Ilmenauer Beiträge zur Wirtschaftsinformatik

Herausgegeben von U. Bankhofer, V. Nissen
D. Stelzer und S. Straßburger

Dieter W. Joenssen

Hot-Deck Imputation unter Verwendung eines Donor-Limit: Ein ganzzahliges Optimierungsproblem

Arbeitsbericht Nr. 06, Dezember 2013

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Dieter W. Joenssen

Titel: Hot-Deck Imputation unter Verwendung eines Donor-Limit: Ein ganzzahliges Optimierungsproblem

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2013-06, Technische Universität Ilmenau, Dezember 2013

ISSN 1861-9223

ISBN 978-3-938940-51-8

URN urn:nbn:de:gbv:ilm1-2013200251

Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.
http://www.db-thueringen.de/servlets/DocumentServlet?id=5507

Gliederung

Tabellenverzeichnis

Zusammenfassung: Hot-Deck Methoden imputierten fehlende Daten durch eine Zuordnung von vollständigen Objekten zu den Objekten, bei denen Werte fehlen. Fehlende Beobachtungen innerhalb des Rezipienten werden dann durch Verdopplung der Werte des zugeordneten Donor behoben. Einige Hot-Deck Verfahren begrenzen die Häufigkeit, mit der ein vollständiges Objekt seine Werte spenden kann, um die Präzision von der postimputationalen Parameterschätzung zu erhöhen. Diese Beschränkung, auch Donor-Limit gennant, beschränkt das Risiko, dass ein Spender exklusiv oder „zu häufig“ zur Imputation herangezogen wird. Trotz dieser erstrebenswerten Eigenschaften sind in Konsequenz die Ergebnisse eines spenderbegrenzten Hot-Deck abhängig von der Reihenfolge, in der die Objekte mit fehlenden Werten imputiert werden. Dies ist eine unerwünschte Eigenschaft, da nun nichtmehr die Gesamtähnlichkeit zwischen allen Donor und Recipients maximiert wird durch die schrittweise Wahl des ähnlichsten Donor für einen Recipient. Daher kann die Impuataionsqualität durch eine global optimale Zuordnung von Spendern zu Empfängern verbessert werden. In dieser Arbeit wird das ganzzahlige Optimierungsproblem formuliert sowie eine Simulation präsentiert die zeigt, dass eine bessere Lösung dieses Optimierungsproblems nie zu schlechteren Ergebnissen führt.

Abstract: Hot deck methods impute missing data by matching records that are complete to those that are missing values. Observations absent within the recipient are then replaced by replicating the values from the matched donor. Some hot deck procedures constrain the frequency with which any donor may be matched to increase the precision of post-imputation parameter-estimates. This constraint, called a donor limit, also mitigates risks of exclusively using one donor for all imputations or using one donor with an extreme value or values ``too often''. Despite these desirable properties, imputation results of a donor limited hot deck are dependent on the recipients' order of imputation, an undesirable property. For nearest neighbor type hot deck procedures, the implementation of a constraint on donor usage causes the stepwise matching between each recipient and its closest donor to no longer minimize the sum of all donor-recipient distances. Thus, imputation results may further be improved by procedures that minimize the total donor-recipient distance-sum. The discrete optimization problem is formulated and a simulation detailing possible improvements when solving this integer program is presented.

Schlüsselworte: Hot-Deck Verfahren, Imputation, Donor-Limit, ganzahlige Optimierung

1 Einführung

Fehlende Daten sind ein nahezu allgegenwärtiges Problem in den Sozialwissenschaften und stellen eine Herausforderung in jedem Kontext, wo Informationen gesammelt werden, dar. Der richtige Umgang mit diesen fehlenden Daten ist nicht nur für jede Analyse entscheidend, weil typische statistische Verfahren für vollständige Daten konzeptioniert sind, sondern auch weil jegliche Verzerrung, die durch den falschen Umgang mit fehlenden Werten bei der Datenvorverarbeitung entsteht, sich negativ auf die Ergebnisse einer nachgelagerten Analyse auswirkt.

Während die Abwesenheit von fehlenden Daten in jeder Erhebung von Daten wünschenswert ist, verhindern technische und ökonomische Einschränkungen, dass alle fehlenden Werte durch logische Inferenz oder eine Nacherhebung behoben werden. Deshalb müssen explizit Maßnahmen zum korrekten Umgang mit den verbleibenden fehlenden Daten ergriffen werden. Diese Maßnahmen, die auch Missing-Data Methoden genannt werden, können unter Eliminierungs-, Imputations- oder Parameterschätzverfahren subsumiert werden. Während Eliminierungsverfahren einfach jegliche Objekte oder Variablen, bei denen Merkmalsausprägungen fehlen, eliminieren und Parameterschätzverfahren unter Einbeziehung von Verteilungsannahmen Verteilungsparameter direkt schätzen, ersetzen Imputationsverfahren die fehlenden mit geschätzten Werten. Die Wahl der richtigen Missing-Data Methode hängt jedoch von dem stochastischen Prozess, der das Auftreten der fehlenden Werte bedingt, ab. Ein Verständnis dieses Prozesses ist essentiell zur Vermeidung jeglicher durch den Datenausfall entstehenden Reduktion in Präzision und Erhöhung des Bias. Grundlegend für dieses erforderliche Verständnis ist Rubins (1976) Betrachtung der möglichen stochastischen Prozesse. Die stochastischen Prozesse werden in drei Gruppen unterteilt. Welche dieser drei Ausfallmechanismen in den Rohdaten vorhanden ist, bestimmt, welche Missing-Data Methoden anwendbar sind.

Der erste Ausfallmechanismus wird Missing-Completely-At-Random (MCAR) genannt. Wenn die Daten MCAR sind, dann entsprechen die vorhandenen Daten einer Stichprobe der gewünschten Daten. Wird dieser Ausfallmechanismus festgestellt, können die Werte der fehlenden Daten mit Hilfe der vorhandenen Daten prognostiziert werden, dieses ist jedoch entbehrlich. Wird die, doch sehr stringente, Annahme des MCAR-Ausfalls getroffen, können Objekte oder Variablen, die fehlende Werte aufweisen, eliminiert werden, wobei lediglich die Konsequenzen einer kleineren Stichprobe verbleiben. Die zweite Klasse von

Ausfallmechanismus wird unter dem Begriff Missing-At-Random (MAR) subsumiert. Hier existiert eine Abhängigkeit zwischen dem Auftreten von fehlenden Werten in einer Variable und den Werten einer weiteren, vollständig vorhandenen Variable. Daher können die fehlenden Werte nicht nur vorhergesagt werden, sie müssen vorhergesagt werden. Eine Auswertung ausschließlich der vorhandenen Werte führt unweigerlich zu verzerrten Ergebnissen. Der letzte Ausfallmechanismus wurde nicht explizit von Rubin (1976) benannt, und ist daher entweder unter der Bezeichnung Not-Missing-At-Random (NMAR) (Little and Rubin (2002)) oder Missing-Not-At-Random (MNAR) (Collins et al. (2001), Enders (2010)) in der Literatur zu finden. Ist dieser letzte Ausfallmechanismus vorhanden, können die fehlenden Werte nicht aus den vorhandenen bestimmt oder geschätzt werden, da beispielsweise die Wahrscheinlichkeit, dass eine Merkmalsausprägung nicht beobachtet wird, von dem Wert dieser Ausprägung abhängig ist. Ist ein NMAR Ausfallmechanismus vorhanden, so müssen Daten und Ausfallmechanismus weitgehend und korrekt modelliert werden, damit die Auswertung der Daten korrekte Schlüsse zulässt. Alternativ kann durch Inklusion hoch korrelierter, vollständiger Kovariaten in die Analyse das Vorliegen eines MAR Mechanismus plausibler gemacht werden (vgl. Collins et al. (2001), Schafer and Graham (2002), Enders (2010)).

Eine Gruppe von Imputationsalgorithmen sind die Hot-Deck Methoden. Die Definition von dieser Klasse an Verfahren geht auf Ford (1983) und Sande (1983) zurück. Beide definieren Hot-Deck Verfahren als Imputationsmethoden, bei denen verfügbare Werte von einem oder mehreren Objekten (Donors) verwendet werden, um ein anderes, unvollständiges Objekt (Recipient) des gleichen Datensatzes zu vervollständigen. Die Auswahl eines willkürlich gewählten Donor kann zwar zu validen Ergebnissen führen, wenn die Daten MCAR sind, jedoch ist dies wenig zweckmäßig, da die MCAR Annahme wohl in den meisten Fällen nicht gegeben ist. Nicht nur können die Ergebnisse für MCAR Daten verbessert, sondern auch Hot-Deck Methoden bei MAR Daten angewandt werden, sofern dem Recipient ein ähnlicher Donor zugewiesen wird. Ähnlichkeit wird in diesem Kontext entweder mittels einer Distanzfunktion oder Mitgliedschaft in einer Imputationsklasse definiert, wobei Andridge und Little (2010) zeigten, dass die Mitgliedschaft in einer Impuationsklasse in eine entsprechende Distanzfunktion überführt werden kann. Daher können die bestehenden Methoden, um Ähnlichkeit zu definieren, auf eine Donor-Recipient Zuordnung mittels einer Kostenmatrix reduziert werden.

Ferner haben Hot-Deck Verfahren eine Anzahl an erstrebenswerten Eigenschaften. Hot-Deck Verfahren sind nicht nur berechnungstechnisch einfach durchzuführen, sondern sie bedürfen auch nur minimale Verteilungsannahmen, welches sie robust gegen Modelmisspezifikation macht (Andridge and Little (2010)). Des Weiteren garantiert die Verdoppelungseigenschaft dafür, dass Eigenschaften der Datenverteilung erhalten bleiben und, dass die imputierten Werte innerhalb des Wertebereichs des betroffenen Merkmals liegen. Trotzdem führt die bei den Hot-Deck Verfahren entstehende Verdoppelung von Werten zu unerwünschten Situationen, da grundsätzlich ein Donor für mehrere Recipients verwendet werden kann. In Extremfällen kann ein Donor für „zu viele“ oder gar alle Recipients verwendet werden oder ein Ausreißer wird „zu häufig“ zum Spenden seiner Werte verwendet. Die Wahrscheinlichkeit, dass diese Extremfälle auftreten, wird insbesondere durch die Art und Weise, wie Imputationsklassen meist definiert werden, erhöht. Zwar existieren anspruchsvollere Methoden zur Bestimmung von Imputationsklassen, trotzdem werden Imputationsklassen meist durch eine Kreuzklassifikation mehrerer kategorialer Kovariaten definiert (Andridge and Little (2010)). Diese Kreuzklassifikation, auch als Adjustment-Cell-Method bekannt, führt häufig zu vielen Imputationsklassen bei denen es an potenziellen Spender-Objekten mangelt.

Um die Risiken dieser Extremfälle zu beschränken, begrenzen einige Hot-Deck Methoden die Häufigkeit, mit der ein bestimmter Donor seine Werte spenden darf (Donor-Limit). Ein weiterer Vorteil des Donor-Limits ist, dass je strenger das Donor-Limit gewählt wird, desto geringer wird die Imputationsvarianz. Während dieses Donor-Limit mit den meist berechnungstechnisch simplen Lösungsansätzen, mit denen Praktiker praktische Probleme bewältigen, im Einklang steht, entsteht durch die Einführung eines Donor-Limit ein neues Problem. Grundsätzlich führt eine Beschränkung, wie häufig ein Donor seine Werte spenden kann, dazu, dass die Resultate der Hot-Deck Algorithmen abhängig von der Imputationsreihenfolge der Recipients werden (Bankhofer and Joenssen (2013), Kovar and Whitridge (1995)). Dieses Verhalten führt zu zwei Nachteilen. Zum einen, wenn die Recipient streng nach der Erscheinungsreihenfolge in den Daten abgearbeitet werden, kann eine einfache Umsortierung zu anderen Imputationen führen. Dies ist insbesondere für deterministische Hot-Decks von Nachteil, da Reproduzierbarkeit und stabile Ergebnisse Schlüsselargumente für deterministische Algorithmen sind. Zum anderen führt eine in jedem Schritt optimale Zuordnung von Spender zu Empfänger nicht mehr unbedingt zu einer global-optimalen Lösung, unabhängig von der Abarbeitungsreihenfolge der Recipients.

Die Zielstellung dieses Paper ist zweierlei. Zunächst wird das ganzzahlige Optimierungsproblem definiert, dass, wenn gelöst, nicht nur eine Reihenfolge unabhängige, sondern auch global optimale Donor-Recipient Zuordnung bei Hot-Deck Methoden mit Donor-Limit garantiert. Danach sollen mittels einer Simulationsstudie mögliche Verbesserungen abgeschätzt werden, die sich durch die Lösung des aufgestellten Optimierungsproblems mittels einer sequenzunabhängigen Heuristik ergeben. Die Aufstellung des Optimierungsproblems erfolgt in Abschnitt 2, während die Simulation und ihre Resultate in Abschnitt 3 diskutiert werden. Im letzten Abschnitt folgt eine Zusammenfassung und eine Diskussion möglicher weiterer Forschungsansätze.

2 Formulierung des Optimierungsproblems

Ohne Beschränkung der Allgemeinheit sei anzunehmen, dass Donor und Recipients zwei disjunkte Mengen an Objekten der Größe d bzw. r bilden. Ferner seien c_{ij} die Elemente der $d \times r$ Zuordnungskosten matrix der $n \times m$ Datenmatrix. Daraus folgt bei einem für alle Donor gleichen, konstanten Donor-Limit, dl, folgendes ganzzahliges Optimierungsproblem:

$$\begin{aligned} & g(x_{ij}) = \sum_{i=1}^{d} \sum_{j=1}^{r} c_{ij} x_{ij} \rightarrow \min \\ & \sum_{j=1}^{r} x_{ij} \leq dl, \quad \forall i = 1, \ldots, d \\ & \sum_{i=1}^{d} x_{ij} = 1, \quad \forall j = 1, \ldots, r \\ & x_{ij} \in \{0;1\}, \quad dl \in \left\{ \left\lceil \tfrac{r}{d} \right\rceil, \ldots, r \right\} \end{aligned} \tag{1}$$

wobei x_{ij} die Zuweisung von Donor i zu Recipient j angibt und $\lceil \bullet \rceil$ die Aufrundungsfunktion ist. Die zweite Zeile der Formeln (1) zeigt die Beschränkungen die durch die Einführung eines Donor-Limit entstehen. Es sind diese Beschränkungen, die eine sequentielle Abarbeitung der Empfänger suboptimal macht. Die dritte Zeile der Formeln (1) beinhaltet die fundamentale Forderung, dass alle fehlenden Daten zurch die Donor-Recipient Zuordnung behoben werden müssen. Durch diese Kombination von Anforderungen und Restriktionen lässt sich zudem der Wertebereich ermitteln, in dem ein Donor-Limit grundsätzlich variiert werden kann. Die obere Grenze eines möglichen Donor-Limit ist dadurch gegeben, dass ein Spender maximal jedem Empfänger zugeordnet werden kann. Wird als Donor-Limit r gewählt, so wird das Optimierungsproblem unbeschränkt. Die untere Grenze des Donor-Limit wird durch die Notwendigkeit, dass jedem Recipient mindestens ein Donor zugewiesen wird, festgelegt.

Eine Optimierung dieser Art, d.h. die Optimierung der gesamten Donor-Recipient Distanzsumme hat weitere attraktive Eigenschaften, zusätzlich zur Auflösung der obig beschriebenen Reihenfolgenproblematik. Erstens, ist es die logische Konsequenz dessen, dass das Angebot an Spendern beschränkt wird. Wäre kein Donor-Limit vorhanden, würden Hot-Deck Verfahren, die sequentiell den besten Donor jeden Recipient zuweisen, auch sequenzunabhängig und würde zudem auch die Donor-Recipient Distanzsumme minimieren. Zweitens, während weniger Recipients den bestmöglichen Donor zugewiesen kriegen, sind die Zuweisungen im Mittel besser. Dies sollte sich positiv auf die Imputationswerte kleiner Subpopulationen auswirken, da extreme Abweichungen zwischen den wahren, unbeobachteten Werten und den imputierten Werten weniger häufig werden. Drittens, wenn

Donor selten sind, oder ein stringenteres Donor-Limit gewählt wird, wird die Auswahl von Ausreißern unumgänglich. Dies ist nahezu paradox, da ein Donor-Limit eingentlich gegen die Verwendung von Ausreißern schützen sollte. In diesem Falle garantiert die Optimierung, dass die Ausreißer unter den Recipients so verteilt werden, dass die geringste Verzerrung in den Daten verursacht wird.

Die grundlegende Struktur des in (1) gegebenen ganzzahligen Optimierungsprogramms ist das des klassischen Transportproblems. Obwohl dieses Programm wahrscheinlich eine spezielle Struktur aufweist, kann es durchaus mittels der Algorithmen, die für das Standardproblem entwickelt wurden, gelöst werden. Eine erwähnenswerte Heuristik für die Lösung des klassischen Transportproblems ist die Spaltenminimummethode. Wenn die Zuordnungskostenmatrix so definiert ist wie hier, so dass die Zeilen die Donor und die Spalten die Recipients darstellen, ist die Spaltenminimummethode äquivalent zu der naiven Vorgehensweise die heute bei Hot-Deck Methoden mit einem Donor-Limit verwendet wird. Eine weitere erwähnenswerte Heuristik zur Lösung dieser Zuordnungsproblematik ist Vogels-Approximationsmethode (Reinfeld and Vogel (1958)). Vogels-Approximationsmethode ist zwar eine Heuristik, führt aber zu einer reihenfolgenunabhängigen Lösung, da das Auswahlkriterium für eine Zuweisung itterativ über die Distanzmatrix für alle Zeilen und Spalten in jedem Schritt neu berechnet wird (vgl. Domschke (1995)). Algorithmen, die garantiert optimale Lösungen liefern, sind beispielsweise die MODI-Methode oder Graphen-basierte Vorgehensweisen. Für eine exhaustive Beschreibung und Diskusion der optimalen Methoden und Heuristiken sei hier auf Domskte (1995) verwiesen.

3 Eine Simulationsstudie

Für eine erste Einschätzung möglicher Vorteile, die durch die Lösung des in Abschnitt 2 präsentierten Optimierungsproblems zu erzielen sind, wird im folgenden Abschnitt 3 eine Simulationsstudie durchgeführt. Abschnitt 3.1 präsentiert die verwendeten Designparameter der Studie sowie das verwendete Qualitätskriterium. Die Resultate der Simulation werden in Abschnitt 3.2 beschrieben.

3.1 Studiendesign

Zur Abschätzung, welche Verbesserungen ein komplexeres, reihenfolgenunabhängiges nearest-neighbor Hot-Deck Verfahren bietet, wird die Imputationsqualität für simulierte Daten verglichen. Die Simulation wurde in der Statistiksoftware R in der Version 2.15.2 (R Core Team (2013)) implementiert. Imputationsalgorithmen sind über die Funktion *impute.NN_HD*, als Teil des R-Pakets **HotDeckImputation** (Joenssen (2013)), verfügbar. Die im folgenden beschriebenen Faktoren entsprechen, mit einiger Vereinfachung, denen in Bankhofer and Joenssen (2014).

Die Daten bestehen strukturell aus zwei gleich großen bivariat-normalverteilten Clustern mit den Zentren (-1; -1) bzw. (1; 1). Zur Generierung von (pseudo-) Zufallszahlen die dieser Verteilung entsprechen, wurde die Funktion *rmvnorm* aus dem R-Paket **mvtnorm** (Genz et al. (2013)) verwendet. Die Größe der Datenmatrix wird zwischen 50 und 100 Objekten mit drei verschiedenen Innerklassenkorrelationen (0,00; 0,35; 0,75) variiert. Diese Datenmatrizen, welche je 1.000 mal generiert wurden, können entweder als eine einzelne Stichprobe verstanden werden oder als eine einzelne Imputationsklasse eines größeren Datensatzes.

Die hieraus resultierenden 6.000 vollständigen Datenmatrizen wurden dann drei verschiedenen Ausfallmechanismen ausgesetzt, jeweils mit vier unterschiedlichen Anteilen von fehlenden Werten. Dies wurde 1.000 mal wiederholt, so dass 90.000.000 Matrizen mit fehlenden Werten entstanden. Die relativen Häufigkeiten fehlender Werte, bezogen auf die einzige von Ausfall betroffene Variable, sind 10%, 20%, 30%, 40% und 50%. Der erste betrachtete Ausfallmechanismus ist MCAR. Hier wird zufällig der vorgegebene Anteil an Werten gelöscht. Der zweite Ausfallmechanismus, vom Typ MAR, löscht Daten, so dass die Menge an fehlenden Werten in einem Cluster immer doppelt so hoch ist wie in dem anderen. Der letzte Ausfallmechanismus ist dem zweiten ähnlich, jedoch ist der Ausfall in dem zweiten Cluster vier mal so groß wie in dem ersten.

Distanzen zwischen Donor und Recipients wurden mittels einer Euklidischen Distanz über die binäre Clusterindikatorvariable dem zweiten vollständigen Merkmal berechnet. Die Zuordnung der Donor zu den Recipients wurde dann für alle 90.000.000 Distanzmatrizen durchgeführt, einmal mit der standard-naiven Zuordnungsmethode und einmal mittels der Vogel‘schen Approximationsmethode.

Da Hot-Deck Verfahren am häufigsten bei Umfragen verwendet werden, und hier die Schätzung von Verteilungsparametern meist von Interesse ist (vgl. Little and Rubin, 2002), wird die Qualität der Impuation auf Basis von sieben Parametern bewertet. Diese Parameter wurden für alle 6.000 vollständigen und alle 180.000.000 imputierten Datenmatrizen, oder ca. 308,4 Gigabyte an Daten, berechnet. Die berechneten Parameter beinhalten vier univariate Parameter und drei multivariate Parameter, welche zwischen der imputierten Variable und der vollständigen Kovariate berechnet wurden. Die Parameter sind der Mittlelwert ($\bar{x}$), die Standardabweichung (sd_x), die Schiefe (s_x), die Kurtosis (k_x), die Mardia Schiefe ($b_{1,xy}$) und Kurtosis ($b_{2,xy}$) und die Pearson Korrelation (r_{xy}). Für jeden dieser Parameter wurde die Wurzel aus der mittleren Fehlerquadratsumme (RMSE) zwischen dem wahren Parameter (p_T), berechnet, auf Basis der vollständigen Datenmatrix, und dem auf Basis der imputierten Datenmatrix geschätzten Parameter (p_I), d.h.

$$RMSE = \sqrt{\frac{1}{1000000} \sum_{i=1}^{1000000} (p_T - p_I)^2} \tag{2}$$

Die hieraus entstehenden RMSE werden dann gemittelt, um Schätzungen der Haupteffekte zu berechnen.

3.2 Resultate

Der folgende Abschnitt beschreibt die Resultate der in Abschnitt 3.1 beschriebenen Monte Carlo Simulation. Die Werte in den Tabellen 1 bis 3 sind die Unterschiede in RMSE zwischen den beiden betrachteten Zuordnungsmethoden. Die Differenzenbildung wurden zwischen den RMSE für die naive und der Vogel’schen Approximationsmethode gebildet. Daher bedeuten positive Werte, dass Vogels Approximationsmethode besser ist als die naive Vorgehensweise. Werte in den Tabellen sind auf die ersten drei Nachkommastellen beschränkt. Zur Verbesserung der Lesbarkeit werden jene Werte, die auf die ersten drei Nachkommastellen Null sind, durch Bindestriche dargestellt.

3.2.2 Ergebnisse der Simulation für *MCAR* Daten

Die Ergebnisse aus Tabelle 1 zeigen, dass die Optimierung der Donor-Recipient Distanzsumme niemals unterlegen ist, sofern der Ausfallmechanismus *MCAR* ist. Darüber hinnaus ist das auffälligste Ergebnis, dass die Schätzung der univariaten Parameter nur marginal duch die Zuordnungsmethode beeinflusst wird. Weder Mittelwert, Standardabweichung, Schiefe noch Kurtosis der imputierten Variable werden merklich verbessert. Bei den multivariaten Parametern zeigt sich ein anders Bild. Schätzungen der Pearson Korrelation und der Mardia Schiefe und Kurtosis sind in jeder Situation deutlich besser, wenn die Daten *MCAR* fehlen. Erreichbare Verbesserungen steigen monoton mit der Innerklassenkorrelation und dem Anteil fehlender Werte. Des Weiteren scheinen kleine Datensätze etwas von der globalen Optimierungsmethode zu profitieren.

		Univariat				Multivariat		
Faktor	Faktorstufen	$\bar{x}$	sd_x	s_x	k_x	r_{xy}	$b_{1,xy}$	$b_{2,xy}$
Objekt-anzahl	50	—	—	—	—	0,004	0,086	0,121
	100	—	—	—	—	0,002	0,067	0,122
Innerklassen-korrelation	0,00	—	—	—	—	0,001	0,018	0,034
	0,35	—	—	—	—	0,003	0,051	0,092
	0,70	—	—	—	—	0,005	0,16	0,238
Anteil fehlender Werte	10%	—	—	—	—	—	—	0,001
	20%	—	—	—	—	—	0,001	0,003
	30%	—	—	—	—	—	0,004	0,013
	40%	—	—	—	—	0,001	0,039	0,080
	50%	—	—	—	—	0,013	0,338	0,510

Tabelle 1: RMSE Differenzen für *MCAR* Daten, Haupteffekte

3.2.3 Ergebnisse der Simulation für *MAR 1:2* Daten

Tablle 2 zeigt die Ergebnisse für den *MAR 1:2* Ausfallmechanismus, den Fall, dass ein Cluster doppelt soviele fehlende Werte hat wie der andere. Wieder zeigen die Ergebnisse, dass die Optimierung der Donor-Recipient Distanzsumme niemals unterlegen ist, und dass

primär die Schätzung der multivariaten Parameter verbessert wird. Die Schätzung keiner der univariaten Paremeter wird durch die Wahl der Spenderauswahlstrategie substantiell beinflusst. Bei einem Verleich der Tabelle 2 und Tabelle 1 zeigt sich, dass hier die selben Effekte vorhanden sind, nur stärker. Die Optimierung wird wichtiger, je mehr Werte fehlen, je stärker die Innerklassenkorrelation ist und umso weinger Objekte die Datenmatrix umfasst.

		Univariat				Multivariat		
Faktor	Faktorstufen	$\bar{x}$	sd_x	s_x	k_x	r_{xy}	$b_{1,xy}$	$b_{2,xy}$
Objekt-anzahl	50	—	—	—	—	0,009	0,144	0,160
	100	—	—	—	—	0,007	0,132	0,150
Innerklassen-korrelation	0,00	—	—	—	—	0,003	0,038	0,035
	0,35	—	—	—	—	0,008	0,096	0,105
	0,70	—	—	—	—	0,012	0,279	0,324
Anteil fehlender Werte	10%	—	—	—	—	—	—	0,001
	20%	—	—	—	—	—	0,001	0,004
	30%	—	—	—	—	—	0,005	0,015
	40%	—	—	—	—	0,004	0,166	0,342
	50%	—	—	—	—	0,034	0,516	0,414

Tabelle 2: RMSE Differenzen für *MAR 1:2* Daten, Haupteffekte

3.2.4 Ergebnisse der Simulation für *MAR 1:4* Daten

Bei der Betrachtung von Tablle 3 setzt sich das homogene Bild der Ergebnisse fort. Auch für den dritten Ausfallmechanismus ist Vogels Approximationsmethode niemals schlechter als die naive Vorgehensweise. Beim *MAR 1:4* Mechanismus fehlten in einem Cluster vier Mal soviele Daten wie in dem anderen Cluster, daher stellt dieser Mechanismus eine Verschärfung des *MAR 1:2* Mechanismus dar. Diese Verschärfung führt zu noch deutlicheren Verbesserungen, wenn Vogels Approximationsmethode verwendet wird. Wieder sind die Vorteile bei der Schätzung der multivariaten Parameter zu finden, wobei die Tendenzen genau denen in den *MCAR* und *MAR 1:2* Fällen entsprechen.

Faktor	Faktorstufen	Univariat				Multivariat		
		$\bar{x}$	sd_x	s_x	k_x	r_{xy}	$b_{1,xy}$	$b_{2,xy}$
Objekt-anzahl	50	––	––	––	––	0,018	0,149	0,133
	100	––	––	––	––	0,015	0,150	0,141
Innerklassen-korrelation	0,00	––	––	––	––	0,008	0,040	0,030
	0,35	––	––	––	––	0,016	0,107	0,095
	0,70	––	––	––	––	0,025	0,302	0,286
Anteil fehlender Werte	10%	––	––	––	––	––	––	0,001
	20%	––	––	––	––	––	0,002	0,005
	30%	––	––	––	––	0,001	0,028	0,069
	40%	––	––	––	––	0,013	0,357	0,120
	50%	––	––	––	––	0,069	0,360	0,492

Tabelle 3: RMSE Differenzen für *MAR 1:4* Daten, Haupteffekte

4 Schlussfolgerungen

Dieser Aufsatz erläutert den als Donor-Limit bekannten Aspekt von Hot-Deck Imputationsmethoden und wie eine Verwendung des Donor-Limit zu einem Versagen konventioneller Zuordnungsmethoden führt. Zusätzlich zu der Formulierung des Optimierungsproblems wird eine Simulationsstudie, bei der die jetzige Standardmethode und eine Heuristik miteinander verglichen werden, präsentiert.

Die Ergebnisse der Monte Carlo Simulation zeigen, dass die optimierte Methode niemals schlechter ist als die naive Zuordnungsmethode, die heutzutage verwendet wird, wenn ein Donor-Limit eingesetzt wird. Tendenzen, dass die Vorteile stärker werden, sind über alle drei betrachteten Ausfallmechanismen konsistent. Die Vorteile steigen immer, je weniger Donor in den Daten vorhanden sind und je stärker die Korrelation der Variablen innerhalb der Klassen ist. Vorteile des neuen Ansatzes werden kleiner, je kleiner die Datenmatrix ist und je weniger schlimm der Ausfallmechanismus ist.

Am interessantesten ist wahrscheinlich, dass nur die Schätzung multivariater Parameter substantiell durch die Optimierung verbessert wird. Dies deutet darauf hin, dass der primäre Wert des erhöhten Aufwands die gesamten Donor-Recipient Zuordnungskosten zu minimieren in der besseren Erhaltung von multivariaten Eigenschaften der Daten liegt. Dies ist besonders dann wichtig, wenn die Daten, nach Imputation, mittels multivariater Verfahren, z.B. multiple Regression, ausgewertet werden sollen.

Die Ergebnisse dieser Arbeit lassen sich für eine breite Menge an Situationen und Anwendungen verallgemeinern. Nicht nur wurden die wichtigsten Faktoren in der Simulationsstudie berücksichtigt, sondern auch die Menge an Annahmen, die getroffen wurden, ist klein. Trotzdem existieren Einschränkungen der Verallgemeinerbarkeit der Ergebnisse. Erstens, während die Ausfallmechanismen, die hier betrachtet wurden, dass zwei Untergruppen unterschiedliche Ausfallwahrscheinlichkeiten haben, viele Situationen gut abdeckt, sind weitaus mehr MAR Mechanismen denkbar. Zweitens, kann in der Praxis eine Hot-Deck Imputation mit einem Donor-Limit von eins nicht durchführbar sein. Durch wie in der Praxis die Imputationsklassen konstruiert werden und wenn mehrere Variablen fehlende Werte aufweisen, kann die Menge an Recipients leicht größer werden als die Menge an Donor. Eine Überprüfung weniger stringenter Donor-Limits sollte weitere interessante Ergebnisse liefern. Drittens, es wurde lediglich eine Datenstruktur betrachtet. Datenstrukturen, die in der Realität existieren, sind komplexer, und daher (insbesondere wegen der Präsenz von Ausreißern) reagieren sie sensib-

ler auf die Zuordnung schlechter Donor. Viertens, in dieser Arbeit wurden nur die Ergebnisse der konventionellen Methode mit jenen einer Heuristik, die reihenfolgeunabhängige Ergebnisse erzeugt, verglichen. Zukünftige Arbeiten könnten in einem weiteren Schritt nicht nur die zwei Heuristiken, sondern auch eine optimale Methode in Hinblick auf Imputationsgüte, aber auch in Hinblick auf Berechnungszeit und –komplexität, vergleichen. Des Weiteren könnte die spezielle Struktur des ganzzahligen Optimierungsproblems ausgenutzt werden und spezialisierte, effiziente Algorithmen zur Lösung dieses speziellen Problems entwickelt werden.

Die Schlussfolgerungen und Resultate dieser Arbeit sollten nicht nur für Akademiker von Interesse sein, sondern auch für Praktiker, die multivariate Analysemethoden auf Hot-Deck imputierten Daten verwenden.

Literaturverzeichnis

Andridge, R.R. and Little, R.J.A. (2010): A Review of Hot Deck Imputation for Survey Nonresponse. *International Statistical Review*, 78, pp. 40-64.

Bankhofer U. and Joenssen, D.W. (2014): On Limiting Donor Usage for Imputation of Missing Data via Hot Deck Methods. *In: M. Spiliopoulou, L. Schmidt–Thieme, and R. Jannings (Eds.): Data Analysis, Machine Learning and Knowledge Discovery*. Berlin: Springer, pp. 3-11.

Collins, L., Schafer, J. and Kam, C. (2001): A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, 6, pp. 330-351.

Domschke, W. (1995): *Logistik: Transport*. Oldenbourg, München.

Enders, C.K. (2010): *Applied Missing Data Analysis*. New York: Guilford Press.

Ford B. (1983): An Overview of Hot-Deck Procedures. *In: W. Madow, H. Nisselson, and I. Olkin (Eds.): Incomplete Data in Sample Surveys*. New York: Academic Press, pp. 185-207.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2013): mvtnorm: Multivariate Normal and t Distributions. *R package version 0.9-9995*. URL http://CRAN.R-project.org/package=mvtnorm

Joenssen, D.W. (2013): HotDeckImputation: Hot Deck Imputation Methods for Missing Data. *R package version 0.1.0*. URL http://CRAN.R-project.org/package=HotDeckImputation

Kalton, G. and Kish, L. (1984): Some Efficient Random Imputation Methods. *Communications in Statistics - Theory and Methods*, 13, pp. 1919-1939.

Kovar, J.G. and Whitridge, J. (1995): Imputation of Business Survey Data. *In: B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (Eds.): Business Survey Methods*. New York: Wiley, pp. 403-423.

Little, R.J.A. and Rubin, D.B. (2002): *Statistical Analysis with Missing Data*. Hoboken: Wiley.

R Core Team (2013): *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org/

Reinfeld, N.V. and Vogel, W.R. (1958): *Mathematical Programming*. New Jersey: Prentice–Hall.

Rubin, D.B. (1976): *Inference and Missing Data (with Discussion)*. Biometrika, 63, pp. 581-592.

Sande I. (1983): Hot–Deck Imputation Procedures. *In: W. Madow, H. Nisselson, and I. Olkin (Eds.): Incomplete Data in Sample Surveys*. New York: Academic Press, pp. 339-349.

Schafer, J. and Graham, J. (2002): Missing Data: Our View of the State of the Art. *Psychological Methods*, 7, pp. 147–177.