebung „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 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 zusätzlich das Ziel, Armutsrisiken und Armutsindikatoren präziser abbilden zu können als mit dem bisherigen Messinstrument, das sich weitgehend nur auf das Einkommen selbst beschränkt und die Konsumausgaben außer Acht lässt (Serafino/Tonkin, 2017).

Um den primären und initialen Fokus der amtlichen Statistik aufzugreifen, der sich auf die gemeinsame Verteilung von Einkommen und Konsumausgaben bezieht, untersucht der vorliegende Beitrag die mögliche Datenfusion von EU-SILC und HBS. Er baut auf dem aktuellen Forschungsstand des Statistischen Amtes der Europäischen Union (Eurostat) auf. Um EU-SILC und HBS zu fusionieren, verfolgt Eurostat die Ergänzung des vorhandenen EU-SILC-Datensatzes um die im HBS enthaltenen Konsumvariablen. Daraus soll ein vollständiger, fusionierter Mikrodatenfile aus Einkommens- und Konsuminformationen resultieren. Daher 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.

Die von Eurostat vorgeschlagene Datenfusionsmethode entspricht dem Random Hot-Deck-Verfahren, ein nichtparametrischer Ansatz, welcher über Ähnlichkeitsmaße in den gemeinsamen Variablen maximal ähnliche Beobachtungen zu identifizieren versucht (D’Orazio und andere, 2006, hier: Kapitel 2.4.1; Lamarche, 2017). Predictive Mean Matching (PMM) nach Rubin (1986) und Little (1988) könnte hingegen eine vielversprechende Fusionsalternative darstellen, da PMM ein semiparametrisches Verfahren ist und Beobachtungen des Rezipienten- und Donorendatensatzes anhand modellbasierter Distanzen verknüpft. Ziel dieses Artikels ist es, die Performance von Random Hot-Deck und PMM hinsichtlich des Erhalts gemeinsamer Verteilungen zwischen den nicht gemeinsam beobachteten Variablen Einkommen und Konsum im Zuge einer Simulationsstudie zu evaluieren.

Der vorliegende Artikel gliedert sich dabei wie folgt: Kapitel 2 bietet zunächst eine kurze Einführung zu Datenfusionen und stellt die relevanten Fusionsalgorithmen, Random Hot-Deck und PMM, allgemein vor. Kapitel 3 skizziert das Simulationsdesign und die Evaluationskriterien, während in Kapitel 4 die Ergebnisse der Simulationsstudie diskutiert werden. Kapitel 5 fasst die Ergebnisse und Implikationen für die amtliche Statistik zusammen.

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.

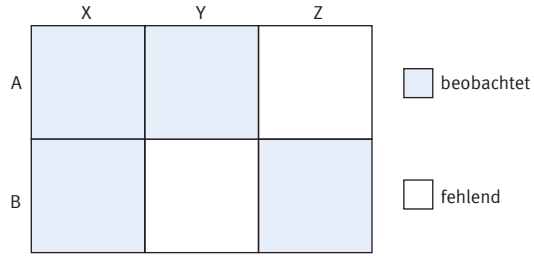
2

Methodischer Überblick und Fusionsalgorithmen

2.1 Datenfusion: eine Einführung

Datenfusionen bezeichnen allgemein gesprochen ein spezifisches Datenausfallmuster, konkret ein Missing-By-Design-Pattern, welches durch „Übereinanderstapeln“ zweier oder mehrerer Datenquellen entsteht (Rässler, 2002; Meinfelder, 2013). Während in Datenfile A und B die *X*-Variablen gemeinsam beobachtet sind, enthält Datenfile A lediglich Informationen zu *Y* (jedoch nicht zu *Z*), dagegen sind in Datenfile B lediglich Informationen zu *Z* vorhanden (jedoch nicht zu *Y*). Dementsprechend wird fortlaufend die Menge an gemeinsamen Variablen als *X*, die Menge an spezifischen, nicht gemeinsam beobachteten Variablen hingegen als *Y* und *Z* bezeichnet. Das Analyseziel einer Datenfusion bezieht sich stets auf die gemeinsame Verteilung der ursprünglich nicht gemeinsam beobachteten Merkmale *Y* und *Z* (Kiesl/Rässler, 2005). ↘ Grafik 1

Grafik 1
Datenfusion als spezifisches Datenausfallmuster



Quelle: Meinfelder (2013), hier: Seite 85

2021 - 0283

Zwar wäre es möglich, wie Grafik 1 suggeriert, beide fehlenden Datenblöcke, *Z* in A und *Y* in B, zu ergänzen. Allerdings wird in der Praxis meist lediglich einer der beiden Variablenblöcke mittels Datenfusionsverfahren ergänzt. Hinsichtlich der Datenfusion von EU-SILC und HBS entspricht Datenfile A EU-SILC mit den beobachteten Einkommensinformationen (*Y*), während Datenfile B den HBS widerspiegelt, welcher die Konsuminformationen (*Z*) liefert, die in der Rezipientenstichprobe EU-SILC imputiert werden sollen.

Datenfusionen werden mithilfe der Informationen der gemeinsamen Variablen (*X*) durchgeführt, häufig durch einfache, nicht-parametrische Distanzberechnung (Koschnick, 1995; van der Putten, 2002), wobei ebenfalls semi-parametrische Verfahren (wie etwa PMM) oder sogar rein parametrische Ansätze (etwa basierend auf Regressionen) zum Einsatz kommen können (D’Orazio und andere, 2006, hier: Seite 14 ff.; Gilula und andere, 2006).[3] Herkömmliche Fusionsverfahren, auch Random Hot-Deck und PMM, unterstellen jedoch implizit die Conditional Independence Assumption, also die Unabhängigkeit von *Y* und *Z* gegeben *X*, woraus $corr(Y, Z|X) = 0$ resultiert. Diese Annahme der bedingten Unabhängigkeit wurde zuerst von Sims (1972) in einem Kommentar zu Okner (1972) aufgezeigt. Die Problematiken dieser Annahme wurden ausführlich in Rodgers (1984) thematisiert, während in den letzten Jahren auch vermehrt mittels Hilfsinformationen versucht wird, die Conditional Independence Assumption abzuschwächen oder zu umgehen (Singh und andere, 1993; Fosdick und andere, 2016). Kiesl und Rässler (2006) diskutieren hingegen die Möglichkeit, mittels Fréchet-Hoeffding-Grenzen für jedes Variablenpaar von *Y* und *Z* eine Ober- und Untergrenze für die marginale, gemeinsame kumulative Verteilungsfunktion zu berechnen. Im Rahmen dieses Beitrags werden Annahmeverletzungen nicht explizit thematisiert. Hinsichtlich des angestrebten Performancevergleichs von Random Hot-Deck und PMM wird demnach die Conditional Independence Assumption unterstellt, da beide Algorithmen bedingte Unabhängigkeit im fusionierten Datenfile herstellen.

2.2 Random Hot-Deck

Nach der kurzen methodischen Einführung in das Datenfusionsthema folgt nun die Vorstellung der beiden relevanten Fusionsalgorithmen, Random Hot-Deck und PMM. Der Random Hot-Deck-Ansatz basiert auf einer zufälligen Zuweisung von Rezipienten- und Donorenbeobachtungen, welche in der Regel innerhalb homogener Teilgruppen durchgeführt wird (D’Orazio und andere, 2006, hier: Kapitel 2.4.1).

3 Einen ausführlichen Überblick klassischer Fusionsalgorithmen liefern Rässler (2002) sowie D’Orazio und andere (2006). Rässler (2002) thematisiert zusätzlich alternative, bayesianische Fusionsalgorithmen, während D’Orazio und andere (2006) auch makrobasierte Ansätze aufgreift.

Konkret stellt sich die von Eurostat vorgeschlagene Random Hot-Deck-Methode nach Lamarche (2017) für jede spezifische, zu fusionierende Variable Z_r wie folgt dar: Zunächst erfolgt eine Kategorisierung aller metrischen Variablen. Anschließend werden mithilfe von Backward-Selection aus einer Regression von Z_r auf X für den Fusionsprozess relevante gemeinsame X-Variablen ausgewählt. Hieraus werden Schichten bestehend aus Zellkombinationen der ausgewählten X-Variablen für jede Beobachtung des Rezipienten- und Donorendatensatzes gebildet. Sofern beispielsweise Alter und Geschlecht als gemeinsame Variablen ausgewählt werden, fasst die Schichtung deren Ausprägungen, also die Zellkombination zusammen (Geschlecht „1“ und Altersgruppe „3“ werden etwa zu „13“). Das Random Hot-Deck, also die zufällige Zuweisung der Rezipienten zu potenziellen Donoren, wird nun innerhalb derselben Schichtungsausprägung durchgeführt. Um einen ausreichend großen Donorenpool sicherzustellen, wird ein Schwellwert gesetzt, gemäß dem auf einen Donoren höchstens drei Rezipienten fallen dürfen. Dabei wird für 10 % der Rezipienten eine Abweichung dahingehend toleriert, dass ein Donor auf mehr als drei Rezipienten fällt (Lamarche, 2017).

Sollten Rezipienten übrig bleiben, für die entlang der ausgewählten X-Variablen noch kein merkmalsgleicher Donor zur Verfügung steht, wird die Variablenauswahl so lange weiter iteriert, bis entlang der gebildeten Schichtungsausprägungen alle verbliebenen Rezipienten mit mindestens einem Donor übereinstimmen. Dabei wird auf den oben genannten Schwellwert nun verzichtet und die Variablenauswahl in jedem Iterationsschritt um eins reduziert. Sofern alle Rezipienten einem Donor zugewiesen werden konnten, bekommt jede Rezipientenbeobachtung den realen Wert der jeweiligen Z_r-Ausprägung des zugewiesenen Donors. Dieser Fusionsansatz basiert lediglich auf Distanzen innerhalb der X-Variablen, ohne modellbasierte Einbeziehung weiterer Informationen. Daher kann das Random Hot-Deck als nicht-parametrisches Verfahren betrachtet werden (Lamarche, 2017).

2.3 Predictive Mean Matching

Predictive Mean Matching (PMM) ist hingegen ein semi-parametrisches Verfahren mit der Grundidee, für jeden fehlenden Wert einer Variablen ihren prädiktiven Mittelwert, der sich mithilfe einer OLS-Regression berechnet, mit den prädiktiven Mittelwerten der beobachteten Werte zu vergleichen. Die Beobachtung mit dem ähnlichsten prädiktiven Mittelwert fungiert als Donor, dessen realer Wert sodann den fehlenden Wert ersetzt (Rubin, 1986; Little, 1988).

Konkret kann das PMM-Verfahren für jede spezifische, zu fusionierende Variable Z_r wie folgt beschrieben werden: Zunächst werden, wie beim Random Hot-Deck-Ansatz, für die Fusion relevante X-Variablen mithilfe von Backward-Selection ausgewählt. Hierbei entfällt jedoch die Kategorisierung metrischer Variablen und alle Merkmale können auf ihrem ursprünglichen Skalenniveau verbleiben. Anschließend wird, basierend auf einer Regression von Z_r auf X im Donorendatensatz, für jede Beobachtung im Rezipienten- und Donorendatensatz der prädiktive Mittelwert berechnet. Die Suche nach entsprechenden Donoren erfolgt nun unter Verwendung der von Little (1988) vorgeschlagenen Mahalanobis-Distanz:

$$D_{i,j} = (\hat{z}_i - \hat{z}_j)^T \, S_{Z|X}^{-1} \, (\hat{z}_i - \hat{z}_j)$$

mit $i = 1, \ldots, n_{silc}$ und $j = 1, \ldots, n_{hbs}$, wobei $\hat{z}_i$ dem prädiktiven Mittelwert der i-ten Rezipientenbeobachtung und $\hat{z}_j$ dem prädiktiven Mittelwert der j-ten Donorenbeobachtung entspricht (Meinfelder, 2013, hier: Seite 89). $S_{Z|X}^{-1}$ stellt die inverse Varianz-Kovarianzmatrix der Residuen der Regression von Z_r auf X dar, die als Gewichtungsmatrix fungiert. Dabei kommt den gemeinsamen X-Variablen, welche eine gute Erklärungskraft bezüglich der zu fusionierenden Variable Z_r aufweisen, ein stärkerer Einfluss auf die Gesamtdistanz zu als jenen X-Variablen mit geringerer Erklärungskraft. Distanzen werden also bei guter Erklärungskraft stärker, bei schwacher Erklärungskraft weniger stark bestraft (Koller-Meinfelder, 2009, hier: Seite 33 f.; Meinfelder, 2013).

Nach Berechnung der Distanzen $D_{i,j}$ wird der real beobachtete Z_r-Wert des maximal ähnlichen Donors im Rezipientendatensatz imputiert. Sofern ein Rezipient zu mehreren Donoren die minimalste Distanz aufweist, erfolgt eine zufällige Auswahl dieser Donoren. Da PMM Distanzen aufgrund modellbasierter Regressionen in parametrischer Weise bildet, jedoch anschließend den realen Wert in nicht-parametrischer Weise imputiert, kann PMM als semi-parametrisches Verfahren angesehen werden.

2.4 Diskussion beider Fusionsverfahren

Beim methodischen Vergleich beider Ansätze, Random Hot-Deck und PMM, sind insbesondere zwei elementare Unterschiede zu erkennen: Erstens müssen bei Random Hot-Deck alle gemeinsamen *X*-Variablen kategorisiert werden, was für metrische Variablen einen Informationsverlust induziert, während PMM auch nicht kategoriale *X*-Variablen verarbeiten kann. Zweitens besteht ein entscheidender Unterschied darin, inwiefern die *X*-Variablen gemäß ihrer Erklärungskraft auf die zu fusionierenden, spezifischen *Z*-Variablen abgestuft beziehungsweise gewichtet werden. Bei Random Hot-Deck fließt jede gemeinsame, ausgewählte *X*-Variable mit dem gleichen Gewicht in den Fusionsprozess mit ein, ungeachtet dessen, ob diese *X*-Variablen untereinander eine unterschiedliche Erklärungskraft auf die zu fusionierende *Z*-Variable aufweisen. Wenn etwa eine durch die Backward-Selection ausgewählte *X*-Variable im Vergleich zu den anderen, ebenfalls ausgewählten gemeinsamen Variablen die in diesem Artikel thematisierten Konsuminformationen *Z* schlechter erklären kann, so spielt dies bei der Distanzberechnung beziehungsweise bei der Suche nach Übereinstimmungen keine Rolle. PMM hingegen löst dieses Optimierungsproblem mittels Distanzgewichtung anhand der Inversen der Varianz-Kovarianzmatrix der Residuen einer Regression von *Z* auf *X*. Je besser beziehungsweise schlechter dabei die gemeinsamen Variablen die Konsumangaben *Z* erklären können, desto stärker beziehungsweise schwächer ist ihr Einfluss in der Distanzberechnung (Koller-Meinfelder, 2009, hier: Seite 33 f.; Meinfelder, 2013). Dementsprechend wird aufgrund der nicht notwendigen Kategorisierung gemeinsamer Merkmale sowie der Gewichtung der *X*-Variablen entlang ihres Erklärungsbeitrags als Hypothese unterstellt, dass PMM ein besseres Fusionsergebnis produziert als Random Hot-Deck.

3

Simulationsdesign

Um die unterstellte Arbeitshypothese zu überprüfen und zu evaluieren, wird eine Simulationsstudie durchgeführt. Grund dafür ist, dass andernfalls keine validen Benchmarks über die wahre Verteilung der jeweils nicht gemeinsam beobachteten Merkmale, Einkommen (*Y*) und Konsum (*Z*) im untersuchten Fall, vorliegen. Dabei soll das Simulationsdesign so realitätsbezogen wie möglich bezüglich der begehrten Datenfusion von EU-SILC und HBS konzipiert werden. Dieses Kapitel stellt die dafür relevante Datengrundlage vor und skizziert Details der Monte-Carlo-Studie.

3.1 Datengrundlage

Als Datengrundlage dienen zugangsbedingt Public Use Files von EU-SILC aus dem Jahr 2013; für eine ausreichend große Datenbasis werden die Datensätze für Deutschland, Frankreich und die Niederlande gemeinsam betrachtet, was einer Simulationsgrundlage von $N = 25\,857$ Beobachtungen entspricht. Hieraus werden für jeden in Grafik 1 dargestellten Variablenblock (*X*, *Y*, *Z*) Variablen ausgewählt, die dem realen Fusionsszenario von EU-SILC und HBS nahe kommen. Zu den Konzepten der in diesem Abschnitt spezifizierten Variablen bietet die Dokumentation in Eurostat (2013) umfangreiche Informationen.

Als gemeinsame *X*-Variablen werden die von Eurostat hinsichtlich der Datenfusion von EU-SILC und HBS bereits für Deutschland, beispielsweise aufgrund von Verteilungsuntersuchungen anhand der Hellinger-Distanz, vorgeschlagenen Merkmale ausgewählt (Leulescu/Agafitei, 2013; Lamarche, 2017). ↘ **Übersicht 1** zeigt die entsprechenden *X*-Variablen einschließlich Wertebereich und Skalenniveau. Hierbei ist ersichtlich, dass die metrischen Variablen Alter und Einkommen für Random Hot-Deck kategorisiert werden müssen, während

Übersicht 1
Übersicht der gemeinsamen Variablen

Gemeinsame *X*-Variablen	Wertebereich/Skalenniveau	
	Eurostat	PMM
X_1: Activity Status of RP[1|2]	1 bis 7 / kategorial	1 bis 7 / kategorial
X_2: Age of RP[2]	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[1|3|4]	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[1]	1 bis 2 / kategorial	1 bis 2 / kategorial
X_7: Income	1 bis 5 / kategorial	gem. X_7 / metrisch

1 Zusätzlich bilden hier die fehlenden Werte eine Kategorie (kodiert als 9).
2 RP: „Reference Person" (befragte Person des gezogenen Haushalts).
3 Approximiert durch Variable „Dwelling Type".
4 Eigentlicher Wertebereich 1 bis 5, Kategorie 5 ist jedoch leer.
Quelle: Public Use Files der Erhebung EU-SILC 2013 für Deutschland, Frankreich und die Niederlande

bei PMM beide Variablen auf dem metrischen Skalenniveau verbleiben können.

Als spezifische *Y*- und *Z*-Variablen werden nun Stellvertretermerkmale ausgewählt, welche Einkommen und Konsum approximativ widerspiegeln sollen. Für Einkommen ist dies trivial, da die Datengrundlage auf EU-SILC basiert, worin detaillierte Einkommensmerkmale enthalten sind. Als Substitute für die Konsumvariablen muss hingegen auf ähnliche Variablengegebenheiten geachtet werden; deshalb ist es entscheidend, hierfür metrische Variablen auszuwählen. Die ausgewählten *Y*- und *Z*-Variablen zeigt ↘ Übersicht 2.

Übersicht 2

Übersicht der spezifischen Variablen für EU-SILC (*Y*) und Household Budget Survey (*Z*)

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: Public Use Files der Erhebung EU-SILC 2013 für Deutschland, Frankreich und die Niederlande

Eine inhaltlich exakte Überschneidung mit Konsum kann aufgrund der Datenbeschaffenheit nicht gewährleistet werden. Um die zugrundeliegenden Verfahren valide zu vergleichen, müssen jedoch Informationen über die gemeinsame Verteilung von ursprünglich nicht gemeinsam beobachteten Variablen vorliegen. Daher kann ein reales Fusionsszenario nicht als Performanceevaluation für Random Hot-Deck und PMM herangezogen werden. Dementsprechend ist hier simulationsorientiert vorzugehen und es scheint, dass die *Z*-Variablensubstitute für Konsum tatsächlich Konsuminformationen umfassten, obwohl dies rein logisch nicht der Fall ist. Dies dient jedoch dem Zweck der Evaluation der (sonst unbekannten) gemeinsamen Verteilung von *Y* und *Z*.

3.2 Monte-Carlo-Simulation

Die eben beschriebene Datengrundlage stellt die Hilfsdatenbasis für die Monte-Carlo-Simulationsstudie (Morris und andere, 2019) dar. Hieraus werden $k = 1\,000$ Zufallsstichproben (ohne Zurücklegen) gezogen. Anschließend wird für jeden Simulationszug das Datenfusionsmuster gemäß Grafik 1 erzeugt und die in Datenfile A (entspricht hier EU-SILC) fehlenden Z_1- und Z_2-Werte werden mittels beider Verfahren, Random Hot-Deck und PMM, entsprechend ergänzt.

Für die darauffolgende Evaluation jedes Simulationszugs sind besonders die Korrelationen zwischen den nicht gemeinsam beobachteten Merkmalen *Y* und *Z* im fusionierten Datenfile relevant. Denn Datenfusionen werden in der Regel zu genau diesem Zwecke durchgeführt, zur Erfassung und Analyse unbeobachteter Korrelationen jeweils nicht gemeinsam beobachteter Merkmale (Kiesl/Rässler, 2005). Die geschätzten Korrelationen der $k = 1\,000$ Zufallszüge werden anschließend jeweils mit den wahren Korrelationen[4] der Hilfsdatenbasis verglichen, um so die Performance beider Verfahren abschätzen zu können. Der Erhalt der bereits im Spenderdatensatz beobachteten Korrelationen zwischen den gemeinsamen Variablen und den spezifischen Konsumvariablen des HBS gilt als eine Art Mindestanforderung an eine Datenfusion und wird daher ebenso kurz beleuchtet. Um die Sensitivität beider Verfahren auf eine gleich- und übermäßige Anzahl an Beobachtungen des Spenderdatensatzes im Vergleich zu Beobachtungen des Empfängerdatensatzes abzuschätzen, wird die Monte-Carlo-Simulation sowohl mit einem gleichmäßigen Stichprobenverhältnis aus Empfänger- und Spenderbeobachtungen durchgeführt ($n_{silc} = n_{hbs} = 300$), als auch unter einem Stichprobenumfang, der neunmal so viele Spender- wie Empfängerbeobachtungen enthält ($n_{silc} = 300$ und $n_{hbs} = 2\,700$).

Die Monte-Carlo-Simulation wird mit dem Statistikprogramm *R* (R Core Team, 2019) durchgeführt. Die wichtigsten Packages sind StatMatch (D'Orazio, 2020) für Random Hot-Deck sowie BaBooN (Meinfelder/Schnapp, 2015) für Predictive Mean Matching. Das folgende Kapitel diskutiert die Ergebnisse der Monte-Carlo-Simulation.

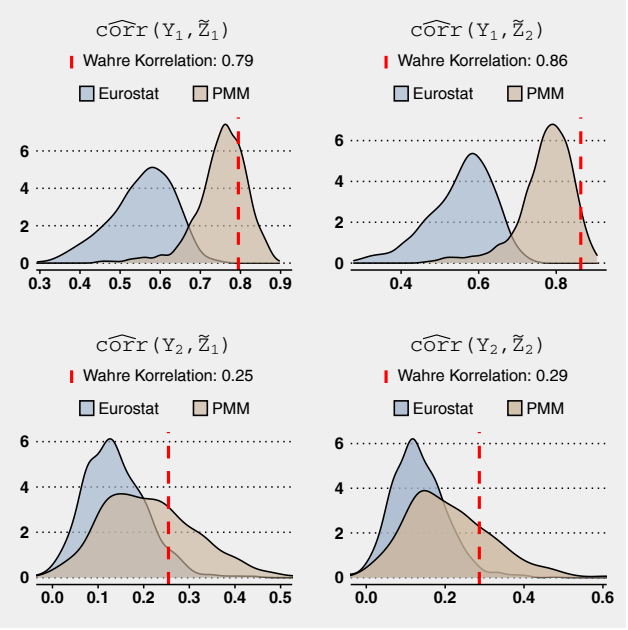
4 Hier ist zu beachten, dass die „wahre Korrelation" die Benchmark-Korrelation der in EU-SILC beobachteten Variablensubstitute für *Y* und *Z* widerspiegelt, nicht jedoch die tatsächliche, aber unbeobachtete Korrelation zwischen Einkommen und Konsum.

4

Ergebnisse der Monte-Carlo-Simulation

Die Verteilungen der jeweils $k = 1\,000$ geschätzten Korrelationen zeigen unter einem gleichmäßigen Verhältnis an Empfänger- und Spenderbeobachtungen, dass PMM die tatsächlichen Zusammenhänge zwischen den simulierten Einkommens- und Konsumvariablen deutlich besser reproduziert als das Random Hot-Deck-Verfahren. ↘ Grafik 2 Für hohe Originalkorrelationen (hier in Höhe von 0.79 und 0.86) optimiert PMM den Korrelationserhalt erheblich: Während beim Random Hot-Deck die (über alle $k = 1\,000$ Monte-Carlo-Simulationszüge) gemittelten Korrelationsschätzer 0.55 und 0.56 betragen, produziert PMM im Mittel Korrelationsschätzer von 0.75 und 0.77. Damit kommen diese relativ nah an die Originalzusammenhänge von 0.79 und 0.86 heran. Für mittlere Zusammenhänge in der Grundgesamtheit (hier in Höhe von 0.25 und 0.29) produziert PMM durchschnittliche Monte-Carlo-Korrelationen von 0.21 und ebenfalls 0.21, während die Korrelationsschätzer unter Random Hot-Deck im Mittel 0.14 und 0.13 betragen. Jedoch sind hier die Monte-Carlo-Varianzen für PMM höher und induzieren damit eine stärkere Unsicherheit, wenngleich PMM auch mittlere Originalkorrelationen besser abdeckt. Bei einer übermäßigen Anzahl an Spenderbeobachtungen zeigt sich ein kaum abweichendes Ergebnis, was auf eine geringe Sensitivität beider Verfahren mit Blick auf das Empfänger- und Spenderverhältnis hindeutet. ↘ Grafik 3

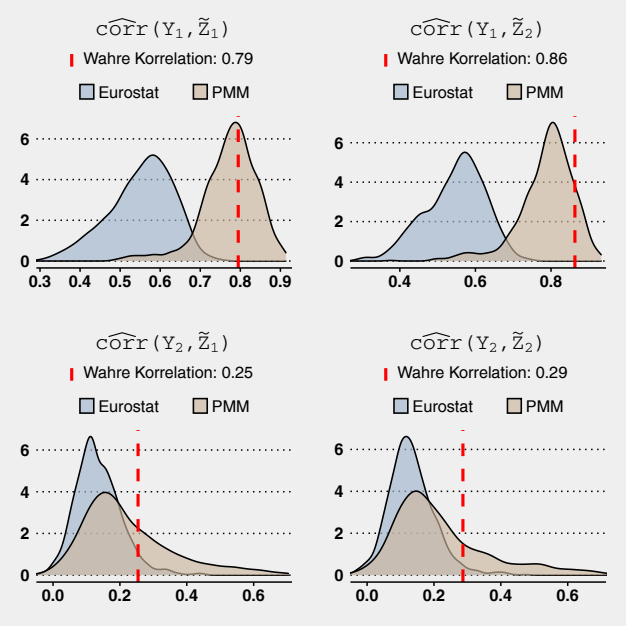
Grafik 2
Dichte der Korrelationen von *Y* und *Z* bei gleichmäßigem Empfänger- und Spenderverhältnis



Quelle: Public Use Files der Erhebung EU-SILC 2013 für Deutschland, Frankreich und die Niederlande
PMM: Predictive Mean Matching

2021 - 0284

Grafik 3
Dichte der Korrelationen von *Y* und *Z* bei übermäßiger Donorenanzahl

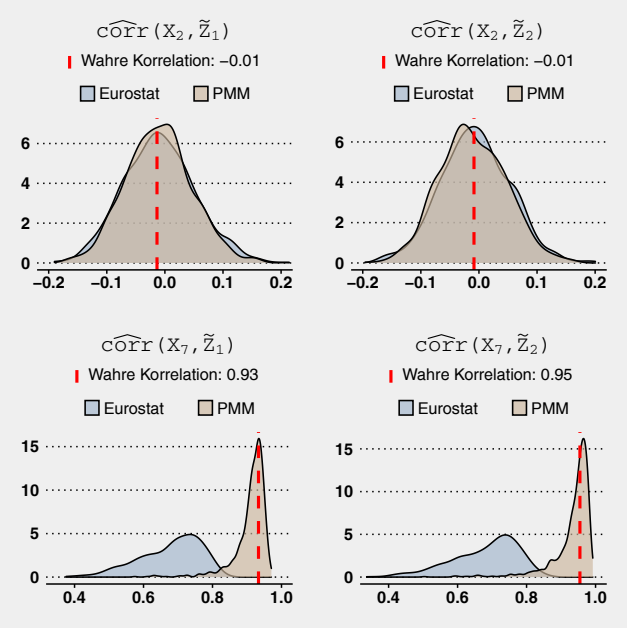


Quelle: Public Use Files der Erhebung EU-SILC 2013 für Deutschland, Frankreich und die Niederlande
PMM: Predictive Mean Matching

2021 - 0285

Die Reproduktion der bereits im Spenderdatensatz beobachteten Korrelationen zwischen *X* und *Z* stellt eine Art Mindestanforderung an eine Datenfusion dar (Kiesl/Rässler, 2005). Hier wird ersichtlich, dass PMM sowohl bei gleichmäßigem Empfänger- und Spenderverhältnis als auch bei einer übermäßigen Donorenanzahl dieser Mindestanforderung deutlich besser nachkommt als das Random Hot-Deck-Verfahren. Allerdings kann dies aufgrund der Datenbeschaffenheit lediglich für hohe Originalkorrelationen (hier in Höhe von 0.93 und 0.95) untersucht werden. Originalzusammenhänge nahe 0 (wie sie in diesem Fall in Höhe von –0.01 vorliegen) eignen sich hingegen für keinen Performancevergleich, da die Kor-

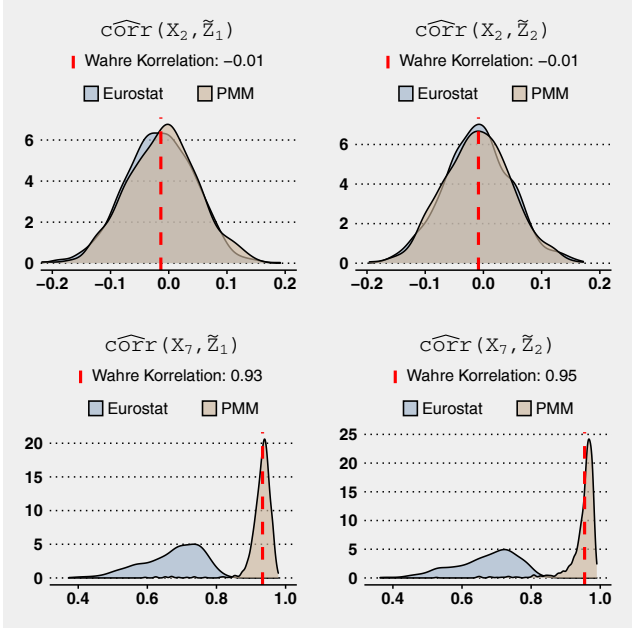
Grafik 4
Dichte der Korrelationen von *X* und *Z* bei gleichmäßigem Empfänger- und Spenderverhältnis



Quelle: Public Use Files der Erhebung EU-SILC 2013 für Deutschland, Frankreich und die Niederlande
PMM: Predictive Mean Matching

2021 - 0286

Grafik 5
Dichte der Korrelationen von *X* und *Z* bei übermäßiger Donorenanzahl



Quelle: Public Use Files der Erhebung EU-SILC 2013 für Deutschland, Frankreich und die Niederlande
PMM: Predictive Mean Matching

2021 - 0287

relationsschätzung eher einem Zufallsereignis gleicht. Deshalb erzielen hier Random Hot-Deck und PMM aggregiert relativ ähnliche Ergebnisse. ↘ Grafik 4, Grafik 5

5

Fazit

Für die validere und umfassendere Beurteilung des sozialen und wirtschaftlichen Lebensstandards sowie zur präziseren Abbildung von Armutsindikatoren strebt die amtliche Statistik die gemeinsame Betrachtung von Einkommen und Konsumausgaben privater Haushalte an. Beide Informationen sind über zwei Datensätze, EU-SILC und Household Budget Survey, verteilt, weshalb beide Datenquellen mittels einer geeigneten Datenfusionsmethode verbunden werden sollen. Die Ergebnisse der Simulationsstudie zeigen, dass die nicht gemeinsam beobachtete Verteilung von Einkommen und Konsum sowie im Besonderen deren Zusammenhangsstruktur mittels Predictive Mean Matching präziser reproduziert werden kann als mit der bisher verwendeten Random Hot-Deck-Methode.

Simulationsstudien können allerdings keine Allgemeingültigkeit gewährleisten. Dennoch ergibt sich aus den Ergebnissen für die amtliche Statistik die Erkenntnis, dass PMM zumindest für vergleichbare Datensituationen (also hinsichtlich einer Fusionierung von stetigen Variablen) eine vielversprechende Fusionsalternative zu herkömmlichen, nicht-parametrischen Verfahren darstellt. Eine Schwierigkeit in sämtlichen Datenfusionsszenarien ist hingegen die Annahme der bedingten Unabhängigkeit. Sie muss unterstellt werden, weil Datenfusionsverfahren (auch Random Hot-Deck und PMM) bedingte Unabhängigkeit im fusionierten Datenfile herstellen. Künftige Forschungsarbeiten könnten zur Bewältigung dessen die adäquate Einbeziehung von Hilfsinformationen aufgreifen.

Die Erkenntnisse des vorliegenden Beitrags kann die amtliche Statistik für künftige Datenfusionsvorhaben nutzen. Denn die amtliche Statistik ist heute mehr denn je bestrebt, lange Fragebogen zu vermeiden – zugunsten

ten der Entlastung der Auskunftgebenden und einer höheren Datenqualität. Dennoch soll es möglich sein, aus den Datenbeständen ein möglichst hohes Analysepotenzial zu generieren, um den steigenden amtlichen Daten- und Analysebedarf zu decken. Die Implementierung eines zielführenden Datenfusionsverfahrens spielt daher auch in Zukunft eine wichtige Rolle.

LITERATURVERZEICHNIS

D'Orazio, Marcello/Di Zio, Marco/Scanu, Mauro. *Statistical Matching. Theory and Practice.* 2006. [Zugriff am 6. Juli 2021]. Verfügbar unter: epdf.pub

D'Orazio, Marcello. *StatMatch: Statistical Matching or Data Fusion. R-Package. Version 1.4.0.* 2020. [Zugriff am 6. Juli 2021]. Verfügbar unter: cran.r-project.org

Eurostat. *Description of Target Variables: Cross-sectional and Longitudinal.* EU-SILC 065, 2013 operation (Version May 2013). 2013. [Zugriff am 6. Juli 2021]. Verfügbar unter: circabc.europa.eu

Fosdick, Bailey K./DeYoreo, Maria/Reiter, Jerome P. *Categorical Data Fusion Using Auxiliary Information.* In: The Annals of Applied Statistics. Jahrgang 10. Ausgabe 4/2016, Seite 1907 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.jstor.org

Gilula, Zvi/McCulloch, Robert E./Rossi, Peter E. *A Direct Approach to Data Fusion*. In: Journal of Marketing Research. Jahrgang 43. Ausgabe 1/2006, Seite 73 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: asu.pure.elsevier.com

Kiesl, Hans/Rässler, Susanne. *Techniken und Einsatzgebiete von Datenintegration und Datenfusion*. In: König, Christian/Stahl, Matthias/Wiegand, Erich (Herausgeber). Datenfusion und Datenintegration, 6. wissenschaftliche Tagung. Bonn 2005, Seite 17 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.iab.de

Kiesl, Hans/Rässler, Susanne. *How Valid Can Data Fusion Be?* In: IAB Discussion Paper. Ausgabe 15/2006. [Zugriff am 6. Juli 2021]. Verfügbar unter: doku.iab.de

Koller-Meinfelder, Florian. *Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. Dissertation. Otto-Friedrich-Universität Bamberg. 2009. [Zugriff am 6. Juli 2021]. Verfügbar unter: fis.uni-bamberg.de

Koschnick, Wolfgang J. *Standard-Lexikon für Markt- und Konsumforschung*. Band 1, A – K. München 1995.

Lamarche, Pierre. *Measuring Income, Consumption and Wealth jointly at the micro-level*. 2017. [Zugriff am 6. Juli 2021]. Verfügbar unter: ec.europa.eu

Leulescu, Aura/Agafitei, Mihaela. *Statistical matching: a model based approach for data integration*. In: Eurostat. Methodologies & Working papers. 2013. [Zugriff am 6. Juli 2021]. Verfügbar unter: ec.europa.eu

Little, Roderik J. A. *Missing-Data Adjustments in Large Surveys*. In: Journal of Business & Economic Statistics. Jahrgang 6. Ausgabe 3/1988, Seite 287 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.tandfonline.com

Meinfelder, Florian. *Datenfusion: Theoretische Implikationen und praktische Umsetzung*. In: Riede, Thomas/Bechtold, Sabine/Ott, Notburga (Herausgeber). Weiterentwicklung der amtlichen Haushaltsstatistiken. Auflage 1. 2013. Seite 83 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.konsortswd.de

LITERATURVERZEICHNIS

Meinfelder, Florian/Schnapp, Thorsten. *Package BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and Single Imputation for Discrete Data. R-Package. Version 0.2-0*. 2015. [Zugriff am 6. Juli 2021]. Verfügbar unter: cran.r-project.org

Morris, Tim P./White, Ian R./Crowther, Michael J. *Using simulation studies to evaluate statistical methods*. In: Statistics in Medicine. Jahrgang 38. Ausgabe 11/2019, Seite 1 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: onlinelibrary.wiley.com

Okner, Benjamin A. *Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File*. In: Annals of Economic and Social Measurement. Jahrgang 1. Ausgabe 3/1972, Seite 325 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.nber.org

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Wien 2019. [Zugriff am 6. Juli 2021]. Verfügbar unter: https://www.R-project.org/

Rässler, Susanne. *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York 2002.

Rodgers, Willard. L. *An Evaluation of Statistical Matching*. In: Journal of Business & Economic Statistics. Jahrgang 2. Ausgabe 1/1984, Seite 91 ff.

Rubin, Donald B. *Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations*. In: Journal of Business & Economic Statistics. Jahrgang 4. Ausgabe 1/1986, Seite 87 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.jstor.org

Serafino, Paola/Tonkin, Richard. *Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey*. In: Statistisches Amt der Europäischen Union. Statistical working papers. Doi: 10.2785/933460. 2017. [Zugriff am 6. Juli 2021]. Verfügbar unter: ec.europa.eu

Sims, Christopher A. *Comments (zu Okner 1972)*. In: Annals of Economic and Social Measurement. Jahrgang 1. Ausgabe 3/1972, Seite 343 ff.

Singh, A. C./Mantel, H. J./Kinack M. D./Rowe, G. *Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption*. In: Survey Methodology. Jahrgang 19. Ausgabe 1/1993, Seite 59 ff. [Zugriff am 6. Juli 2021]. Verfügbar unter: www150.statcan.gc.ca

Stiglitz, Joseph E./Sen, Amartya/Fitoussi, Jean-Paul. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. 2009. [Zugriff am 6. Juli 2021]. Verfügbar unter: www.economie.gouv.fr

Van der Putten, Peter/Kok, Joost N./Gupta, Amar. *Data Fusion Through Statistical Matching*. In: MIT Sloan School of Management. Working Paper 4342-02. 2002. [Zugriff am 6. Juli 2021]. Verfügbar unter: liacs.leidenuniv.nl

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im August 2021
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-21004-4, ISSN 1619-2907