Bewerbung auf den Gerhard-Fürst-Preis Wiesbaden, den 24.03.2020
Verfasser: Jannik Schaller

Kurzfassung der Masterarbeit

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

Motivation: Datenfusionen werden in der amtlichen Statistik immer häufiger Gegenstand wissenschaftlicher Untersuchungen, da sich interessierende Merkmale oft nicht in einer, sondern in verschiedenen Datenquellen wiederfinden und immer umfangreichere Befragungen aufgrund der Belastung der Auskunftgebenden sowie erhöhten finanziellen Kosten suboptimal sind. Ein Verknüpfen der Datenquellen mittels Record Linkage (also anhand eindeutiger Identifikatoren) wäre wünschenswert, ist jedoch mit Stichproben, die unterschiedliche Erhebungseinheiten beinhalten, sowie aufgrund rechtlicher Restriktionen für die amtliche Statistik in Deutschland kaum möglich. Dadurch geraten Datenfusionsverfahren immer stärker in den allgemeinen Interessen- und Forschungsfokus der amtlichen Statistik.
So auch im konkreten Anwendungsfall der vorliegenden Arbeit, die im Rahmen des EU-OECD-Projekts „Income, Consumption and Wealth" (kurz: ICW), worin von Seiten Deutschlands auch das Statistische Bundesamt involviert ist, entstand. Das ICW-Projekt hat in einem ersten Schritt zum Ziel, die Determinanten Einkommen und Konsumausgaben der privaten Haushalte gemeinsam zu betrachten. Da bisher keine einheitliche Datenbasis für eine gemeinsame Analyse der umfassenden Einkommens- und Konsumangaben der privaten Haushalte vorliegt, strebt die amtliche Statistik eine Fusionierung von EU-SILC und dem HBS, in Deutschland die Einkommens- und Verbrauchsstichprobe (EVS), an. Beide Datenquellen erfassen die interessierenden Determinanten besonders detailliert: EU-SILC die Einkommensangaben, der HBS wiederum vielfältige Konsumausgaben privater Haushalte. Konkret soll der EU-SILC-Datenbestand über die spezifischen Konsuminformationen des HBS-Datensatzes erweitert werden, weshalb EU-SILC den Empfängerdatensatz darstellt, dem die Konsummerkmale aus dem HBS hinzugefügt werden sollen. Der HBS wiederum fungiert als Spenderdatensatz, der die jeweiligen Konsuminformationen enthält. Die gegenwärtige Fusionsmethode des Statistischen Bundesamtes entspricht dem von Eurostat für das ICW-Projekt vorgeschlagenen Random Hot-Deck-Ansatz. Ziel der Arbeit ist es, die Random Hot-Deck-Fusionsmethode zu evaluieren, sie gegen ein alternatives Verfahren, Predictive Mean Matching (PMM), zu testen und dadurch, wenn möglich, das im ICW-Projekt bisher verwendete Verfahren zu optimieren.

Relevante Fusionsalgorithmen und Hypothese: In der Regel werden Datenfusionen, auch unter dem Terminus „Statistical Matching" bekannt, entlang ausgewählter, gemeinsamer Variablen, die in beiden Datensätzen erhoben wurden (zum Beispiel Alter, Beruf, etc.), durchgeführt, wobei auf Basis dieser gemeinsamen Merkmale über Distanzberechnungen möglichst ähnliche Beobachtungen in beiden Datensätzen ausfindig gemacht und verbunden werden sollen. Während beim derzeit verwendeten Random Hot-Deck-Verfahren alle zuvor (hier via Backward Selection) ausgewählten gemeinsamen Variablen mit dem gleichen Gewicht in die Distanzberechnung zwischen Beobachtungen aus EU-SILC und HBS miteingehen, ist der Vorteil von PMM, dass die gemeinsamen Variablen anhand ihres Erklärungsbeitrags auf die zu fusionierenden Merkmale (hier die Konsumausgaben) abgestuft werden. Je höher die Relevanz der gemeinsamen Variablen für die Erklärung der spezifischen Konsumvariablen, desto höher ist deren Gewicht bei der Distanzberechnung. Ebenso lässt das gegenwärtige Random Hot-Deck-Verfahren aufgrund der Kategorisierung sämtlicher metrischer Variablen nur Nulldistanzen (vermeintlich

„exakte“ Matches) zu, während PMM auf eine solche Kategorisierung verzichtet und den dadurch induzierten Informationsverlust vermeidet. Sowohl die Abstufung erklärungsrelevanter Variablen als auch das Vermeiden eines Informationsverlustes dürften eine präzisere Distanzmessung von PMM implizieren. Dementsprechend wird folgende Arbeitshypothese unterstellt: Da von PMM eine präzisere Distanzberechnung ausgeht, führt PMM zu einem besseren Fusionsergebnis als das gegenwärtige Random Hot-Deck-Verfahren.

Simulationsdesign und Evaluationskriterien: Die Überprüfung der Hypothese erfolgt mithilfe einer Simulationsstudie, da andernfalls keine validen Benchmarks über die wahre Verteilung der jeweils nicht gemeinsam beobachteten Merkmale Einkommen und Konsum vorliegen. Deshalb werden aus einer ausreichend großen Hilfsdatenbasis k = 1000 Zufallszüge im Rahmen einer Monte-Carlo-Simulation (kurz: MC-Simulation) gezogen. Anschließend wird für jeden Zufallszug das Datenausfallmuster einer Datenfusion künstlich generiert, sodass eine Fusionierung von EU-SILC und dem HBS unter Verwendung von Random Hot-Deck und PMM mit je zwei Einkommens- und Konsumvariablen simuliert und darauffolgend evaluiert werden kann. Für diese Evaluation sind besonders die Korrelationen zwischen den nicht gemeinsam beobachteten Merkmalen Einkommen und Konsum im fusionierten Datenfile relevant, denn Datenfusionen werden in der Regel zu genau diesem Zwecke durchgeführt, zur Erfassung und inferenzstatistischen Analyse unbeobachteter Korrelationen jeweils nicht gemeinsam beobachteter Merkmale. Die geschätzten Korrelationen der k = 1000 Zufallszüge werden anschließend jeweils mit den wahren Korrelationen 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, wurde die MC-Simulation nicht nur mit einem gleichmäßigen Stichprobenverhältnis aus Empfänger- und Spenderbeobachtungen durchgeführt, sondern ebenso unter einem Stichprobenumfang, der neunmal so viele Spender- wie Empfängerbeobachtungen beinhaltet.

Ergebnisse: Die Verteilungen der jeweils k = 1000 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, wie in der Abbildung dargestellt. Für hohe Originalkorrelationen (hier in Höhe von 0.79 und 0.86) optimiert PMM

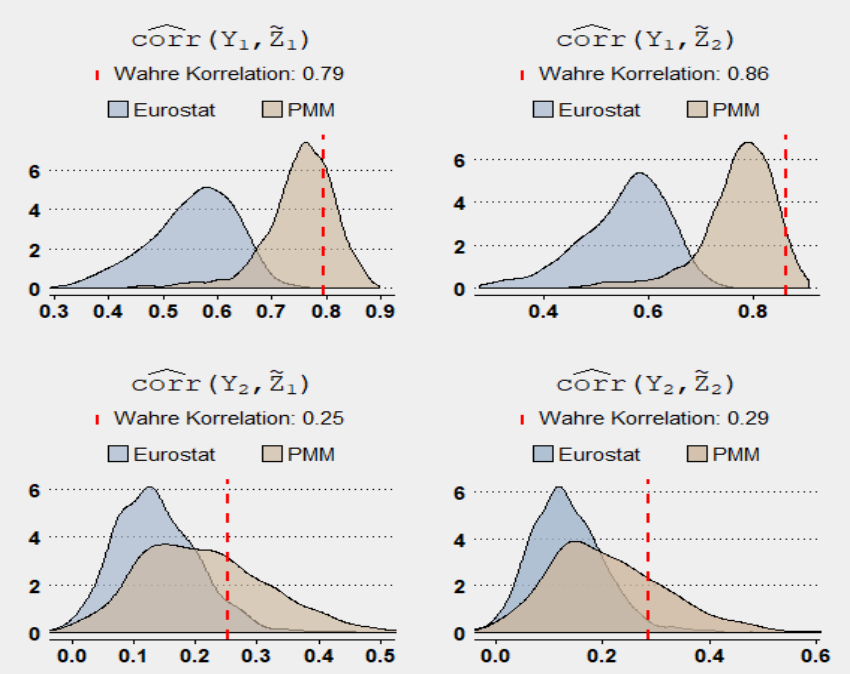

***Abbildung**: Dichte der Korrelationen bei gleichmäßigem Empfänger- und Spenderverhältnis*

den Korrelationserhalt erheblich: Während beim Random Hot-Deck die (über alle k = 1000 MC-Simulationszüge) gemittelten Korrelationsschätzer 0.55 und 0.56 betragen, produziert PMM im Mittel Korrelationsschätzer von 0.75 und 0.77, womit diese relativ nah an die Originalzusammenhänge von 0.79 und 0.86 herankommen. Für mittlere Zusammenhänge in der Grundgesamtheit (hier in Höhe von 0.25 und 0.29) produziert PMM durchschnittliche MC-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 MC-Varianzen für PMM höher und induzieren damit eine stärkere Unsicherheit, wenngleich PMM auch mittlere Originalkorrelationen besser abdeckt. Auch hinsichtlich der Reproduktion der bereits im Spenderdatensatz beobachteten Korrelationen wird ersichtlich, dass PMM dieser Mindestanforderung deutlich besser nachkommt als das Random Hot-Deck-Verfahren. 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.

Nutzen für die amtliche Statistik: Die Ergebnisse der vorliegenden Arbeit implizieren für die amtliche Statistik den Nutzen, dass die nicht gemeinsam beobachtete Verteilung von Einkommen und Konsum sowie im Besonderen deren Zusammenhangsstruktur präziser reproduziert werden kann. Dies ist von enormer Relevanz für die Validität der nachgelagerten Analysevorhaben im ICW-Projekt, die unter anderem eine präzisere Messung des sozialen und wirtschaftlichen Lebensstandards der privaten Haushalte in der EU sowie eine fundiertere Abschätzung von Armutsrisiken und Armutsindikatoren (etwa als Ergänzung oder Alternative zum AROPE-Indikator) umfassen. Ebenso werden der amtlichen Statistik darüber hinausgehende Handlungsperspektiven aufgezeigt, etwa die Einbeziehung von Hilfsvariablen, um Probleme mit der zentralen Annahme der bedingten Unabhängigkeit, die herkömmlichen Fusionsalgorithmen, auch Random Hot-Deck und PMM, zugrunde liegt, abzuschwächen oder Multiple Imputation für inferenzstatistische Analysen in Erwägung zu ziehen.
Zudem resultiert aus den Ergebnissen der vorliegenden Arbeit für die amtliche Statistik der generelle Nutzen, dass die gegenwärtig von Eurostat und dem Statistischen Bundesamt verwendete Fusionsmethode zumindest für vergleichbare Datensituationen, also hinsichtlich der Fusionierung stetiger, metrischer Merkmale, mit PMM optimiert werden kann. Diese Erkenntnisse sind für die amtliche Statistik gerade in der aktuellen Diskussion von enormer Relevanz. So ist die amtliche Statistik heute mehr denn je bestrebt, lange Fragebögen zugunsten der Entlastung der Auskunftgebenden und einer höheren Datenqualität zu vermeiden, aus ihren Datenbeständen aber dennoch ein möglichst hohes Analysepotential zu generieren, um den steigenden amtlichen Daten- und Analysebedarf zu decken. Die Implementierung eines zielführenden Datenfusionsverfahrens ist hierfür unabdingbar.