OZP

Research Article

Datenfusion in der sozialwissenschaftlichen Wahlforschung – Begründeter Verzicht oder ungenutzte Chance? Theoretische Vorüberlegungen, Verfahrensüberblick und ein erster Erfahrungsbericht

Johann Bacher[1,*] & Dimitri Prandner[1]

1 Institut für Soziologie, Johannes Kepler Universität Linz , Austria

* Johann.Bacher@jku.at.

Zusammenfassung

Datenarchive stellen umfangreiches Individualdatenmaterial in hoher Qualität zur Verfügung. Theoretisch böte sich damit die Möglichkeit für eine Datenfusion: Auf Ebene der Befragten wird ein neuer Datensatz erzeugt, der Variablen aus unterschiedlichen Datensätzen enthält. Dieses Potential wird in den Sozialwissenschaften aber kaum genutzt. Es existieren nur wenige Ausnahmen, in denen das Verfahren der Datenfusion zur Anwendung kommt.

Der Beitrag geht daher der Frage nach, ob dieser Verzicht begründet ist oder ob dadurch Chancen ungenutzt bleiben. Zur Beantwortung wird zunächst ein formales Modell entwickelt, das aufzeigt, unter welchen Bedingungen eine Datenfusion zu einer Verbesserung oder Verschlechterung der Datenqualität führen kann. Daran anschließend wird ein Überblick zu den Techniken der Datenfusion gegeben, bevor anhand eines konkreten Beispiels aus der Wahlforschung die Frage untersucht wird, ob Datenfusion mit einem der derzeit verfügbaren Standardstatistikpaketen möglich ist und zu welchen Ergebnissen sie führt.

Schlüsselwörter

Datenfusion, Record Linkage, Statistical Matching, Multiple Imputation, Wahlforschung

Data fusion in social science electoral research – justified absence or unused opportunity? Theoretical considerations, methods and preliminary empirical results

Abstract

Data archives provide a large number of high quality data sets. Therefore, the opportunity for data fusion exists theoretically. Data fusion generates a new respondent level dataset, containing variables coming from different donor datasets. However, apart from a few exceptions, this procedure is rarely used in the social sciences.

The article examines the question whether this absence is justified or whether opportunities remain unused. In a first step to answer this question a formal model is developed. This model outlines the conditions under which data fusion can lead to an improvement or deterioration of data quality. Subsequently, an overview of techniques of data fusion is given. The final section of the article uses an example from the field of electoral research to examine the question, whether it is possible to complete data fusion with a standard statistical package or not and discusses the results provided.

Keywords

Data fusion, record linkage, statistical matching, multiple imputation, electoral research

The authors have declared that no competing interests exist.

October 31, 2018 I *innsbruck* university press, Innsbruck
OZP – Austrian Journal of Political Science I ISSN 2313-5433 I http://oezp.at/
Vol. 47, issue 2 I DOI 10.15203/ozp.2025.vol47iss2

1. Problemstellung

Internationale Umfrageprogramme, wie der European Social Survey (ESS), der World Value Survey (WVS) oder das International Social Survey Project (ISSP), aber auch nationale Erhebungen, wie der Soziale Survey Österreichs (SSÖ) oder die Austrian National Election Study (AUTNES), generieren umfangreiche sozialwissenschaftliche Datenbestände, die ausführlich dokumentiert und über Datenarchive (z.B. DAS oder AUSSDA) frei für Sekundäranalysen verfügbar sind. Gängige Datenformate, ausführliche Handbücher sowie kontrollierte Datenqualität ermöglichen die Untersuchung einer Vielzahl an Themen, ohne dass eigene empirische Erhebungen notwendig sind. Die in den unterschiedlichen Datensätzen verfügbaren Informationen reichen dabei von Angaben zur Wohnsituation über individuelle Wertehaltung bis hin zur Freizeitgestaltung und politischen Orientierung. Hinzu kommen in Zukunft Geo- und Social-Media-Daten (Stichwort Big Data; Breur 2011).

Theoretisch böte sich somit die Möglichkeit für eine Datenfusion, bei der auf Ebene der Befragten ein neuer Datensatz erzeugt wird, der Variablen aus unterschiedlichen Datensätzen enthält (Denk/Hackl 2003, 311; Rässler 2004; Kiesl/Rässler 2006, 4). Im Unterschied zum Record Linkage auf Individual- (Personen) oder Aggregatebene (Parteien, Länder usw.) ist zu beachten, dass es sich bei der Datenfusion um Informationen unterschiedlicher Personen handelt, die verknüpft werden, während beim Record Linkage Angaben derselben Untersuchungseinheit in unterschiedlichen Datensätzen gesucht und verlinkt werden.

Abbildung 1: Problemstellung der Datenfusion

Datensatz 1 (Empfänger) zu imputierende Varialen X*	Variale, die sowohl im Datensatz 1 als auch im Datensatz 2 enthalten sind (gemeinsame Variablen Z)	Datensatz 1 (Spender) spezifische Variablen Y
⬆		⬇
Datensatz 2 (Spender) spezifische Variablen X		Datensatz 2 (Empfänger) zu imputierenden Variablen Y*

In dem Beispiel aus Abbildung 1 stammen die Variablen X aus dem Datensatz 2, die Variablen Y aus dem Datensatz 1. Die Variablen Z sind in beiden Datensätzen vorhanden. Datensatz 1 könnte z.B. eine Umfrage zu einer aktuellen Wahl sein, Datensatz 2 eine Umfrage zu einer vorausgehenden Wahl. Die gemeinsamen Variablen Z könnten sozio-demographische Eckdaten, Wertorientierungen und zentrale politische Einstellungen enthalten. Mittels Datenfusion könnten beide Datensätze um die fehlenden Informationen (Datensatz 1 um die Variablen X, Datensatz 2 um die Variablen Y) zur Durchführung von Längsschnittanalysen erweitert werden.

Ein anderer potentieller politikwissenschaftlicher Anwendungsfall wäre, wenn Datensatz 2 Variablen zum Medienkonsum während eines Wahlkampfes enthält und Datensatz 1 wiederum das berichtete Wahlverhalten erfasst. Nach einer erfolgreichen Datenfusion könnten mediale Einflussfaktoren auf das Wahlverhalten identifiziert werden.

Der Einsatz von Datenfusionen ist in der (kommerziellen) Markt- und Medienforschung durchaus üblich (Czaia 2005). In der Marktforschung beispielsweise waren methodische Experimente zur Datenfusion bereits in den späten 1960er Jahren üblich (Baker et al. 1989; Rässler 2002) und mit Ende der 1990er Jahre wurden fusionierte Datensätze zur Analyse der Werbewirkung genutzt (Rius et al. 1999; Breur 2011; Baker 2007). Heute wird Datenfusion in der Marketing-Forschung als sinnvolle Verknüpfungsmöglichkeit von Big-Data-Beständen unterschiedlicher Herkunft – z.B. von sozialen Netzwerken, Cloud-Diensten etc. – gesehen (Zerr et al. 2011; Breur 2011). Ein weiteres bekanntes Beispiel ist die MA-Intermedia aus dem Bereich der Mediennutzung, wo in Deutschland seit 1987 mit fusionierten Daten gearbeitet wird (Babic et al. 2011).

Außerhalb der Markt- bzw. Marketingforschung und der Kommunikationswissenschaft findet man im Bereich der amtlichen Statistik Beispiele und Anwendungen von Datenfusion. So werden beispielsweise in Italien (D'Orazio et al. 2001; D'Orazio et al. 2006; Conti et al. 2017) sowie in Kanada (Liu/Kovacevic 1997) seit längerem Datenfusionsverfahren zur Verknüpfung von Einkommens- und Konsuminformationen aus unterschiedlichen Erhebungen eingesetzt. Auch in Österreich werden von der Statistik Austria Daten fusioniert (Wegscheider-Pichler/Haslinger 2015). Ansonsten wird bisher Datenfusion in der sozialwissenschaftlichen Forschung nur vereinzelt angewandt (z.B. im Bildungsbereich bei Kaplan/McCarty 2013 oder in den Gesundheitswissenschaften Aluja-Banet et al. 2015).

Mit dem vorliegenden Beitrag wird am Beispiel der Wahlforschung der Frage nachgegangen, ob dieser implizierte Verzicht auf Datenfusionen in den Sozialwissenschaften berechtigt ist oder ob durch ihn Erkenntnischancen brachliegen.

Dafür wird zunächst in Abschnitt 2 ein formales Modell entwickelt, das eine Antwort darauf gibt, unter welchen Bedingungen eine Datenfusion zu einer Verbesserung oder Verschlechterung der Datenqualität führen kann. Abschnitt 3 gibt einen Überblick über Techniken der Datenfusion. Abschnitt 4 untersucht dann für ein konkretes Beispiel aus der Wahlforschung die Frage, ob

Datenfusion mit einem der derzeit verfügbaren Standardstatistikpaket möglich ist und zu welchen Ergebnissen sie führt. Der abschließende Abschnitt 5 fasst die gewonnenen Erkenntnisse und Einsichten zusammen. In drei ergänzenden Dokumenten (Bacher/Prandner 2018a, 2018b; Prandner/Bacher 2018) sind Details nachlesbar.

Die Darstellung wird sich dabei auf die Fusion von zwei Datensätzen konzentrieren, die entsprechend der Literatur (z.B. D'Orazio et al. 2001; Rässler 2004) als Empfänger („Recipient") und Spender („Doner") bezeichnet werden. Der Empfängerdatensatz R = (Y, Z) enthält die Variablen Y und Z, der Spenderdatensatz D = (X, Z) die Variablen X und Z. Die Variablen X und Y werden als spezifische Variablen bezeichnet, die Variable Z als gemeinsame Variable. In der Literatur wird für die gemeinsamen Variable noch die Bezeichnung überlappende Variablen („overlapping variables", z.B. Breur 2011) verwendet. Die hier präsentierten Überlegungen für die Fusion von zwei Datensätzen lassen sich problemlos auf die Fusion von mehr als zwei Datensätzen übertragen. Da die Datenfusion formal ein spezifisches Imputationsproblem von fehlenden Werten darstellt (siehe unten), wird im Folgenden auch von „imputierten" Variablen, von „imputieren" und von „Imputation" und „Imputationsmodellen" gesprochen.

2. Ein einfaches formales Modell zur Erklärung möglicher Genauigkeitsgewinne oder -verluste

Aus methodischer Sicht kann eine Datenfusion sowohl zu einem Genauigkeitsgewinn als auch zu einem -verlust führen (siehe Supplement 1, Prandner/Bacher 2018). Konkret wird nachfolgend untersucht, ob die Korrelation von zwei Variablen bei einer Datenfusion mit einer geringeren Abweichung (Genauigkeitsgewinn) oder einer größeren Abweichung (Genauigkeitsverlust) von der „wahren" Korrelation im Vergleich zu einer simultanen Erhebung beider Variablen geschätzt wird.

Das Modell nimmt an, dass die Variablen X und Y eine Funktion der gemeinsamen Variablen Z sind. Für die empirisch erhobenen Variablen X und Y soll für den Fall, dass beide in einer Befragung erfasst werden, gelten:

$$X = Z + \varepsilon_1$$
$$Y = \rho \cdot Z + \varepsilon_2$$

wobei ε_1 und ε_2 zufällige Messfehler mit Varianzen $VAR(\varepsilon_1)$ und $VAR(\varepsilon_2)$ sind. Die theoretische („wahre") Korrelation zwischen den Variablen ist ρ.

Unter der Annahme zufälliger Messfehler lässt sich die empirische Korrelation zwischen den Variablen X und Y darstellen als

$$COR(X,Y) = \frac{\rho}{\sqrt{1+VAR(\varepsilon_1)} \cdot \sqrt{1+VAR(\varepsilon_2)}},$$

wenn ohne Einschränkung der Allgemeinheit angenommen wird, dass die gemeinsame Variable Z standardisiert ist. Der Ausdruck

$$R_{XX} = \frac{1}{1+VAR(\varepsilon_1)} = 1 - \frac{VAR(\varepsilon_1)}{1+VAR(\varepsilon_1)}$$

lässt sich als Zuverlässigkeit der Messung von X interpretieren. Analog ist

$$R_{YY} = \frac{1}{1+VAR(\varepsilon_2)} = 1 - \frac{VAR(\varepsilon_2)}{1+VAR(\varepsilon_2)}$$

die Zuverlässigkeit von Y. Die Korrelation lässt sich damit bekanntlich (erstmals Spearman 1904) darstellen als

$$COR(X,Y) = r \times \sqrt{R_{XX} \times R_{YY}}.$$

Die Variable X soll nun aus einem anderen Datensatz imputiert werden. Zur Abgrenzung von der oben erörterten Situation verwenden wir für Y die Spezifikation Y*. Für Y* soll gelten:

$$Y^* = \rho \cdot Z + \varepsilon_2^*,$$

wobei ε_2* der zufällige Messfehler von Y* ist, wenn X nicht erhoben wird. Die Varianz dieses Zufallsfehlers $VAR(\varepsilon_2^*)$ kann gleich, größer (Reduktion der Datenqualität) oder kleiner (Erhöhung der Datenqualität) sein, wenn X und Y gemeinsam befragt werden. Für die imputierte Variable X* soll gelten:

$$X^* = Z + \delta,$$

wobei δ der durch die Imputation bedingte Fehler ist. Unter der Annahme, dass es sich bei dem Imputationsfehler um einen Zufallsfehler handelt, nimmt die Korrelation zwischen X* und Y* folgenden Wert an:

$$COR(X^*,Y^*) = \frac{\rho}{\sqrt{1+VAR(\delta_1)} \cdot \sqrt{1+VAR(\varepsilon_2^*)}} = \rho \cdot \sqrt{R^2_{X^*/Z} \cdot R_{Y^*Y^*}},$$

wobei $R^2_{X^*/Z}$ die durch die gemeinsamen Variablen erklärte Varianz in X ist. R_{Y^*/Y^*} ist wiederum die Zuverlässigkeit der Messung von Y*.

Ein Genauigkeitsgewinn mit der Datenfusion lässt sich somit dann erzielen, wenn gilt:

$$R^2_{X/Z} \times R_{Y^*Y^*} > R_{XX} \times R_{YY}.$$

Diese Bedingung wäre z.B. erfüllt, wenn die durch das Imputationsmodell erklärte Varianz in etwa der Messgenauigkeit (Zuverlässigkeit der Messung) von X ($R^2_{X/Z} \approx R_{XX}$) entspricht und sich die Messqualität von Y ($R_{Y^*Y^*} > R_{YY}$), z.B. durch Einsatz einer Langfassung einer Skala zur Messung von Y oder durch Vermeidung eines Konzentrationsabfalls, (leicht) erhöht. Bei gleichbleibender Messgenauigkeit von Y ($R_{Y^*Y^*} \approx R_{YY}$) müsste für

einen Genauigkeitsgewinn die erklärte Varianz größer sein als die Messgenauigkeit von X ($R^2_{X/Z} > R_{XX}$).

Wenn dagegen nur ein sehr schlechtes Imputationsmodell mit einer hohen Imputationsfehlervarianz vorliegt, würde ein deutlicher Genauigkeitsverlust die Folge sein. Beispiel: Eine empirische Korrelation COR(X,Y) von 0,40 ("wahre" Korrelation $\varphi = 0{,}67$) würde sich auf 0,16 reduzieren, wenn die Zuverlässigkeit der Messung für alle Variablen 0,60 beträgt, das Imputationsmodell aber nur 0,10 erklären würde, da gilt:

$$COR(X^*, Y^*) = 0{,}16 = 0{,}67 \times \sqrt{0{,}1 \times 0{,}6} < COR(X, Y) = 0{,}40 = 0{,}67 \times \sqrt{0{,}6 \times 0{,}6}.$$

Entscheidend ist somit, dass ein Imputationsmodell mit einer geringen Fehlerkomponente bzw. einer hohen erklärten Varianz spezifiziert werden kann. Als Faustregel lässt sich festhalten, dass die durch das Imputationsmodell erklärte Varianz in den zu imputierenden Variablen in etwa so hoch sein sollte wie die vermutete Zuverlässigkeit der Messung der zu imputierenden Variablen. Vor diesem Hintergrund ist nachvollziehbar, dass für die Datenimputation erklärte Varianzen von 0,50 bzw. sogar von 0,60 gefordert werden (Cielebak/Rässler 2014; Rässler 2004), da dies in etwa den Schwellenwerten von Cronbachs α entspricht (George/Mallery 2010, 231f).

3. Techniken der Datenfusion

Wie bereits ausgeführt, lässt sich das Vorgehen bei der Datenfusion als Imputationsproblem auffassen. Die im Empfängerdatensatz R = (Y, Z) fehlenden Variablen X sollen auf der Grundlage der gemeinsamen Variablen Z im Spenderdatensatz D = (X, Z) geschätzt werden. Erfolgt die Schätzung in Richtung der Empfängerdatei R wird von einer asymmetrischen Datenfusion gesprochen. Der Empfängerdatensatz R wird um die geschätzten Variablen X* erweitert zu R* = (X^*, Z, Y). Die Schätzung kann auch in beide Richtungen erfolgen mit dem Ziel, zwei erweitere Datensätze R* = (X^*, Z, Y) und D* = (X, Z, Y*) zu genieren. In diesem Fall wird von einer symmetrischen Datenfusion gesprochen. Unter technischen Gesichtspunkten ist diese Unterscheidung irrelevant. R und D lassen sich beliebig vertauschen. Daher wird hier nur der Fall der klassischen Datenfusion der Variablen X in die Empfängerdatei R behandelt.

Für die Datenfusion, eignen sich prinzipiell alle für die Imputation entwickelten Verfahren (z.B. Enders 2010; Graham 2012; Yucel 2011). Es lassen sich zwei grundlegende Ansätze unterscheiden (siehe Tabelle 1):

- **Fallorientierte bzw. implizite / nicht-parametrische Verfahren.** Bei diesen Verfahren wird hinsichtlich der gemeinsamen Variablen Z für jeden Fall der Empfängerdatei nach einem oder mehreren ähnlichen Fällen in der Spenderdatei gesucht, für die gelten soll: $d(Z)_{i,j} \rightarrow \min$, wobei $d(Z)$ eine Distanzfunktion für $i \in R = (Y, Z)$ und $j \in D = (X, Z)$ ist. Die Variablenwerte in den spezifischen Variablen X des Falles j (oder der Fälle j) aus der Spenderdatei werden anschließend als Schätzwerte für die fehlenden Variablenwerte des Falles i in der Empfängerdatei unter Beachtung bestimmter Rahmenbedingungen (siehe z.B. Babic et al. 2011) verwendet. Mitunter wird vorab eine Segmentierung bzw. Schichtung vorgenommen, damit gewährleistet wird, dass in den Segmentierungsvariablen (z.B. Geschlecht, Bundesland) eine vollständige Übereinstimmung vorliegt. Angenommen wird, dass Fälle, die sich in den gemeinsamen Variablen Z nicht unterscheiden, auch hinsichtlich der Variablen X ähnlich sind. Dieser Gruppe gehören Verfahren an, die auf das Statistical Matching und/oder auf Clusteranalysen zurückgreifen, wobei zu beachten ist, dass das in der Kausalforschung oft verwendete Propensitiy-Score-Matching nicht geeignet ist (Cielebak/Rässler 2014, 380f).
- **Variablenorientierte bzw. modellbasierte Verfahren.** Bei diesen Verfahren wird nicht nach ähnlichen Fällen gesucht, sondern in der Spenderdatei wird ein funktionaler Zusammenhang $X = f(Z)$ zwischen den spezifischen Variablen X und den gemeinsamen Variablen Z definiert und geschätzt. Die ermittelte Funktion wird anschließend zur Schätzung von Werten der spezifischen Variablen X in der Empfängerdatei verwendet. Es wird angenommen, dass in der Empfängerdatei derselbe funktionale Zusammenhang besteht wie in der Spenderdatei. Dieser Gruppe gehören die nachfolgend beschriebenen Verfahren an.

Für beide Ansätze und die ihnen zugehörenden Verfahren sind in der Literatur unterschiedliche Bezeichnungen gebräuchlich. In beiden Ansätzen kann die Fusion nur einmal („single") oder mehrmals („multiple") erfolgen, was in m Datensätzen ($m = 1$ für „single" und $m > 1$ für „multiple") resultiert (siehe Tabelle 1).

Von den dargestellten Verfahren ist aus unserer Sicht die variablenorientierte multiple Imputation zu bevorzugen, da sie von weniger und schwächeren Annahmen ausgeht. Variablenorientierte Verfahren treffen nämlich nur die Annahme, dass die funktionalen Zusammenhänge stabil sind, was i.d.R. der Fall ist, während bei den fallorientierten Verfahren wie bei der Clusteranalyse (Bacher et al. 2010, 195–232) die wesentlich schwieriger zu erfüllende Annahme getroffen wird, dass die Variablenwerte in den gemeinsamen Variablen weitgehend fehlerfrei gemessen werden. Zudem muss bei fallorientierten Verfahren der Anwender/die Anwenderin Gewichte für die gemeinsamen Variablen definieren (D'Orazio et al. 2006, 167–170) und ein geeignetes Distanz- oder Ähnlichkeitsmaß auswählen. Diese Ent-

Tabelle 1: Übersicht: erfahren der Datenfusion[a)]

Zugang zur Datenfusion	*Single* *(Fehlende Werte in den spezifischen Variablen werden nur einmal geschätzt.)*	*Multiple* *(Fehlende Werte in den spezifischen Variablen werden mehrfach geschätzt.)*
Fallorientiert (implizite Verfahren, distance based Verfahren), zentrale Annahme: Ähnliche Fälle in den gemeinsamen Variablen sind auch in den spezifischen Variablen ähnlich.)	Statistical Matching (z.B. Baker 2007)[b)] Nächste-Nachbarn-Verfahren Hot-Dock-Techniken (Enders 2010, 49-50) Clusteranalyse (Babic et al. 2011)	Statistical Matching[b)] häufig in Kombination mit Re-Sampling-Techniken (Yucel 2011, Van der Putten et al. 2002)
Variablenorientiert (modellbasierte (explizite) Verfahren); zentrale Annahme: Funktionale Zusammenhänge zwischen den spezifischen Variablen und den gemeinsamen Variablen in der Spenderdatei gelten auch in der Empfängerdatei.)	Single Imputation (z.B. Enders 2010, 44-48)	Multiple Imputation (z.B. Enders 2010, 187-216)

a) Synonyme Bezeichnungen von „Datenfusion" in der Literatur sind „Statistical Matching" (D'Orazio et al. 2006) oder „Datenintegration" (Baker 2007; Bleiholder/Naumann 2008) usw.

b) D'Orazio et al. (2006) verwenden „Statistical Matching" als Oberbegriffe im Sinne des hier verwendeten Begriffs „Datenfusion". Die hier als „Statistical Matching" bezeichneten Verfahren werden von den Autoren unter der Bezeichnung „Nonparametric Micro Approach" zusammengefasst. Weitere synonyme Bezeichnungen sind „distanced based" Verfahren (Breur 2011).

scheidungen können zu schwer erklärbaren Ergebnisse führen (siehe dazu z.B. Kim et al. 2004). Bei den variablenorientierten Verfahren werden dagegen die Variablen aufgrund des gewählten statistischen Modells nach formalen Kriterien „automatisch" gewichtet. Die Definition eines Distanzmaßes ist nicht erforderlich. Daher sind aus unserer Sicht variablenorientierte Verfahren zu bevorzugen. Innerhalb dieser sollte ein multipler Ansatz gewählt werden, da bei ihm die Unsicherheiten der statistischen Schätzung mitberücksichtigt werden. In diesem Sinn sind multiple Verfahren statistisch „korrekter". Hinzukommt, dass multiple fallorientierte Imputationsverfahren in den Standardstatistikprogrammen IBM-SPSS (Modul Multiple Imputation), STATA (Modul MI) und SAS (Modul MI) verfügbar sind. Erwähnt werden muss aber, dass sich in der Praxis auch fallorientierte Verfahren bewährt haben, wie etwa in der deutschen Mediennutzungsstudie MA-Intermedia (Babic et al. 2011). Nach Saporta (2000) führen fallorientierte Verfahren zu besseren Ergebnissen, wenn mit der Datenfusion die Kovarianzen geschätzt werden sollen, während modellbasierte Verfahren bei Schätzung von Individualwerten besser abschneiden. Auf der Grundlage mehrerer Fusionsexperimente ziehen Soong/Montigny (2004) den Schluss, dass eine allgemeine Methodenempfehlung nicht möglich ist. Baker (2007) spricht dagegen modellbasierten Verfahren eine größere Genauigkeit zu.

Neben den hier dargestellten Verfahren, bei denen auf Personenebene fusioniert wird, gibt es sogenannte „Makroansätze", bei denen „nur" die Momente erster und zweiter Ordnung (Mittelwerte, Varianzen und Kovarianzen) geschätzt werden (D'Orazio et al. 2006).

Nachfolgend soll das Vorgehen bei einer variablenorientierten multiplen Imputation skizziert werden. Zunächst wird in der Spenderdatei ein funktionaler Zusammenhang zwischen den zu imputierenden Variablen X und den gemeinsamen Variablen Z spezifiziert:

$$X = f(Z;\theta;\delta),$$

wobei die Funktion $f(Z;\theta;\delta;)$ von den zu schätzenden Populationsparametern θ abhängt. δ ist eine Zufallsvariable, die zusätzlich die Unsicherheit der Schätzung für einen Befragten/eine Befragte erfasst. Handelt es sich beispielsweise bei X um eine intervallskalierte Variable, dann kann für die Imputation eine lineare Regressionsgleichung spezifiziert werden:

$$X = \beta_0 + \beta_1 Z_1 + \ldots + \beta_p Z_p + \delta \cdot$$

Die unbekannten Parameter umfassen in diesem Beispiel die Regressionskoeffizienten β sowie die Verteilungsparameter der Zufallsfunktion δ (bei der linearen Regression eine Normalverteilung mit Erwartungswert 0 und Varianz σ^2_δ, bei der logistischen Regression eine im Intervall 0,1 gleichverteilte Zufallszahl).

Auf der Grundlage der geschätzten Modellparameter werden die nicht vorhandenen Variablenwerte in X, die auch als plausible Variablenwerte bezeichnet werden, geschätzt mit:

$$X^* = f(Z;\hat{\theta};\hat{\delta})$$

bzw. im Falle einer linearen Regression als

$$X^* = \hat{\beta}_0 + \hat{\beta}_1 Z_1 + \ldots + \hat{\beta}_p Z_p + \hat{\delta}.$$

Bei einer multiplen Imputation wird die Schätzung mehrfach vorgenommen. Dabei wird die Tatsache berücksichtigt, dass die Parameterschätzungen selbst Schwankungen unterliegen. Für jeden Befragten/jede Befragte wird nicht ein Schätzwert in den zu imputierenden Variablen ermittelt, sondern mehrere Schätzwerte. Dies führt dazu, dass nach erfolgreicher Durchführung m Datensätze vorhanden sind, wobei m die Zahl der vorgenommenen Wiederholungen ist. Bezüglich der Zahl der Wiederholungen m empfiehlt Enders (2010, 213) auf der Basis einer Simulationsstudie von Graham et al. (2007) die Verwendung von 20 Imputationen. Mit Vorgriff auf unser Anwendungsbeispiel ist aber anzumerken, dass eine noch größere Zahl an Wiederholungen zu empfehlen ist.

Die Schätzung der plausiblen Werte erfolgt mittels Bayes-Verfahren (für eine Übersicht siehe Cielebak/Rässler 2014 oder Conti et al. 2017), wobei zwei Modellansätze zur Schätzung (Yucel 2011) unterschieden werden. Bei der sogenannten „joint estimation"-Methode werden die Parameter des funktionalen Zusammenhangs für alle zu imputierenden Variablen, also für X_1, X_2 usw., simultan in einem einzigen Schritt geschätzt. Beim „variable-to-variable"-Ansatz erfolgt die Schätzung schrittweise für jede Variable getrennt, also zunächst für X_1, dann für X_2 usw. Der „variable-to-variable"-Ansatz wird oft dahingehend kritisiert, dass die Schätzergebnisse von der Anordnung der Variablen abhängen können. In der Praxis hat sich dieser Ansatz aber – insbesondere bei Variablen mit gemischtem Messniveau – bewährt (Yucel 2011).

Die Schätzung der plausiblen Werte innerhalb der beiden genannten Methoden kann sequentiell – es wird mit einem Startwert gearbeitet und eine lange Schätzkette erzeugt, von der jede x-te Schätzung verwendet wird – oder parallel mit unterschiedlichen Startwerten erfolgen.

4. Ein Anwendungsbeispiel – Nationalratswahl 2013 und Bundespräsidentschaftswahl 2016

Die bei der Anwendung der Datenfusion erforderlichen Schritte (siehe Abbildung 2) werden im Supplement 2 (Bacher/Prandner 2018a) ausführlich dargestellt und nachfolgend anhand eines konkreten Beispiels aus der Forschung beschrieben. Die Vorablektüre von Supplement 2 ist daher hilfreich. Die Datenfusion verlangt derzeit vom Anwender/von der Anwenderin Entscheidungen, für die klare und eindeutige Empfehlungen fehlen, sodass eine weitgehend automatische Nutzung (noch) nicht möglich ist.

Abbildung 2: Anwendungsschritte bei der variablenorientierten Datenfusion mittels multipler Imputation

Schritt	*Aufgabe*	*Voraussetzungen bzw. Verfahren*
1	Auswahl eines geeigneten Spenderdatensatzes	• Spenderdatei sollte Stichprobe aus derselben Population sein und muss gemeinsame Variablen mit hoher Prognosekraft enthalten. • Der funktionale Zusammenhang zwischen den gemeinsamen Variablen Z und den zu imputierenden Variablen X in der Spenderdatei muss auch in der Empfängerdatei gelten.
2	Auswahl und Bildung geeigneter gemeinsamer Variablen	• Gemeinsame Variable müssen in beiden Datensätzen vergleichbar sein. • Entscheidung, ob mit ursprünglichen oder abgeleiteten Variablen gerechnet wird. • Methoden zur Bildung abgeleiteter Variablen: Hauptkomponentenanalyse oder Korrespondenzanalyse bei gemischten Merkmalen
3	Spezifikation und Schätzung eines geeigneten Datenfusions- bzw. Imputationsmodell	• Auswahl einer geeigneten Regressionsmethode • Prüfung, ob Regressionsmodell zur Imputation ausreichende Erklärungskraft besitzt. Falls nicht, Hinzunahme weiterer Variablen.
4	Anwendung des Imputationsmodells zur Schätzung der plausiblen Werte	• Spezifikation von mindestens 20 Imputationen • Größere Zahl an Imputationen wünschenswert

5	Validitätsprüfung bzw. Evaluation der Datenfusion	Folgende Strategien werden vorgeschlagen: • Prüfung des Vorliegens der lokalen bzw. bedingten Unabhängigkeit • Prüfung des Erhalts der Zusammenhänge zwischen gemeinsamen und imputierten Variablen in der Empfängerdatei • Prüfung der Annahme der Strukturerhaltung zwischen den gemeinsamen Variablen • Kriterienbezogene bzw. externe Validitätsprüfung / Evaluation
6	Datenanalysen zur Beantwortung der Forschungsfrage	• Entsprechend üblichem Vorgehen • Auswahl geeigneter statistischer Methoden • Prüfung der lokalen Unabhängigkeit von gefundenen Zusammenhängen

Erörterungen siehe Supplement 2 (Bacher/Prandner 2018a) sowie Text

4.1 Untersuchte Fragestellung und Datenbasis

Im Rahmen des Sozialen Surveys Österreich (SSÖ) 2016 (Bacher/Prandner 2017) sollte auch der Frage nachgegangen werden, ob eine Datenfusion in den Sozialwissenschaften mit Standardstatistikprogrammen möglich ist und ob sie – im Idealfall mit einem allgemeinen Imputationsmodell - zu brauchbaren Ergebnissen führt. Dazu wurden unterschiedliche inhaltliche Fragestellungen spezifiziert, u.a. die für diesen Beitrag ausgewählte Frage, für welchen Bundespräsidentschaftskandidaten im Jahr 2017 die Wähler der letzten Nationalratswahl 2013 votierten. Der Fokus wurde dabei auf FPÖ- und Grünen-Wähler_innen gelegt, da hier eine unmittelbare Plausibilitätsprüfung möglich ist: Die FPÖ-Wähler_innen der letzten Nationalratswahl müssten signifikant häufiger Hofer gewählt haben als Van der Bellen und umgekehrt die Grünen-Wähler_innen signifikant häufiger Van der Bellen. Die Information zur Bundespräsidentenwahl stand im SSÖ zur Verfügung, jene zur Nationalratswahl 2013 wurde im European Social Survey (7. Welle, ESS 2015) erfasst (siehe Tabelle 2).
Beide Datensätze sind repräsentativ für die österreichische Bevölkerung ab 16 Jahren. Die Anforderung, dass die beiden Datenquellen Stichproben aus derselben Grundgesamtheit sind, ist erfüllt.

Tabelle 2: Spezifische Variablen des Anwendungsbeispiels

Datensatz	*Variable*	*Variablenbezeichung / Inhalt*
ESS	X1	Berichtetes Wahlverhalten bei der Nationalratswahl 2013
SSÖ	Y1	Berichtetes Wahlverhalten bei der ersten Runde der Bundespräsidentschaftswahl 2016
SSÖ	Y2	Berichtetes Wahlverhalten bei der zweiten Runde der Bundespräsidentschaftswahl 2016

4.2 Auswahl gemeinsamer Variablen

Entsprechend der allgemeinen Zielsetzung des Forschungsprojekts wurde zunächst die Entwicklung eines allgemeinen Imputationsmodells, das eine vollständige Fusion des ESS in den SSÖ erlaubt, angestrebt, um die Daten möglichst vielen Nutzer_innen unterschiedlicher Fachdisziplinen verfügbar zu machen. Dafür wurden acht soziodemographische Items und zehn Items des PVQ – Portrait Value Questionaire – von Schwartz et al. (2001) als gemeinsame Variable herangezogen. Die Verwendung dieser Variablen ist in der empirischen Sozialforschung Standard. Von ihnen wird angenommen, dass sie zur Prognose unterschiedlicher Variablen (Einstellungen und Verhaltensweisen in unterschiedlichen Lebensbereichen) geeignet sind. Da sich dieses allgemeine Modell als zu wenig valide erwies (siehe unten), wurden zusätzlich spezifische Imputationsmodelle entwickelt, in denen die Einstellung zur Immigration, die Links-Rechts-Einstufung und die Parteipräferenz hinzugenommen wurden.

Anzumerken ist, dass aus methodischer Perspektive keine Richtlinien ableitbar sind, welche und wie viele Variablen (Faktenfragen, Einstellungsitems oder Werte) zur Anwendung kommen sollten. Formal wichtig ist, dass Variablen mit einer hohen Prognosekraft ausgewählt werden. Die verwendeten Variablen müssen vergleichbar sein, d.h. sie müssen entweder in beiden Datensätzen in identer Form gemessen werden oder auf eine einheitliche Kodierung transformierbar sein.

Zur Datenfusion wurden mittels einer multiplen Korrespondenzanalyse abgeleitete Variablen gebildet und alle Komponenten mit Eigenwerten größer 1 in die weiteren Analysen einbezogen. Durch die Verwendung von abgeleiteten Variablen (Details siehe Supplement 2, Bacher/Prandner 2018a) anstelle der direkt erhobenen Variablen können Schätzprobleme (Mulitkolinearität, fehlende Werte, Rechenaufwand) vermieden werden. Zudem ist es möglich, auf zur Fusion verwendete Variablen in späteren Analysen zurückzugreifen.[1]

4.3 Spezifikation und Auswahl eines geeigneten Datenfusions- bzw. Imputationsmodells

Entsprechend der Überlegungen im Abschnitt 3 wurde festgelegt, ein variablenorientiertes multiples Imputationsmodell einzusetzen. Da die zu imputierende Variable „Berichtetes Wahlverhalten bei der letzten Nationalratswahl" nominalskaliert ist, wurde als geeignetes Regressionsmodell die multinominale logistische Regression spezifiziert. Die Ergebnisse der Regressionsanalysen sind in Tabelle 3 zu finden, wobei hier sowohl die Pseudo-R^2-Werte nach Nagelkerke als auch jene nach McFadden wiedergegeben werden. Mit Rückgriff auf die Simulationsstudien von Smith/McKenna (2013) unterschätzen beide Maßzahlen R^2 aus der linearen Regression, wobei das Pseudo-R^2-Wert nach Nagelkerke das R2 aus der linearen Regression besser abbildet.

Mit Bezug auf die in Abschnitt 3 angeführte Faustregel, liegt das Pseudo-R^2 für das allgemeine Modell unter dem Schwellenwert von 0,5, während dieser Schwellenwert durch das spezifische Imputationsmodell 1 geringfügig und durch das spezifische Imputationsmodell 2 mit einem Pseudo-R^2-Wert nach Nagelkerke von 0,813 deutlich überschritten wird.

Problematisch erscheint für die vorliegende Forschungsfrage die schlechte Prognosequalität des allgemeinen Modells bezüglich der FPÖ-Wähler_innen mit knapp 4%. Da ein Fokus auf den FPÖ-Wähler_innen liegt, muss mit Vorgriff auf die nachfolgend berichteten inhaltlichen Ergebnissen die Prognosekraft des allgemeinen Modells als nicht ausreichend betrachtet werden. Die Ergebnisse legen ein spezifisches Modell zur Schätzung von den zu imputierenden X-Werten nahe. Besonders gut schneidet das spezifische Imputationsmodell 2 ab, das auch die Parteipräferenz als gemeinsame Variable einbezieht.

4.4 Anwendung des Imputationsmodells zur Schätzung der plausiblen Werte

Die Datenfusion wurde mit der multiplen Imputationsfunktion von IBM-SPSS 23 durchgeführt und entsprechend der Literaturempfehlung wurden 20 Schätzungen gerechnet. Die mittels Korrespondenzanalyse gebildeten gemeinsamen Variablen wurden als unabhängige Variablen genutzt, die zu imputierende Variable als abhängige.

Wir haben die Imputation mehrfach gerechnet. Dabei wurden für das allgemeine Imputationsmodell immer leicht abweichende Ergebnisse erzielt. Um stabilere Ergebnisse zu generieren, erscheint eine größere Zahl erforderlich.

Mit IBM-SPSS 23 ist es zwar möglich, Daten aus der multiplen Imputation zu analysieren. Allerdings

Tabelle 3: Modellgüte der verwendeten Imputationsmodelle

Kenngröße	*Allgemeines Modell*	*Spezifisches Modell I (Allgemeines Modell + Links/ Rechtsskala + Einstellung zu Migration)*	*Spezifisches Modell II (Allgemeines Modell + Links/ Rechtsskala + Einstellung zu Migration + Parteipräferenz*
Nagelkerkes R^2	0,448	0,545	0,813
McFaddens R^2	0,154	0,203	0,422
Korrekte Schätzung in %	39,7	44,4	63,3
Korrekte Schätzung der FPÖ-Wähler_innen in %	3,9	17,8	59,5
Korrekte Schätzung der Grünen-Wähler_innen in %	36,6	46,3	67,4
n=	1795	1795	1795

1 Eine Parallelanalyse mit den ursprünglich erhobenen Variablen führt in dem Beispiel zu einer singulären Hessematrix, also zu dem genannten Multikolinearitätsproblem.

bietet IBM-SPSS nur bei wenigen Verfahren statistische Signifikanztests an. Daher wurden die von Van Ginkel (2010, 2016) entwickelten Makros genutzt.

4.5 Validitätsprüfung bzw. Evaluation der Datenfusion

Für die formale und inhaltliche Validitätsprüfungen wurden sieben Hypothesen aufgestellt (siehe Tabelle 4). Mit den Hypothesen 1 bis 5 wird eine kriterienbezogene Validitätsprüfung vorgenommen. Die Hypothesen beruhen auf den Ergebnissen von Wahlanalysen einschließlich von Wählerstromanalysen der Nationalratswahl 2013 und der Bundespräsidentschaftswahl 2016 (Sora 2013, 2016). Diese weisen z.B. aus, dass die FPÖ mehrheitlich von Männern und Personen mit formal niedrigeren Bildungsabschlüssen gewählt wurde (Sora 2013, 5ff), für die die Zuwanderungsfrage ein zentrales Thema war (Sora 2013, 10). Hypothesen mit Bezug zu Tradition und politisch rechter Verortung sind durch aktuelle wissenschaftliche Literatur gestützt (z.B. rezent Heinisch 2017, 449).

Mit den Hypothesen 6 bis 7 wird die formale Gültigkeitsprüfung vorgenommen. Hypothese 6 prüft die Forderung, dass Zusammenhänge, die in der Spenderdatei bestehen, in der Empfängerdatei erhalten bleiben, Hypothese 7 die lokale Unabhängigkeit.

Für die Hypothesen 1 bis 5, die der inhaltlichen Gültigkeitsprüfung dienen, können beim allgemeinen Imputationsmodell drei der fünf Hypothesen angenommen werden (siehe Tabelle 5). Die Wähler_innen der freiheitlichen Partei zeigen signifikant niedrigeres Bildungsniveau, sind öfter traditionsbewusst und verorten sich auch eher rechts auf der politischen Einstellungsskala. Die Hypothese zum Geschlecht, nämlich dass FPÖ-Wähler_innen signifikant öfter männlich sind als Grüne-Wähler_innen, lässt sich nicht bestätigen. Dies entspricht aber auch der Feststellung von Sora (2013), dass Grün-Wähler_innen nicht mehr eindeutig mehrheitlich weiblich sind und somit das Geschlecht nicht mehr als signifikante Trennlinie zwischen FPÖ- und Grün-Wähler_innen gesehen werden kann. Widersprüchlich ist aber das Ergebnis, dass die FPÖ-Wähler_innen keine signifikant negativere Einstellung zur Immigration berichten als die Grünen-Wählerinnen. Das allgemeine Imputationsmodell ist offensichtlich zu ungenau, sodass der in der Spenderdatei vorhandene Zusammenhang zwischen FPÖ-Wähler_innen und Einstellung zur Immigration (r = 0,510 in der Spenderdatei, aber nur r = 0,144 beim allgemeinen Imputationsmodell) insignifikant wird.

Durch die beiden spezifischen Imputationsmodelle werden die in der Spenderdatei enthaltenden Zusammenhänge gut reproduziert. Es bestehen keine signifikanten Differenzen zwischen den auf der Basis der imputierten Variablen berechneten Korrelationen und den ursprünglichen Korrelationen. Die Annahme der Erhaltung der Zusammenhänge (H6) ist in beiden Modellen erfüllt, wobei hier auch das spezifische Imputationsmodell 1 fast ebenso gut abschneidet wie das spezifische Imputationsmodell 2, das mitunter sogar numerisch höhere Korrelationen aufweist.

Tabelle 4: Hypothesen für die formale und kriterienbezogene Gültigkeitsprüfung

	Hypothese	*Kriterium*
	Kriterienbezogene Gültigkeitsprüfung bzw. Evaluierung	
H1	Die FPÖ hatte signifikant mehr männliche Wähler als die Grünen	Kriteriumsvalidität
H2	Die FPÖ-Wähler_innen hatten ein signifikant niedrigeres Bildungsniveau als die Grünen-Wähler_innen.	Kriteriumsvalidität
H3	Die FPÖ-Wähler_innen sehen Tradition in einem signifikanten höheren Ausmaß als wichtig an als die Grünen-Wähler_innen	Kriteriumsvalidität
H4	Die FPÖ-Wähler_innen haben in einem signifikanten Ausmaß eine negativere Einstellung zur Immigration als die Grünen-Wähler_innen	Kriteriumsvalidität
H5	FPÖ-Wähler_innen verorten sich signifikant öfter rechts auf der Links-Rechts-Skala als die Grünen-Wähler_innen	Kriteriumsvalidität
	Formale Gültigkeitsprüfung bzw. Evaluierung	
H6	Statistische Tests ergeben ähnliche Ergebnisse für H1 bis H5, wenn sie mit den Daten des ESS ohne imputierte Werte getestet werden und wenn sie mit Daten aus dem SSÖ mit imputierten Werten getestet werden.	Beibehaltung von Zusammenhängen
H7	Alle signifikanten Ergebnisse der Hypothesentests H1 bis H5 werden insignifikant, wenn die gemeinsamen Variablen in die Analyse aufgenommen werden.	Lokale Unabhängigkeit

Tabelle 5: Inhaltliche Gültigkeitsprüfung und Beibehaltung der Zusammenhänge

		Allgemeines Modell	*Spezifisches Modell I*[a)]	*Spezifisches Modell II*[b)]	*Spenderdatei*
H1	Die FPÖ hatte signifikant mehr männliche Wähler als die Grünen	0,041 -1,186[c)]	0,034 -1,282(*)	0,167 0,050	0,161**
H2	Die FPÖ-Wähler_innen hatten ein signifikant niedrigeres Bildungsniveau als die Grünen-Wähler_innen.	-0,337** 1,507 (*)	-0,364** 1,233	-0,491** -0,265	-0,466***
H3	Die FPÖ-Wähler_innen sehen Tradition in einem signifikanten höheren Ausmaß als wichtig an als die Grünen-Wähler_innen	0,256** 0,291	0,281** 0,566	0,296* 0,597	0,228***
H4	Die FPÖ-Wähler_innen haben in einem signifikanten Ausmaß eine negativere Einstellung zur Immigration als die Grünen-Wähler_innen	0,144(*) -4,082***	0,561*** 0,715	0,613*** 1,233	0,510***
H5	FPÖ-Wählerinnen verorten sich signifikant öfter rechts auf der Links-Rechts-Skala als Grün-Wähler_innen	0,187* -5,560***	0,574*** -1,046	0,715*** 1,137	0,640***
	n (für Differenzentest)	141	150	88	313

*(*) p < 10%, * p < 5%, ** p < 1%, ***p <0,1%.*
a) Spezifisches Modell I = Allgemeines Modell (siehe Text) + Links-Rechts-Skala + Einstellung zur Migration
b) Spezifisches Modell II = Spezifisches Modell I + Parteipräferenz
c) Erste Zeile = bivariate Korrelation, signifikante Abweichungen von 0 notiert, z.B. r=0,41 ist nicht signifikant von 0 verschieden, r=-0,337 zu 1%. Für die Tests in Zeile 1 wurden einseitige Fehlerniveaus verwendet, da gerichtete Hypothesen vorliegen.
Zweite Zeile = z-Wert nach Fisher und Signifikanz für Abweichung vom Zusammenhang in der Spenderdatei, z.B. r=0,041 weicht nicht signifikant von r=0,161 ab. Der z-Wert nach Fisher beträgt -1,186 und ist nicht signifikant. Der Wert von r=0,144 weicht dagegen signifikant (p<0,1%) von r=0,510 ab. Der z-Wert beträgt -4,082. Für die Tests in Zeile 2 wurden zweiseitige Fehlerniveaus verwendet, da keine gerichteten Hypothesen vorliegen.

Tabelle 6: Prüfung der lokalen Unabhängigkeit

		Allgemeines Modell	*Spezifisches Modell I*[a)]	*Spezifisches Modell II*[b)]
H1	Die FPÖ hatte signifikant mehr männliche Wähler als die Grünen	-0,098	-0,037	-0,049
H2	Die FPÖ-Wähler_innen hatten ein signifikant niedrigeres Bildungsniveau als die Grünen-Wähler_innen.	0,018	0,051	-0,007
H3	Die FPÖ-Wähler_innen sehen Tradition in einem signifikanten höheren Ausmaß als wichtig an als die Grünen-Wähler_innen	0,006	0,125	0,020
H4	Die FPÖ-Wähler_innen haben in einem signifikanten Ausmaß eine negativere Einstellung zur Immigration als die Grünen-Wähler_innen	-0,105	-0,150	0,098
H5	FPÖ-Wählerinnen verorten sich signifikant öfter rechts auf der Links-Rechts-Skala als Grün-Wähler_innen	-0,084	0,047	0,021

*(*) p(einseitig) < 10%, * p(einseitig) < 5%, ** p(einseitig) < 1%, ***p(einseitig) <0,1%.*
a) Spezifisches Modell I = Allgemeines Modell (siehe Text) + Links-Rechts-Skala + Einstellung zur Migration
b) Spezifisches Modell II = Spezifisches Modell I + Parteipräferenz, wegen der geringen Fallzahl konnten nur die ersten 20 gemeinsamen abgeleiteten Variablen in die Testung aufgenommen werden

Die Annahme der lokalen Unabhängigkeit (H7) ist in allen drei Imputationsmodellen erfüllt. Die ermittelten Zusammenhänge werden insignifikant, wenn der Einfluss der gemeinsamen Variablen statistisch kontrolliert wird.

4.6 Datenanalysen zur Beantwortung der Forschungsfrage

Die getroffene Annahme, dass Grün-Wähler_innen signifikant öfter Van der Bellen wählten und FPÖ-Wähler_innen signifikant öfter Hofer, kann für das allgemeine Imputationsmodell im fusionierten Datensatz nicht bestätigt werden kann. Wie Tabelle 7 zu entnehmen

Tabelle 7: Ergebnisse für Forschungsfrage

Modell	*X1: Nationalratswahl 2013 (imputiert*	*n*	*Hofer*	*Van der Bellen*	*Andere Parteien*	*Antwort-verweigerung*	*t-Wert*
				Y1: Erste Runde der Bundespräsidentschaftswahl			
Allgem.	FPÖ	76	30,0	15,3	37,0	17,7	1,502(*)
	Grüne	65	18,6	30,5	35,4	15,5	-1,057
Spez. I[a)]	FPÖ	88	42,6	9,6	30,3	17,5	3,997**
	Grüne	62	9,9	42,1	34,2	13,8	-3,034**
Spez. II[b)]	FPÖ	54	67,1	4,1	17,3	11,5	5,929***
	Grüne	33	6,6	58,8	27,8	6,8	-2,840*
				Y2: Zweite Runde der Bundespräsidentschaftswahl			
Allgem.	FPÖ	75	43,1	35,0	-	21,9	0,598
	Grüne	65	28,0	51,0	-	21,0	-1,635(*)
Spez. I[a)]	FPÖ	86	56,3	23,9		19,8	2,314*
	Grüne	62	18,0	62,9		19,1	-2,979**
Spez. II[b)]	FPÖ	55	76,8	10,8		12,4	3,834**
	Grüne	33	12,1	79,1		8,8	-3,642**

*(*) p(einseitig) < 10%, * p(einseitig) < 5%, ** p(einseitig) < 1%, ***p(einseitig) <0,1%.*
a) Spezifisches Modell I = Allgemeines Modell (siehe Text) + Links-Rechts-Skala + Einstellung zur Migration
b) Spezifisches Modell II = Spezifisches Modell I + Parteipräferenz

ist, sind nur tendenzielle Unterschiede feststellbar. Bei den spezifischen Modellen zeigen sich dagegen die erwarteten signifikanten Zusammenhänge. Dieser Befund spricht erneut für die spezifischen Imputationsmodelle.

In allen drei Modellen ist die Annahme der lokalen Unabhängigkeit erfüllt. Der Zusammenhang zwischen berichtetem Wahlverhalten in der Nationalratswahl und der Bundespräsidentenwahl wird insignifikant, wenn der Einfluss der gemeinsamen Variablen statistisch kontrolliert wird.

Betrachtet man die Ergebnisse im Detail, zeigt sich beispielsweise für das spezifische Imputationsmodell 2 folgendes Bild: Auf Grundlage der imputierten Daten ist davon auszugehen, dass 67% (1. Wahldurchgang) bzw. 77% (2. Wahldurchgang) der FPÖ-Wähler_innen für Hofer stimmten bzw. 4% und 11% der FPÖ-Wähler_innen für Van der Bellen votierten. Die Werte liegen immer noch über den in den verfügbaren Wählerstromanalysen ermittelten Übergängen (Sora 2016). Sie kommen aber diesen Werten bereits sehr nahe, wobei zu beachten ist, dass auch Wählerstromanalysen auf Aggregatdaten und Schätzungen beruhen und somit auch hier Schätzfehler vorliegen.

5. Zusammenfassung und Conclusio

Bei der Datenfusion werden fehlende Variablen in einem Datensatz auf der Grundlage gemeinsamer Variablen und eines impliziten oder expliziten statistischen Verfahrens aus einem anderen Datensatz eingefügt. Datenfusion lässt sich statistisch als spezifisches Imputationsproblem definieren. Aktuell kommt Datenfusion in den Sozialwissenschaften mit Ausnahme der Markt- und Medienforschung sowie der amtlichen Statistik nur selten zur Anwendung, obwohl im zunehmenden Ausmaß Daten aus unterschiedlichen Quellen in Datenarchiven vorhanden wären. Datenfusion böte grundsätzlich die Chance, dass sich Umfragen auf bestimmte Themen konzentrieren und die Gefahr von zufälligen und systematischen Antwortfehlern könnte verkleinert werden. Die umfangreichen Voraussetzungen für das Verfahren und die potentielle Gefahr, dass inkorrekte Schlüsse gezogen werden, sind potentielle Erklärungen für den zurückhaltenden Einsatz.

Theoretisch begründbar ist diese Abstinenz nicht. Die Datenfusion kann formal sowohl zu einem Genauigkeitsgewinn als auch zu einem Genauigkeitsverlust führen. Ein Genauigkeitsgewinn könnte dann erreicht werden, wenn ein Datenfusionsmodell spezifiziert

werden kann, dessen Prognosekraft die Zuverlässigkeit der Messung der zu imputierenden Variablen übertrifft. Die Zuverlässigkeit von Messmodellen schwankt in den Sozialwissenschaften und hängt von vielen Faktoren ab, als grober Richtwert kann aber ein Wert von 0,5 bzw. 0,6 für die erklärte Varianz betrachtet werden. Das hier durchgeführte Anwendungsbeispiel spricht dafür, dass der Wert von 0,5 auf jeden Fall überschritten werden muss. Bei einem schlechten Imputationsmodell tritt eine deutliche Reduktion der Zusammenhänge und der Signifikanzen auf.

Für die Datenfusion steht eine Vielzahl an Verfahren zur Verfügung, für die – für den Anwender/die Anwenderin – irritierend unterschiedliche Bezeichnungen verwendet werden. Aus statistischer und Anwendersicht wird von den Autoren dieses Beitrags der Einsatz variablenorientierter (modellbasierter) multipler Imputationsverfahren empfohlen. In der Literatur sind aber auch differierende Vorschläge auffindbar.

Die Anwendung der Datenfusion erfordert ein sorgfältiges, inhaltlich und methodisch begründetes Vorgehen, das sich aus mehreren Schritten zusammensetzt und bei dem zahlreiche Entscheidungen zu treffen sind. Sie erfordert methodische und inhaltliche Expertise.

Durch ein Anwendungsbeispiel wurde geprüft, ob eine Datenfusion mit Standardstatistikpaketen durchführbar ist und ob mit einem allgemeinen Imputationsmodell, das sich für unterschiedliche Fragestellungen eignet, befriedigende Ergebnisse erzielt werden können. Zusammenfassend lässt sich festhalten, dass eine Datenfusion mit Standardstatistikpaketen, konkret verwendet wurde IBM-SPSS, möglich und durchführbar ist. Der Versuch, ein allgemeines Datenfusionsmodell zu entwickeln, das auf sozio-demographische Variablen und allgemeine Wertorientierungen basiert, scheiterte.

Die Beantwortung der Frage, ob die Datenfusion eine zu Unrecht vernachlässigte Methode in den Sozialwissenschaften ist, bleibt weiteren Anwendungen vorbehalten. Das durchgeführte Beispiel, die zunehmende Verfügbarkeit von Daten und die theoretische Möglichkeit eines Genauigkeitsgewinns sollten aber Anlass sein, sich weiterhin wissenschaftlich mit der Datenfusion zu beschäftigen. Dafür wurde in dem Beitrag ein Überblick über Techniken gegeben. Für eine konkrete Umsetzung kann die über AUSSDA verfügbare Syntax adaptiert werden. Aufgrund der gewonnenen Erkenntnisse erscheint es zielführend, in zukünftigen Anwendungsversuchen bereits bei der Planung einer Studie eine beabsichtigte Datenfusion zu berücksichtigen und Spenderdaten auszuwählen bzw. zu erheben, die auch Variablen als gemeinsame Variablen enthalten, die im Fokus der Untersuchung stehen, wie hier z.B. die Links-Rechts-Einstufung, die Einstellung zur Immigration oder die Parteipräferenz. Die Verwendung von allgemeinen soziodemografischen Daten und Wertorientierungen erscheint nicht mehr ausreichend, da ihnen insgesamt eine geringere Erklärungskraft zukommt.

Danksagung

Die diesem Beitrag zugrundeliegende Forschung wurde vom damaligen Bundesministerium für Wissenschaft, Forschung und Wirtschaft (BMWFW) finanziell gefördert (Projekt Nr.: GZ BMWFW-3.020 / 0015-WF / V / 4c / 2015). Wir danken den beiden anonymen Reviewer_innen für ihre wertvollen Anregungen, durch die der Beitrag entscheidend verbessert werden konnte.

Literatur

Aluja-Banet, Tomàs/*Daunis-I-Estadella*, Josep et al. (2015), Improving prevalence estimation through data fusion. Methods and validation, in: *BMC medical informatics and decision making*, Vol. 15, 49. DOI: 10.1186/s12911-015-0169-z.

Babic, Damir/*Hagenah*, Jörg/*Meulemann*, Heiner (2011), Über die Fusionskonzepte zur Bildung der MA-Intermedia, in: *MLFZ-Reihe: Dokumentationen zur Aufbereitung der Media-Analyse* (1), 1–26, Internet: http://www.mlfz.uni-koeln.de/assets/files/Dokumentation/MA-Intermedia.pdf (Zugriff: 18.03.2018).

Bacher, Johann/Andreas *Pöge*/Knut *Wenzig* (2010), Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren,3., erg., vollst. überarb. und neu gestaltete Aufl., München: Oldenbourg, Internet: http://www.oldenbourg-link.com/isbn/9783486710236 (Zugriff: 15.09.2018).

Bacher, Johann/Dimitri *Prandner* (2017), Abschlussbericht zum Sozialen Survey Österreich 2016, Linz: Eigenverlag.

Bacher, Johann/Dimitri *Prandner* (2018a), Anwendungsschritte bei der Datenfusion. Supplement 2 zu Bacher/Prandner (2018) Datenfusion in der sozialwissenschaftlichen Wahlforschung - Begründeter Verzicht oder ungenutzte Chance?, Linz: Abteilung für empirische Sozialforschung - JKU Linz.

Bacher, Johann/Dimitri *Prandner* (2018b), Syntax zur Datenfusion. Supplement 3 zu Bacher/Prandner (2018) Datenfusion in der sozialwissenschaftlichen Wahlforschung - Begründeter Verzicht oder ungenutzte Chance?, Linz: Abteilung für empirische Sozialforschung - JKU Linz.

Baker, Ken (2007), Data integration methodologies in market research: an overview, in: *International Journal of Market Research*, Vol. 49 (4), 435–447.

Baker, Ken/*Harris*, Paul/*O'Brien*, John (1989), Data fusion: An appraisal and experimental evaluation, in: *Journal of the market research society*, Vol. 31 (2), 153–212.

Biemer, P. P. (2011), Total Survey Error. Design, Implementation, and Evaluation, in: *Public Opinion Quarterly*, Vol. 74 (5), 817–848, DOI: 10.1093/poq/nfq058.

Breur, Tom (2011), Data analysis across various media: Data fusion, direct marketing, clickstream data and social media, in: *Journal of Direct, Data and Digital Marketing Practice*, Vol. 13, 95–105.

Bryman, Alan (2012), Social research methods,4. ed., Oxford u.a: Oxford Univ. Press.

Cielebak, Julia/*Rässler*, Susanne (2014), Data Fusion, Record Linkage und Data Mining, in: Baur, Nina; Jörg Blasius (Hg.), Handbuch Methoden der empirischen Sozialforschung, Wiesbaden: Springer VS, 367–382.

Conti, Pier Luigi/*Marella*, Daniela/*Scanu*, Mauro (2017), Statistical Matching Analysis for Complex Survey Data With Applications, in: *Journal of the American Statistical Association*, Vol. 111 (516), 1715–1725. DOI: 10.1080/01621459.2015.1112803.

Czaia, Uwe (2005), Media-Analysen & Fusionen, in: König, Christian (Hg.), Datenfusion und Datenintegration. 6. Wissenschaftliche Tagung, Bonn: Informationszentrum Sozialwissenschaften (Tagungsberichte, 10), 45–52.

Denk, Micahela/*Hackl*, Peter (2003), Data integration and record matching: an Austrian contribution to research in official statistics, in: *Austrian Journal of Statistics*, Vol. 23 (4), 305–321.

D'Orazio, Marcello/*Di Zo*, Marco/*Scanu*, Mauro (2001), Statistical Matching: a tool for integrating data in National Statistical Institutes., in: Proceedings of the Joint ETK and NTTS Conference for Official Statistics.

D'Orazio, Marcello/*Di Zo*, Marco/*Scanu*, Mauro (2006), Statistical matching. Theory and practice, Chichester: John Wiley (Wiley series in survey methodology), Internet: http://dx.doi.org/10.1002/0470023554.

Enders, Craig K. (2010), Applied missing data analysis, New York: Guilford Press (Methodology in the social sciences), Internet: http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10389908 (Zugriff: 16.03.2018).

ESS (2015), ESS Round 7 - European Social Survey Round 7 Documentation Report. Edition 3.1., Bergen: European Social Survey Data Archive, NSD – Norwegian Centre for Research Data for ESS ERIC, Internet: http://www.europeansocialsurvey.org/ (Zugriff: 23.03.2018).

George, Darren/Paul *Mallery* (2010), SPSS for Windows Step by Step: A Simple Guide and Reference 18.0 Update, New Jersey: Prentice Hall Press.

Graham, John W. (2012), Missing data. Analysis and design, New York: Springer (Statistics for social and behavioral sciences). Internet: http://dx.doi.org/10.1007/978-1-4614-4018-5.

Graham, John W./*Olchowski*, Allison E./*Gilreath*, Tamika D. (2007), How many imputations are really needed? Some practical clarifications of multiple imputation theory, in: *Prevention science : the official journal of the Society for Prevention Research*, Vol. 8 (3), 206–213, DOI: 10.1007/s11121-007-0070-9.

Groves, Robert M./*Singer*, Elanor et al. (1999), A Laboratory Approach to Measuring the Effects on Survey Participation of Interview Length, Incentives, Differential Incentives, and Refusal Conversion, in: *Journal of Offcial Statistics*, Vol. 15 (2), 251–268.

Hansen, Kasper M. (2006), The Effects of Incentives, Interview Length, and Interviewer Characteristics on Response Rates in a CATI-Study, in: *International Journal of Public Opinion Research*, Vol. 19 (1), 112–121, DOI: 10.1093/ijpor/edl022.

Heinisch, Reinhard C. (2017), Party politics and the European Union since 1989, in: *Livezenau*, Irina; Árpád von Klimó (Hg.), The Routledge History of East Central Europe Since 1700, London: Routledge, 415–464.

Kaplan, David/*McCarty*, Alyn Turner (2013), Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys, in: *Large-scale Assessments in Education*, Vol. 1 (1), 1–26.

Kiesl, Hans/Susanne *Rässler* (2006), How valid can data fusion be, Nürnberg: IAB, Internet: http://doku.iab.de/discussionpapers/2006/dp1506.pdf (Zugriff: 16.03.2018).

Kim, Jonathan S./*Baek*, Seung/*Chi*, Sungbin (2004), A Preliminary Study on Common Variable Selection Strategy in Data Fusion, in: *NA - Advances in Consumer Research*, Vol. 31, 716–720, Internet: http://www.acr-website.org/volumes/9002/volumes/v31/NA-31 (Zugriff: 16.03.2018).

Liu, Tzen-Ping/*Kovacevic*, Milorad S. (1997), An empirical study on categorically constrained matching, in: *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 167–187.

Loosveldt, Geert/*Beullens*, Koen (2013), The impact of respondents and interviewers on interview speed in face-to-face interviews, in: *Social science research*, Vol. 42 (6), 1422–1430. DOI: 10.1016/j.ssresearch.2013.06.005.

Nisbett, Richard E./*Wilson*, Timothy D. (1977), The halo effect: Evidence for unconscious alteration of judgments, in: *Journal of personality and social psychology*, Vol. 35 (4), 250–256.

OECD (2005), PISA 2003. Technical Report, Paris: OECD.

Piscirelli, Alfonso/*D'Ambrosio*, Antonio (2012), Assession Assumptions for Data Fusion Procedures, in: *46th scientific meeting of the italian statistical society*, Internet: http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/view/2385 (Zugriff: 15.09.2018).

Prandner, Dimitri/Johann *Bacher* (2018), Mögliche positive und negative Effekte der Datenfusion auf die Datenqualität. Supplement 1 zu Bacher/Prandner

(2018) Datenfusion in der sozialwissenschaftlichen Wahlforschung - Begründeter Verzicht oder ungenutzte Chance?, Wien: AUSSDA.

Rässler, Susanne (2002), Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches, New York, NY: Springer (Lecture Notes in Statistics, 168), Internet: http://dx.doi.org/10.1007/978-1-4613-0053-3 (Zugriff: 16.03.2018).

Rässler, Susanne (2004), The impact of multiple imputation for DACSEIS, in: *DACSEIS Research Paper Series*, Vol. 5, 1–23.

Rius, Xavier/*Riu*, Jordi et al. (1999), Estimating uncertainties of analytical results using information from the validation process, in: *Analytica Chimica Acta*, Vol. 391 (2), 173–185.

Saporta, Gilbert (2000), Data Fusion and Data Pruning, Internet: https://pdfs.semanticscholar.org/d544/33346d45c10385262c48da6f06aa16c6d448.pdf (Zugriff: 15.09.2018).

Schwartz, Shalom H./*Melech*, Gila et al. (2001), Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement, in: *Journal of Cross-Cultural Psychology*, Vol. 32 (5), 519–542, DOI: 10.1177/0022022101032005001.

Smith, Thomas J./*McKenna*, Corneluis M. (2013), A Comparison of Logistic Regression Pseudo R2 Indices, in: *Multiple Linear Regression Viewpoints*, Vol. 39 (2), 17–26, Internet: http://www.glmj.org/archives/articles/Smith_v39n2.pdf (Zugriff: 18.03.2018).

Soong, Roland/Montigny, Michelle de (2004), No free lunch in data fusion / integration, in: *ARF/ESOMAR Week of Audience Measurement, 2004, 33-54.*

Sora (2013), Wahlanalyse Nationalratswahl 2013, Internet: http://www.sora.at/fileadmin/downloads/wahlen/2013_NRW_Wahlanalyse.pdf (Zugriff: 24.03.2018).

Sora (2016), Wahlanalyse Stichwahl Bundespräsidentschaft 2016, Internet: http://www.sora.at/fileadmin/downloads/wahlen/2016_BP-Stichwahl_Wahlanalyse.pdf (Zugriff: 18.03.2018).

Spearman, Charles (1904), The Proof and Measurement of Association between Two Things, in: *The American Journal of Psychology*, Vol. 15 (1), 72–101, Internet: http://www.jstor.org/stable/1412159 (Zugriff: 15.09.2018).

Sudman, Seymour/Norman M. *Bradburn* (1982), Asking questions. [a practical guide to questionnaire design],1. ed., San Francisco: Jossey-Bass.

van der Putten, Peter/*Kok*, Joost N./*Gupta*, Amar (2002), Data Fusion through Statistical Matching., in: Center for eBusiness@MIT, Paper 185, Internet: http://fileserver.itb.cnr.it/acalabria/PhD/Materiale/DataQuality/Statistical%20Matching/Gupta%202002%20-%20Data%20Fusion%20Through%20Statistical%20Matching.pdf (Zugriff: 18.03.2018).

van Ginkel, Jost (2010), SPSS Syntax for Applying Roles for Combining Univariate Estimates in Multiple Imputation, Internet: https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel/publications#tab-1 (Zugriff: 18.03.2018).

van Ginkel, Jost (2016), SPSS Syntax for Applying Roles for Combining Multivariate Estimates in Multiple Imputation, Internet: https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel/publications#tab-1 (Zugriff: 18.03.2018).

Wegscheider-Pichler, Alexandra/*Haslinger*, Alois (2015), Statistical Matching of EU-SILC and MC Environment for Analysing Environmental Conditions and Behaviour in Dependence of Household Income, in: *Journal of Statistical Science and Application*, Vol. 3 (7-8), 111–121.

Yucel, Recai M. (2011), State of the Multiple Imputation Software, in: *Journal of Statistical Software*, Vol. 45 (1), 1–7.

Zerr, Konrad/*Linxweiler*, Richard/*Forster*, Anja (2011), Kontextsensitives digitales Marketing zur Steigerung des „Value in Context" und Herausforderungen für die digitale Markenführung, in: *Theobald*, Elke (Hg.), Brand Evolution. Moderne Markenführung im digitalen Zeitalter; mit Praxisbeiträgen von Google, Ferrero, Jägermeister, Mercedes-Benz, EnBW, Otto, Edelight und Anne Korn. 1. Aufl., Wiesbaden: Gabler Verlag / Springer Fachmedien Wiesbaden GmbH Wiesbaden, 167–195.

Autoren

Univ.-Prof. Dr. Bacher ist Professor für empirische Sozialforschung am Institut für Soziologie an der Johannes Kepler Universität Linz. Seine Forschungsschwerpunkte sind die Methoden der empirischen Sozialforschung, Bildungsungleichheitsforschung sowie Soziologie der Kindheit und Jugend und des Abweichenden Verhaltens. Er ist unter Johann.Bacher@jku.at erreichbar.

Dimitri Prandner ist wissenschaftlicher Mitarbeiter am Institut für Soziologie an der Johannes Kepler Universität Linz und Begleitforscher für AUSSDA – the Austrian Social Science Data Archive. Zusätzlich lehrt er als Senior Lecturer Methoden der empirischen Sozialforschung am Fachbereich Kommunikationswissenschaft an der Paris Lodron Universität Salzburg. Seine Forschungsschwerpunkte sind im Bereich der Kommunikationssoziologie und quantitativen Datenerhebung.

Er ist unter Dimitri.Prandner@jku.at erreichbar.

Anhang

Tabelle A1: Gemeinsame Variablen Z aus dem SSÖ und ESS

Gemeinsame Variablen für allgemeines Imputationsmodell	Ausprägungen
Z1 Es ist wichtig neue Ideen zu entwickeln und kreativ zu sein. Z2 Reichtum ist wichtig. Z3 Sicherheit ist wichtig. Z4 Spaß und Vergnügen ist wichtig. Z5 Personen in der Umgebung helfen ist wichtig. Z6 Erfolg ist wichtig. Z7 Auf Abenteuer aus und nimmt dafür Risiken auf sich Z8 Alles vermeiden, was Leute als Fehltritt bezeichnen könnten Z9 Umweltschutz ist wichtig Z10 Traditionen sind wichtig	1=„gleich mir sehr“ bis 6=„gleicht mir gar nicht“
Z11 Geschlecht	1=„weiblich“, 0=„männlich“
Z12 Alter in Jahren	Jahre
Z13 Geburtsland des Vaters Z14 Geburtsland der Mutter Z16 Staatsbürgerschaft Z15 Geburtsland des Befragten	1=„Österreich“; 0=„Anderes Land“
Z17 Aktuelle Tätigkeit	1=„Erwerbstätig“ 2=„Schule oder Studium“ 3=„Arbeitslos“ 4=„Arbeitsunfähig oder in Invaliditätspension“ 5=„Im Ruhestand“ 6=„Hausarbeit / Reproduktionstätigkeit“
Z18 Abgeschlossene Schulbildung	1=„Pflichtschule“ 2=„Lehre“ 3=„BMS“ 4=„BHS“ 5=„AHS“ 6=„Postgrad. Ausbildung“ 7=„Bachelor FH“ 8=„Bachelor Universität“ 9=„Master/Diplom FH“ 10=„Master/Diplom Universität“ 11=„Doktorat“ 12=„andere“

Weitere gemeinsame Variablen für spezifische Imputationsmodelle	Ausprägungen
Z19 Links-Rechts-Skala	1=„links“ bis 11=„rechts“
Z20 Einstellung zur Immigration	1=„positiv“ bis 5=„negativ“
Z21 Parteipräferenz	1=„SPÖ“ 2=„ÖVP“ 3=„FPÖ“ 4=„GRÜNE“ 5=„andere“ 6=„keine Angabe, verweigert“