AUSTRIAN JOURNAL OF STATISTICS
Volume 34 (2005), Number 3, 263–281

Nonresponse in Bevölkerungsumfragen – Österreich-Ergebnisse einer Simulationsstudie im Rahmen des EU-Projekts DACSEIS

Andreas Quatember
Johannes Kepler Universität Linz

Zusammenfassung: Das EU-Projekt DACSEIS beschäftigt sich mit der Datenqualität in komplexen Bevölkerungsstichproben. In einer Simulationsstudie wurden aus einer zu diesem Zweck aus Daten eines österreichischen Mikrozensuses (AMC) erzeugten Pseudo-Grundgesamtheit Stichproben entnommen. Das Ziehen der Stichproben erfolgte nach einem den wichtigsten Bausteinen des komplexen Stichprobendesigns des AMC bis 2003 nachgebildeten Stichprobenverfahren. In diesen Stichproben wurde nach einem vorgegebenen Mechanismus Nonresponse unterschiedlichen Ausmaßes erzeugt.

Die Ergebnisse zeigen, dass der AMC-Schätzer (ohne iteratives proportionales Anpassen) für Anzahlen beim verwendeten Stichprobenverfahren ungenauer ist als er bei uneingeschränkter Zufallsauswahl wäre und dass eine qualitativ gute direkte Varianzschätzung bei vollem Response möglich ist. Die Ausgewichtung des Nonresponses erzeugt verzerrte Schätzer, deren Varianz sich ebenfalls direkt schätzen lässt. Die Verwendung adäquater Hilfsinformationen für verschiedene Imputationstechniken (Single und Multiple Imputation) ermöglicht die „Entzerrung“ der Schätzer. Die beste Varianzschätzung für die „imputierten Schätzer“ erfolgt innerhalb der getesteten Varianzschätzmethoden bei Single Imputation mittels geeigneter Bootstrapverfahren bzw. durch die implizite Schätzung bei Multipler Imputation.

Abstract: The subject of the EU-project DACSEIS is data quality in complex surveys. In a simulation study samples have been selected out of a pseudo-population, which was generated from the data of one Austrian Microcensus (AMC). This selection was done according to the most important facts of the sampling design of the AMC until 2003. Nonresponse of different amount was added to these samples according to a predefined mechanism.

The simulation results show, that the AMC-estimator (without iterative proportional fitting) for a number of units, given the sampling method used, is less efficient than it would be with simple random sampling and that it is possible to give a direct variance estimation for it, when there is full response. Weighting for nonresponse produces biased estimators, for which it is still possible to have a good direct variance estimation. But the bias of the AMC-estimator decreases, when appropriate auxiliary information for different imputation techniques (Single and Multiple Imputation) is used. The best estimation of the variance of such “imputed estimators” within the used variance estimation techniques are appropriate bootstrap techniques for single imputation and the implied variance estimation for multiple imputation.

Keywords: DACSEIS, Sampling Theory, Survey Methodology, Data Quality, Nonresponse, Weighting, Imputation.

1 Einleitung

Das EU-Projekt DACSEIS (IST-2000-26057) ist ein von März 2001 bis Mai 2004 dauerndes Forschungsprojekt im 5. Rahmenprogramm der Europäischen Union im Bereich der Forschung und technologische Entwicklung, an dem Institutionen der amtlichen und der universitären Statistik aus sieben europäischen Ländern beteiligt waren – namentlich aus Finnland, Deutschland, Großbritannien, den Niederlanden, Norwegen, dem Nicht-EU-Mitglied Schweiz und aus Österreich. Als Projekt-Koordinator fungierte Ralf Münnich von der Abteilung Statistik, Ökonometrie und Unternehmensforschung der Eberhard Karls Universität in Tübingen. Der österreichische Anteil an DACSEIS lag in den Händen des IFAS - *Institut für Angewandte Statistik* der Johannes Kepler Universität Linz, an dem der Autor dieses Aufsatzes das Projekt leitete. Außerdem waren die Projektassistentin Doris Eckmair und die IFAS-Mitarbeiterin Helga Wagner von Seiten des IFAS maßgeblich am Projekt beteiligt. Dieses IFAS-Team bekam von amtlicher Seite durch Alois Haslinger von der STATISTIK AUSTRIA die für die Bewältigung der Aufgaben nötigen Basisinformationen über den österreichischen Mikrozensus (AMC; Austrian Microcensus).

Das Projekt DACSEIS – eine Abkürzung für „Data Quality in Complex Surveys within the European Information Society“ – lässt sich dem Bereich der statistischen Stichprobentheorie zuordnen. Es setzte sich mit der Problematik der Datenqualität in komplexen Bevölkerungsstichproben am Beispiel der Arbeitskräfteerhebungen der Europäischen Union in den beteiligten Ländern auseinander (für Details zu den verschiedenen Aufgabenstellungen des Projekts siehe Münnich and Wiegert, 2001). Im Speziellen bedeutete dies für Österreich die Untersuchung des AMC, der zu Projektbeginn einmal jährlich die nationale EU-Arbeitskräfteerhebung Österreichs beinhaltete. Mit Beginn des Jahres 2004 wurde einer EU-Richtlinie folgend auch in Österreich auf eine kontinuierliche Erhebung der Arbeitskräfte umgestellt (vgl. Kytir and Stadler, 2004).

2 Aufgabenstellung

Die Aufgaben des IFAS im Rahmen des Projekts umfassten u.a. folgende Punkte: Zu allererst musste aus den von der Statistik Austria erworbenen Daten des AMC des ersten Quartals 2001 eine Grundgesamtheit generiert werden, da sich die Verwendung von Zensusdaten auf Datenschutzgründen verbot. Aus dieser „Pseudo-Grundgesamtheit“ konnten im darauf folgenden Schritt zu Simulationszwecken Stichproben gezogen werden. Um zu gewährleisten, dass diese Pseudo-Grundgesamtheit der tatsächlichen Bevölkerung hinsichtlich der verwendeten Merkmalsauswahl möglichst nahe kommt, wurden innerhalb der Wohnungen dieser Pseudo-Grundgesamtheit reelle Strukturen hinsichtlich der Anzahlen der Haushalte und der in diesen lebenden Personen nachgebildet. Die Ausprägungen der Merkmale Alter und Geschlecht von Erhebungseinheiten in vorgegebenen Schichten bzw. PSUs (siehe unten) wurden den betreffenden vorliegenden AMC-Teilgesamtheiten nachempfunden, um unplausible Kombinationen dieser Merkmale zu vermeiden. Die übrigen für die Simulationen ausgewählten Personenmerkmale Ausbildung, Erwerbstätigkeit und Staatsbürgerschaft wurden auf Basis der bedingten empirischen Verteilung dieser Merkmale unter der Bedingung Alter und Geschlecht ebenfalls schichten- bzw. PSU-weise generiert (für eine detaillierte Beschreibung des Generie-

rungsvorgangs siehe Wagner, 2003, bzw. Wagner and Eckmair, 2004). So waren schließlich zu jeder Wohnung regionale Merkmale wie das Bundesland und eine regional definierte Schicht und zu jedem Bewohner diese fünf Personenmerkmale in der Pseudo-Grundgesamtheit vorhanden.

Aus dieser Pseudo-Grundgesamtheit wurden zu Simulationszwecken 10.000 Stichproben nach einem komplexen Stichprobenverfahren gezogen, das als Grundgerüst des tatsächlich für den AMC zu diesem Zeitpunkt verwendeten Stichprobenverfahrens bezeichnet werden kann (zu den verschiedenen Stichprobenverfahren der im Projekt untersuchten Arbeitskräfteerhebungen siehe Quatember, 2002, und zu Details des Stichprobenverfahrens des AMC bis 2003 siehe Haslinger, 1996). Manche Eigenheiten der Auswahl der Elemente für die Stichprobe im AMC (z.B. auch die „Rotation“) waren nicht simulierbar. Auch die Gruppe der Personen in Anstalten wurde nicht berücksichtigt. Als Auswahleinheiten fungierten wie im AMC Wohnungen, wobei die Pseudo-Grundgesamtheit genauso wie der tatsächliche Auswahlrahmen für den AMC sowohl bewohnte Wohnungen als auch unbewohnte Wohnungen enthielt.

Die Pseudo-Grundgesamtheit aller Wohnungen wurde zuerst in neun Bundesländer zerlegt. In sieben der neun Bundesländer wurden die Wohnungen weiters geschichtet in einen (eher städtischen) Bereich 1 und einen (eher ländlichen) Bereich 2. In zwei Bundesländern (es waren dies Vorarlberg und Wien) gab es – dem AMC folgend – ausschließlich Wohnungen im Bereich 1.

Im Bereich 1 innerhalb jedes Bundeslandes wurden die Wohnungen nochmals regional geschichtet. Innerhalb dieser Schichten wurde im Bereich 1 eine uneingeschränkte Zufallsauswahl von Wohnungen gezogen. Innerhalb der ausgewählten Wohnungen wurden schließlich, sofern die Wohnung bewohnt war, die Personen voll erhoben.

Im Bereich 2 galt hinsichtlich der Schichtung innerhalb der Bundesländer dasselbe wie für Bereich 1. Innerhalb dieser regionalen Schichten wurde nun aber zweistufig vorgegangen: Zuerst wurde eine uneingeschränkt zufällige Auswahl an PSUs (primary sampling units; Klumpen von Wohnungen) entnommen und innerhalb dieser PSUs eine uneingeschränkt zufällige Auswahl an Wohnungen als SSUs (secondary sampling units). In diesen Wohnungen wurden dann wieder alle Personen, sofern welche vorhanden, erhoben (siehe zur Veranschaulichung Abbildung 1).

Sowohl hinsichtlich der Pseudo-Grundgesamtheit als auch hinsichtlich des Stichprobenverfahrens wurde also versucht, so nahe wie möglich am AMC zu simulieren. Dennoch ist die generierte Grundgesamtheit nicht die tatsächliche und das verwendete komplexe Stichprobenverfahren entspricht auch nicht exakt dem tatsächlichen Auswahlverfahren des AMC (siehe oben). Die Ergebnisse der Simulationsstudie dürfen deshalb *von den Zahlenwerten her* nicht mit den vorhandenen realen Daten werden. Die *Schlussfolgerungen daraus* sind jedoch für die Praxis solcher komplexer Stichprobenerhebungen – sei es nun in Österreich oder anderswo – relevant.

Nonresponse, also Antwortausfälle, wurde für die Simulationen wohnungsweise und nur für die interessierende Variable (Item-Nonresponse) in die Stichprobe eingebaut. Er betraf also jeweils alle Personen einer bewohnten Wohnung. Da über den Nonresponse im AMC keine über sein Ausmaß in den verschiedenen Schichten hinausgehende Informationen zur Verfügung standen, wurden solche über den deutschen Mikrozensus (vgl. dazu Münnich, 2003, S.70) für den verwendeten Nonresponse-Mechanismus auf die vorgege-

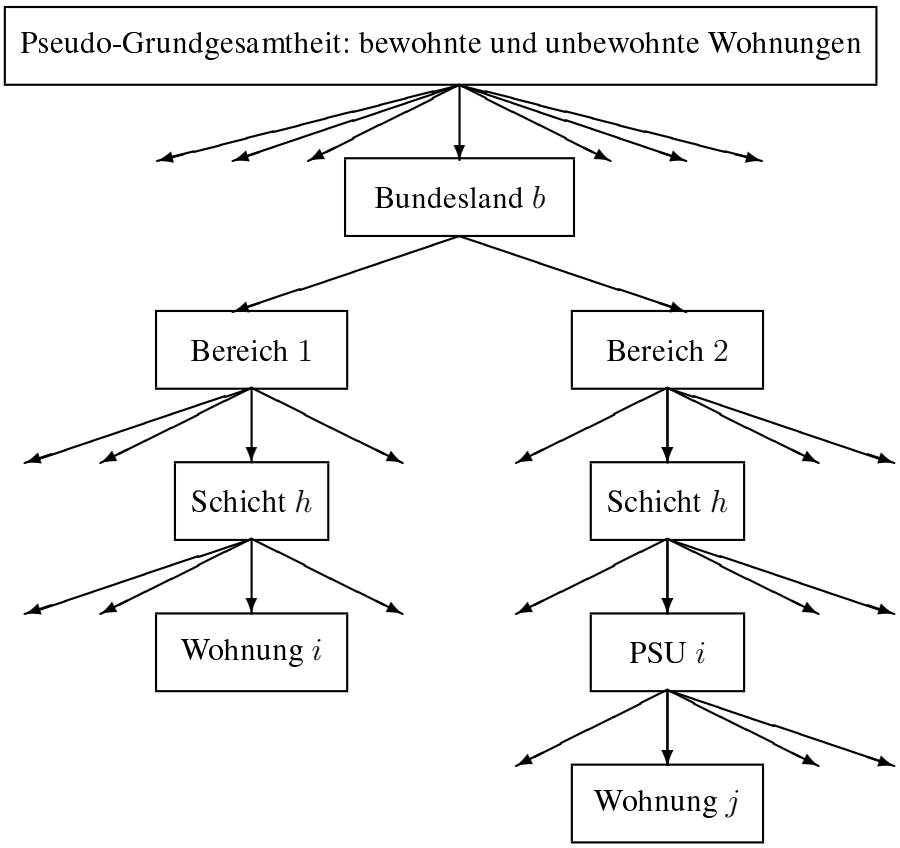

Abbildung 1: Schematische Darstellung des verwendeten Stichprobenverfahrens

benen österreichischen Verhältnisse adaptiert. In der AMC-Stichprobe aus dem Jahr 2001 kam in 3041 von 26512 aufgesuchten bewohnten Wohnungen wegen des Nichtantreffens bzw. der Teilnahmeverweigerung der Bewohner kein Interview zu Stande. Das in der Praxis zu diesem Zeitpunkt auftretende Problem der Interviewerausfälle, das im besagten AMC eine Verminderung der Anzahl an aufgesuchten Wohnungen (bewohnt und unbewohnt) um 2788(!) Wohnungen führte, wurde in den Simulationen nicht nachgebildet, da in den erworbenen Daten Wohnungen nicht einzelnen Interviewern zugeordnet waren und uns dieses Problem zu erhebungsspezifisch erschien. Nach einem „Missing-at-Random-Modell“ für die fehlenden Werte wurden über die Ausprägungen bestimmter Merkmale einer Referenzperson allen Bewohnern einer Wohnung zufällig eine dichotome Zufallsvariable zugewiesen, deren Ausprägung angibt, ob für die betreffende Erhebungseinheit bei der interessierenden Variablen ein Wert vorliegt oder nicht. Die Antwortwahrscheinlichkeit einer Referenzperson wurde bedingt durch ihre Ausprägungen bei den Merkmalen Region REG (Wien/Restösterreich), Staatsbürgerschaft (NAT, österreichisch/nichtösterreichisch), Haushaltsgröße (HS, 1/2/3+), Geschlecht (SEX, m/w) und Alter (AGE, < 60/ 60+). Diese Informationen könnten in der Praxis z.B. aus einem Melderegister stammen.

Die Basis für die Bestimmung der bedingten Wahrscheinlichkeiten bildete ein Antwort-Homogenitäts-Gruppen-Modell (response homogeneity group model), für das diesbezügliche Informationen aus Hamburg für Wien und aus Bayern für den Rest von Österreich auf die österreichischen Verhältnisse adaptiert wurden (Tabelle 1). In diesen Gruppen be-

Tabelle 1: Bedingte wohnungsweise Response-Wahrscheinlichkeiten für die Simulationen des Nonresponses

REG	NAT	HS	SEX	AGE	Response-Wahrscheinlichkeit
Wien	ö	1	m	< 60	0.805
	ö	1	m	60+	0.850
	ö	1	w	< 60	0.851
	ö	1	w	60+	0.847
	ö	2			0.856
	ö	3+			0.859
	nö	1			0.840
	nö	2			0.849
	nö	3+			0.849
Rest	ö	1	m	< 60	0.871
	ö	1	m	60+	0.810
	ö	1	w	< 60	0.893
	ö	1	w	60+	0.860
	ö	2			0.893
	ö	3+			0.894
	nö	1			0.849
	nö	2			0.885
	nö	3+			0.894

saß damit jede Auswahleinheit (Wohnung) die gleiche Response-Wahrscheinlichkeit.

Damit ergab sich eine wohnungsbezogene Nonresponse-Rate von 11.47% in ganz Österreich. Dies ist auch die Rate im AMC des ersten Quartals des Jahres 2001. Für die Simulation weiterer Gesamt-Nonresponse-Raten (5, 25 bzw. 40%), wurden die Antwortwahrscheinlichkeiten aus Tabelle 1 proportional an diese angepasst.

Als interessierendes Merkmal wurde naheliegenderweise die Erwerbstätigkeit nach dem Labour Force Concept bestimmt und zu schätzen war die Anzahl an Erwerbstätigen. Durch die Simulationen sollten die statistischen Eigenschaften (Erwartungswerte, Varianzen) des zu diesem Zweck im AMC verwendeten Schätzers unter vollem Response und bei verschieden hohen Nonresponse-Raten untersucht werden. Auch die Varianzschätzung für diese Schätzer und die Untersuchung ihrer statistischen Eigenschaften waren Aufgabe der Simulationsstudien. Ferner wurden diese Eigenschaften der Schätzer bei verschiedenen Imputationstechniken betrachtet und auch für diese Schätzer die Schätzung der Varianz untersucht. Die für die Simulationen benötigten Programme wurden in `R` erstellt.

3 Ergebnisse der Simulationsstudie

3.1 Erwartungswert und Varianz des AMC-Schätzers für eine Merkmalssumme bei vollem Response

Der Schätzer $\hat{t}$ für eine Merkmalssumme t (z.B. die Anzahl der Erwerbstätigen) im AMC ist wegen der bundesländerweisen Zerlegung und der weiteren Teilung der Bundesländer in zwei Bereiche (siehe Abschnitt 2) darstellbar als (die im Folgenden verwendeten Notationen sind eine Kombination aus den in der Stichprobentheorie von Särndal et al., 1992, und den für den AMC verwendeten von Haslinger, 1996)

$$\hat{t} = \sum_{b=1}^{9} \hat{t}_b = \sum_{b=1}^{9} (\hat{t}_{b1} + \hat{t}_{b2}) \,. \tag{1}$$

Die Merkmalssummenschätzer $\hat{t}_b$ beziehen sich jeweils auf ein Bundesland $b, b = 1, \ldots, 9$. Die Schätzer $\hat{t}_{b1}$ bzw. $\hat{t}_{b2}$ sind die Schätzer für die Merkmalssumme im Bereich 1 bzw. im Bereich 2 des b-ten Bundeslands. Der herkömmliche Horvitz-Thompson-Schätzer $\hat{t}_{HT}$ für eine Merkmalssumme t ist gegeben durch

$$\hat{t}_{HT} = \sum\nolimits_s d_k y_k \tag{2}$$

mit dem Designgewicht d_k, das der Ausprägung y_k des k-ten Elements der Stichprobe s beim Merkmal y zugeordnet ist (vgl. etwa Lundström and Särndal, 2002, S.45). d_k entspricht dem Reziprokwert der Auswahlwahrscheinlichkeit des k-ten Elements und gibt an, wie viele Elemente der Grundgesamtheit durch das k-te Stichprobenelement repräsentiert werden.

$\hat{t}_{b1}$ ist im betrachteten AMC der Horvitz-Thompson-Schätzer für die Merkmalssumme in der Teilgesamtheit $b1$

$$\hat{t}_{b1} = \sum_{h=1}^{H_{b1}} \frac{N_{b1h}}{n_{b1h}} \, q_{b1h} \sum_{i=1}^{n^*_{b1h}} y_{b1hi} \,. \tag{3}$$

Es gilt in (3):

- N_{b1h} ... die Gesamtzahl aller Wohnungen in der h-ten von insgesamt H_{b1} Schichten des Bereichs 1 im Bundesland b
- n_{b1h} ... die Anzahl der Wohnungen in der Stichprobe aus Schicht $b1h$
- q_{b1h} ... dieser Faktor kompensiert den wohnungsweisen Nonresponse der bewohnten Wohnungen in Schicht $b1h$. Es gilt $q_{b1h} = (n^{(1)}_{b1h} + n^{(2)}_{b1h})/n^{(1)}_{b1h}$ mit
- $n^{(1)}_{b1h}$ ... die Anzahl der respondierenden Wohnungen (ohne unbewohnte Wohnungen) in der Stichprobe aus Schicht $b1h$
- $n^{(2)}_{b1h}$ ... die Anzahl der nichtrespondierenden Wohnungen in der Stichprobe aus Schicht $b1h$
- n^*_{b1h} ... die Anzahl der respondierenden Wohnungen plus der unbewohnten Wohnungen in der Stichprobe aus Schicht $b1h$. Es gilt also $n^*_{b1h} = n_{b1h} - n^{(2)}_{b1h}$
- y_{b1hi} ... die Merkmalssumme (z.B. die Anzahl der Erwerbstätigen) in der i-ten respondierenden Stichprobenwohnung aus Schicht $b1h$.

Durch Multiplikation mit dem voran gestellten Faktor q_{b1h} wird der Nonresponse von Wohnungen einer Schicht kompensiert (bei vollem Response ist $n^{(2)}_{b1h} = 0$ und $q_{b1h} = 1$). Dieser Vorgehensweise zur Ausgewichtung des Nonresponses liegt ein Antwort-Homogenitäts-Gruppen-Modell zu Grunde, in dem auftretender Nonresponse genau dann unverzerrt kompensiert wird, wenn innerhalb der – hier geographisch definierten – Gruppen die Antwortwahrscheinlichkeiten der Elemente dieser Gruppe gleich sind. Dieses Nonresponse-Modell wird als „Missing-at-Random" bezeichnet.

Als Merkmalssummenschätzer $\hat{t}_{b2}$ für die Merkmalssumme im Bereich 2 des b-ten Bundeslandes wurde beim AMC der nachfolgende Schätzer verwendet, der *nicht* der Horvitz-Thompson-Schätzer dieser Merkmalssumme ist:

$$\hat{t}_{b2} = \begin{cases} \sum\limits_{h=1}^{H_{b2}} \frac{N_{b2h}}{\sum\limits_{i=1}^{c_{b2h}} N_{b2hi}} \sum\limits_{i=1}^{c_{b2h}} \frac{N_{b2hi}}{n_{b2hi}}\, q_{b2hi} \sum\limits_{j=1}^{n^{(1)}_{b2hi}} y_{b2hij}\,, & \text{für } b = 1,\ldots,7\,, \\ 0\,, & \text{für } b = 8,9\,. \end{cases} \tag{4}$$

In den beiden Bundesländern 8 (Wien) und 9 (Vorarlberg) sind alle Wohnungen der Grundgesamtheit im Bereich 1 enthalten. Es ist in (4):

- N_{b2h} ... die Gesamtzahl aller Wohnungen in der h-ten von insgesamt H_{b2} Schichten des Bereichs 2 im Bundesland b
- N_{b2hi} ... die Gesamtzahl der Wohnungen (inklusive unbewohnte Wohnungen) in der i-ten Stichproben-PSU der Schicht $b2h$
- n_{b2hi} ... die Anzahl der Wohnungen in der Stichprobe aus der i-ten Stichproben-PSU der insgesamt c_{b2h} PSUs (c steht für cluster) aus Schicht $b2h$
- q_{b2hi} ... Dieser Faktor kompensiert den wohnungsweisen Nonresponse der bewohnten Wohnungen in der i-ten Stichproben-PSU der Schicht $b2h$. Es gilt $q_{b2hi} = (n^{(1)}_{b2hi} + n^{(2)}_{b2hi})/n^{(1)}_{b2hi}$ mit
- $n^{(1)}_{b2hi}$ ... die Anzahl der respondierenden Wohnungen (ohne unbewohnte Wohnungen) in der Stichprobe aus der i-ten Stichproben-PSU der Schicht $b2h$
- $n^{(2)}_{b2hi}$ ... die Anzahl der nicht-respondierenden Wohnungen in der Stichprobe aus der i-ten Stichproben-PSU der Schicht $b2h$
- n^{*}_{b2hi} ... die Anzahl der respondierenden Wohnungen plus der unbewohnten Wohnungen in der Stichprobe aus der i-ten Stichproben-PSU der Schicht $b2h$. Es gilt also $n^{*}_{b2hi} = n_{b2hi} - n^{(2)}_{b2hi}$
- y_{b2hij} ... die Merkmalssumme (z.B. die Anzahl der Erwerbstätigen) in der j-ten respondierenden Stichprobenwohnung aus der i-ten Sichproben-PSU der Schicht $b2h$.

Beim verwendeten Stichprobenverfahren wäre beim Horvitz-Thompson-Schätzer für die Merkmalssumme t_{b2} der erste Quotient in (4) nicht

$$\frac{N_{b2h}}{\sum\limits_{i=1}^{c_{b2h}} N_{b2hi}}\,,$$

sondern die reziproke Auswahlwahrscheinlichkeit für eine PSU bei uneingeschränkter Zufallsauswahl von c_{b2h} PSUs aus den C_{b2h} PSUs der Schicht $b2h$

$$\frac{C_{b2h}}{c_{b2h}}\,.$$

Wie man sieht, wurde im AMC bei der Hochrechnung der Stichprobendaten im Bereich 2 jedes Bundeslandes eine Genauigkeitsverbesserung im Vergleich zum Horvitz-Thompson-Schätzer dadurch angestrebt, dass die Größe der Stichproben-PSUs im Schätzer mitberücksichtigt wird (vgl. dazu Särndal et al., 1992, S.144). (Im AMC wurde zur weiteren Erhöhung der Genauigkeit noch eine zusätzliche Korrektur dieser Hochrechungsgewichte durch iteratives proportionales Anpassen (iterative proportional fitting) der aus der Stichprobe hochgerechneten bundesländerweisen Verteilungen des zweidimensionalen Merkmals Alter und Geschlecht bzw. des Merkmals Nationalität an die betreffenden Populationsverteilungen aus der Bevölkerungsfortschreibung durchgeführt (vgl. Haslinger, 1996, S.319f). Darauf wurde aus Zeitgründen in dieser Studie verzichtet.)

Da die Summe $\sum_{i=1}^{c_{b2h}} N_{b2hi}$ eine Zufallsvariable ist, deren Größe davon abhängt, welche PSUs in die Stichprobe gelangt sind, ist eine formale Angabe der exakten Varianz des Schätzers unmöglich. In 10000 Simulationen von AMC-Stichproben bei vollem Response ergaben sich für die ganze Pseudo-Grundgesamtheit (Österreich) bzw. für ein aus dieser Pseudo-Grundgesamtheit ausgewähltes Bundesland mit beiden Bereichen (Burgenland) folgende Kennzahlen der Verteilung des Schätzers $\hat{t}$ nach (1) für die Anzahl an Erwerbstätigen:

Tabelle 2: Kennzahlen der Verteilung von $\hat{t}$

	t	$M_{SIM}(\hat{t})$	$V_{SIM}(\hat{t})$
Österreich	3138606	3139100	$540.54 \cdot 10^6$
Burgenland	106273	106380	$5.28 \cdot 10^6$

$\hat{t}$ nach (1) ist wie der Horvitz-Thompson-Schätzer natürlich unverzerrter Schätzer von t. Die geringen Abweichungen der Mittelwerte $M_{SIM}(\hat{t})$ von $\hat{t}$ vom Parameter t in den Simulationen sind simulationsbedingte Abweichungen. Die in Tabelle 2 enthaltenen, aus den 10000 Simulationen berechneten Schätzervarianzen $V_{SIM}(\hat{t})$ sind die Referenzwerte zur Prüfung der Tauglichkeit der Varianzschätzer im Nachfolgenden. Zur Veranschaulichung der Genauigkeit der Anzahlschätzer in der zu Grunde liegenden Pseudo-Grundgesamtheit, kann die Schwankungsbreite der Stichprobenergebnisse zur Sicherheit 95% angegeben werden, also der approximative 95%-Zufallsstreifen für $\hat{t}$. Dieser beträgt für ganz Österreich 3138606 $\pm$ 45569 und für das Bundesland Burgenland 106273 $\pm$ 4504.

3.2 Direkte Varianzschätzung bei vollem Response

Für eine ohne weitere Manipulation der Stichprobe direkt aus den Daten zu berechnende Schätzung der Referenzwerte für die Varianz von $\hat{t}$ (siehe Tabelle 2) stellte sich nicht überraschend diejenige als beste heraus, die sowohl das verwendete Stichprobenverfahren als auch die verwendete Schätzmethode am stärksten berücksichtigt. In Tabelle 3 wird für einen Horvitz-Thompson-Schätzer die einfache Varianzschätzung $\hat{V}_{SI}(\hat{t})$ bei uneingeschränkter Zufallsauswahl (SI steht für simple random sampling) von Personen (vgl. Eckmair, 2004, S.98ff), bei der die vielfältigen Effekte der Stichprobenziehung und

Tabelle 3: Vergleich von $V_{SIM}(\hat{t})$ mit zwei direkten Schätzverfahren für diese Varianz

	$V_{SIM}(\hat{t})$	$M_{SIM}[\hat{V}_{SI}(\hat{t})]$	$M_{SIM}[\hat{V}^*_{HT}(\hat{t})]$
Österreich	$540.54 \cdot 10^6$	$184.91 \cdot 10^6$ (74.8%)	$597.87 \cdot 10^6$ (96.1%)
Burgenland	$5.28 \cdot 10^6$	$2.56 \cdot 10^6$	$5.57 \cdot 10^6$

Schätzung ignoriert werden, auf Basis ihrer Mittelwerte $M_{SIM}[\hat{V}_{SI}(\hat{t})]$ in den Simulationen mit der komplexen Varianzschätzung beim verwendeten Verfahren verglichen. Der Schätzer $\hat{V}_{HT}(\hat{t})$ für die Varianz eines Horvitz-Thompson-Schätzers wird im Allgemeinen dargestellt durch

$$\hat{V}(\hat{t}_{HT}) = \sum\sum_{s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{5}$$

mit Δ_{kl}, der Kovarianz der Aufnahmeindikatoren zweier Elemente k und l, die angeben, ob ein Element der Stichprobe angehört oder nicht. π_{kl} ist die Aufnahmewahrscheinlichkeit zweiter Ordnung, die angibt mit welcher Wahrscheinlichkeit beide Elemente k und l in die Stichprobe gelangen (vgl. etwa Särndal et al., 1992, S.43f). Wir bezeichnen den Varianzschätzer (5) bei einer uneingeschränkten Zufallsauswahl mit $\hat{V}_{SI}(\hat{t})$ und beim verwendeten komplexen Stichprobenverfahren mit $\hat{V}^*_{HT}(\hat{t})$. Letzterer besitzt folgendes Aussehen, da er sich wegen der Unabhängigkeit der Ziehungen in den Bundesländern und Bereichen natürlich wie der Schätzer $\hat{t}$ selbst bundesländer- und bereichsweise zerlegen lässt

$$\begin{aligned}
\hat{V}^*_{HT}(\hat{t}) &= \sum_b \hat{V}_{HT}(\hat{t}_b) = \sum_b [\hat{V}_{HT}(\hat{t}_{b1}) + \hat{V}_{HT}(\hat{t}_{b2})] \\
&= \sum_b \left\{ \sum_{h=1}^{H_{b1}} \frac{N^2_{b1h}}{n_{b1h}} \left(1 - \frac{n_{b1h}}{N_{b1h}}\right) s^2_{b1h} + \right. \\
&\quad \left. + \sum_{h=1}^{H_{b2}} \left[\frac{C^2_{b2h}}{c_{b2h}} \left(1 - \frac{c_{b2h}}{C_{b2h}}\right) s^2_{\hat{t}b2h} + \frac{C_{b2h}}{c_{b2h}} \sum_{i=1}^{c_{2bh}} \frac{N^2_{b2hi}}{n_{b2hi}} \left(1 - \frac{n_{b2hi}}{N_{b2hi}}\right) s^2_{b2hi} \right] \right\}
\end{aligned} \tag{6}$$

mit s^2_{b1h}, dem Schätzer für die Varianz der Anzahl der Erwerbstätigen pro Wohnung in Schicht $b1h$ und s^2_{b2hi}, dem Varianzschätzer dieser Anzahl in der PSU $b2hi$. $s^2_{\hat{t}b2h}$ schätzt die Varianz der Merkmalssummenschätzer zwischen den PSUs im Bereich 2 der Grundgesamtheit, in dem zweistufig aus der Grundgesamtheit gezogen wird.

Der Design-Effekt eines Stichprobenverfahrens gibt an, um wie viel die Varianz des berechneten Merkmalssummenschätzers beim verwendeten Stichprobendesign größer ist als die des Horvitz-Thompson-Schätzers bei einer uneingeschränkten Zufallsauswahl gleichen Umfangs (vgl. etwa Särndal et al., 1992, S.54). Beim Merkmal Erwerbstätigkeit ergab sich demnach in unseren Simulationen für das komplexe verwendete Stichprobenverfahren ein geschätzter Design-Effekt von mehr als 2. Die Schätzung der Varianz von $\hat{t}$ durch $\hat{V}_{SI}(\hat{t})$ führt also zu einer deutlichen Unterschätzung der tatsächlichen Varianz $V_{SIM}(\hat{t})$.

Der Vergleich der ersten und der dritten Spalte von Tabelle 3 weist ferner nach, dass der Schätzer $\hat{t}$ nach (1) tatsächlich – wie schon in Abschnitt 3.1 behauptet – genauer als

der Horvitz-Thompson-Schätzer ist. Die Differenz $M_{SIM}[\hat{V}^*_{HT}(\hat{t})] - V_{SIM}(\hat{t})$ ist nämlich der Genauigkeitsgewinn, der dadurch erzielt wird, dass bei der Schätzung im zweiten Bereich nach (3) die Größen der Stichproben-PSUs berücksichtigt und an Stelle der Auswahlwahrscheinlichkeiten als Hochrechnungsgewichte herangezogen werden.

Die leichte Überschätzung der tatsächlichen Varianz $V_{SIM}(\hat{t})$ durch $\hat{V}^*_{HT}(\hat{t})$ hat zur Folge, dass in 96.1% der 10000 Simulationen für Österreich das mit diesem Schätzer berechnete approximative Konfidenzintervall zur Sicherheit $1 - \alpha = 0.95$, dessen Grenzen sich jeweils aus $\hat{t} \pm 1.96 \cdot [\hat{V}^*_{HT}(\hat{t})]^{1/2}$ ergeben, den Parameter $t = 3138606$ überdeckt. Somit ist man damit „auf der sicheren Seite". Verwendet man jedoch den „einfacheren" Varianzschätzer $\hat{V}_{SI}(\hat{t})$ statt $\hat{V}^*_{HT}(\hat{t})$ im Konfidenzintervall, so erzielt man lediglich eine Überdeckungshäufigkeit von 74.8% (vgl. Eckmair, 2004, S.109). Vor der Verwendung eines solchen einfachen, die Realität des Stichprobendesigns und der Schätzmethode völlig ignorierenden Varianzschätzers für die Bestimmung eines Konfidenzintervalls ist deshalb abzuraten.

3.3 Nonresponse-Ausgewichtung im AMC-Schätzer für die Merkmalssumme

Der Item-Nonresponse beim Merkmal Erwerbstätigkeit wurde – wie in Abschnitt 2 beschrieben – in unseren Simulationen jeweils einheitlich für alle Bewohner einer Wohnung eingebaut. Treten Antwortausfälle auf, dann gibt es im Wesentlichen zwei Möglichkeiten, diesen in der Schätzphase zu berücksichtigen. Diese bestehen in der Ausgewichtung des Nonresponses und in der Imputation fehlender Werte. Bei der Ausgewichtung wird z.B. im Horvitz-Thompson-Schätzer nach (2) das Designgewicht d_k durch ein den Nonresponse kompensierendes Gewicht w_k ersetzt

$$\hat{t}_{gew} = \sum_s w_k \, y_k \, . \tag{7}$$

Es gilt $w_k \geq d_k$ für alle $k \in s$ (vgl. Lundström and Särndal, 2002, S.66f). In $\hat{t}$ nach (1) werden diese Antwortausfälle innerhalb jeder Schicht des ersten bzw. jeder PSU des zweiten Bereichs durch Erhöhung des Gewichts von respondierenden *Wohnungen* „ausgewichtet". Mit dem verwendeten Nonresponsemechanismus führt dies in den Simulationen zu den in Tabelle 4 dargestellten Ergebnissen hinsichtlich des Mittelwerts bzw. der Varianz von $\hat{t}$.

Ab hier werden wegen der hohen Simulationsdauern für die nachfolgenden Aufgabenstellungen nur mehr die Ergebnisse für das Bundesland Burgenland dargestellt, in dem sich jedoch die AMC-Vorgangsweise für ganz Österreich im Kleinen widerspiegelt.

$\hat{t}$ überschätzt demnach t mit zunehmender Nonresponse-Rate beim verwendeten Nonresponse-Mechanismus immer stärker. Dies zeigt an, dass die Modellannahme für die Ausgewichtung des Nonresponses, dass die Antwortausfälle innerhalb der verwendeten Antwort-Homogenitäts-Gruppen aus Abschnitt 2 unabhängig vom interessierenden Merkmal anfallen, beim gewählten Response-Mechanismus (wie tatsächlich wohl auch in der Realität) nicht zutrifft. Die Antworthäufigkeit von Wohnungen steigt beim verwendeten Nonresponse-Mechanismus mit der Anzahl der Bewohner (und ergo mit der Anzahl der Erwerbstätigen) in der Wohnung (siehe Tabelle 1), da man in Wohnungen mit

Tabelle 4: Verhalten von Mittelwert und Varianz von $\hat{t}$ bei unterschiedlichen wohnungsbezogenen Item-Nonresponse-Raten (Parameter $t = 106237$)

NR-Rate	$M_{SIM}(\hat{t})$	$V_{SIM}(\hat{t})$
0%	106380	$5.28 \cdot 10^6$
5%	106610	$5.48 \cdot 10^6$
11% (Orig.)	106960	$5.79 \cdot 10^6$
25%	107860	$6.50 \cdot 10^6$
40%	109310	$7.45 \cdot 10^6$

mehreren Bewohnern mit größerer Wahrscheinlichkeit jemanden zur Beantwortung des Fragebogens antrifft. Damit sind jedoch zu viele Wohnungen mit mehreren Erwerbspersonen und zu wenig mit wenigen in der Stichprobe. Der Schätzer $\hat{t}$ nach (1) nimmt aber auf die Haushaltsgrößen keine Rücksicht und behandelt alle Wohnungen bei der Hochrechnung auf den Parameter t gleich. Bei einem Item-Nonresponse von 40% hat dies z.B. eine Verzerrung des Anzahlschätzers für das Burgenland um knapp 3000 Personen, also um 2.8%, zur Folge.

Die Auswirkung dieser Verzerrung auf die Überdeckungshäufigkeit von approximativen Konfidenzintervallen zur Sicherheit $1 - \alpha$ ist Gegenstand des nachfolgenden Abschnitts. Dass die Varianz von $\hat{t}$ mit zunehmender Nonresponse-Rate steigt (siehe die rechte Spalte in Tabelle 4) liegt natürlich am mit deren Zunahme abnehmenden tatsächlichen Stichprobenumfang der Erhebung. Mit zunehmender Verzerrung des Schätzers steigt natürlich auch dessen mittlerer quadratischer Fehler im Vergleich zu seiner Varianz. Bei einer Nonresponse-Rate von 40% beträgt dieser z.B. $\text{MSE}_{SIM}(\hat{t}) = 17.67 \cdot 10^6$.

3.4 Direkte Varianzschätzung bei Nonresponse

Um die Auswirkung der Verzerrung von $\hat{t}$ auf die Überdeckungshäufigkeit eines approximativen Konfidenzintervalls zur Sicherheit $1 - \alpha$ zu betrachten, müssen wir uns nochmals mit dem Schätzer für die Varianz des Horvitz-Thompson-Schätzers im AMC beschäftigen. Bei Nonresponse – also praktisch immer – muss man in diesem Schätzer berücksichtigen, dass ein Antwortausfall nur in einer bewohnten Wohnung auftreten kann. Tut man dies nicht, behandelt man also den Ausfall in ausschließlich bewohnten Wohnungen einfach als verkleinerte Stichprobe der bewohnten und unbewohnten Wohnungen, dann wird durch die geschätzten Varianzen s^2_{b1h} und s^2_{b2hi} der Anzahl der Erwerbstätigen pro Wohnung (bewohnt oder unbewohnt) diese Varianz in den Schichten bzw. PSUs klarerweise unterschätzt und somit unterschätzt $\hat{V}^*_{HT}(\hat{t})$ nach (6) bei Nonresponse die Varianz eines Horvitz-Thompson-Schätzers (siehe Tabelle 5).

Zerlegt man jedoch jede Schichtgesamtheit in die zwei Teilgesamtheiten der unbewohnten und der bewohnten Wohnungen, dann ist (8) die erwartungstreue geschätzte Varianz des Horvitz-Thompson-Schätzers im Bereich 1 des Bundeslandes b, $b = 1, \ldots, 9$:

$$\hat{V}_{HT}(\hat{t}_{b1}) = \sum_{h=1}^{H_{b1}} \frac{N^2_{b1h}}{n^*_{b1h}} \left(1 - \frac{n^*_{b1h}}{N_{b1h}}\right) s^2_{b1h}\,. \tag{8}$$

Tabelle 5: Vergleich der unkorrigierten Varianzschätzungen $\hat{V}^*_{HT}(\hat{t})$ und der korrigierten $\hat{V}_{HT}(\hat{t})$ mit den diesbezüglichen Referenzwerten (in Klammern die Überdeckungshäufigkeiten)

NR-Rate	$V_{SIM}(\hat{t})$	$M_{SIM}[\hat{V}^*_{HT}(\hat{t})]$	$M_{SIM}[\hat{V}_{HT}(\hat{t})]$
0%	$5.28 \cdot 10^6$	$5.57 \cdot 10^6$ (95.5%)	$5.57 \cdot 10^6$ (95.5%)
5%	$5.48 \cdot 10^6$	$5.62 \cdot 10^6$ (95.1%)	$5.76 \cdot 10^6$ (96.3%)
11%	$5.79 \cdot 10^6$	$5.68 \cdot 10^6$ (93.8%)	$6.03 \cdot 10^6$ (95.3%)
25%	$6.50 \cdot 10^6$	$5.84 \cdot 10^6$ (88.7%)	$6.72 \cdot 10^6$ (91.2%)
40%	$7.45 \cdot 10^6$	$6.02 \cdot 10^6$ (73.7%)	$7.78 \cdot 10^6$ (78.9%)

Darin ist die nun „korrigierte“ Stichprobenvarianz s^2_{b1h} in der h-ten Schicht von Bereich 1 im Bundesland b gegeben durch

$$s^2_{b1h} = \frac{n_{b1h}}{n_{b1h}-1}\left[\frac{1}{n^{(1)}_{b1h}}\sum_{i=1}^{n^{(1)}_{b1h}}\left(y_{b1hi}-\bar{y}^{(1)}_{b1h}\right)^2 \frac{n^{(1)}_{b1h}+n^{(2)}_{b1h}}{n_{b1h}} + \bar{y}^2_{b1h}\,\frac{n_{b1h}-n^{(1)}_{b1h}-n^{(2)}_{b1h}}{n_{b1h}} + \left(\bar{y}_{b1h}-\bar{y}^{(1)}_{b1h}\right)^2 \frac{n^{(1)}_{b1h}+n^{(2)}_{b1h}}{n_{b1h}}\right]$$

und

$$\bar{y}_{b1h} = \frac{n^{(1)}_{b1h}+n^{(2)}_{b1h}}{n_{b1h}}\,\bar{y}^{(1)}_{b1h}\,.$$

$\bar{y}^{(1)}_{b1h}$ ist der Stichprobenmittelwert der Merkmalssummen in allen respondierenden Wohnungen der Schicht $b1h$ und $\bar{y}_{b1h}$ der Schätzer für den Gesamtmittelwert dieser Schicht. Der Schichtenvarianzschätzer s^2_{b1h} erhält sein Aussehen durch Addition des Erwartungswerts der Varianzen der beiden Gruppen und der Varianz der Erwartungswerte. Wegen $y_{b1hi} = 0$ für alle Wohnungen i unter den unbewohnten Wohnungen ergibt dies dann die oben angeführte Darstellung.

Der Varianzschätzer $\hat{V}_{HT}(\hat{t}_{b2})$ nach (9) ist die geschätzte Varianz eines Horvitz-Thompson-Schätzers in Bereich 2 des Bundeslandes b, $b = 1, \ldots, 7$, bei zweistufiger Auswahl mit uneingeschränkten Zufallsauswahlen auf beiden Stufen und Nonresponse nur in den bewohnten Wohnungen

$$\hat{V}_{HT}(\hat{t}_{b2}) = \begin{cases} \sum_{h=1}^{H_{b2}} \left[\frac{C^2_{b2h}}{c_{b2h}}\left(1-\frac{c_{b2h}}{C_{b2h}}\right) s^2_{\hat{t}b2h} + \frac{C_{b2h}}{c_{b2h}} \sum_{i=1}^{c_{2bh}} \frac{N^2_{b2hi}}{n^*_{b2hi}}\left(1-\frac{n^*_{b2hi}}{N_{b2hi}}\right) s^2_{b2hi}\right], & \text{für } b = 1, \ldots, 7 \\ 0\,, & \text{für } b = 8, 9 \end{cases} \tag{9}$$

mit

$$s^2_{\hat{t}b2h} = \frac{1}{c_{b2h}-1}\sum_{i=1}^{c_{b2h}}\left(N_{b2hi}\,\bar{y}_{b2hi} - \frac{1}{c_{b2h}}\sum_{k=1}^{c_{b2h}} N_{b2hk}\,\bar{y}_{b2hk}\right)^2$$

und

$$\bar{y}_{b2hi} = \frac{n^{(1)}_{b2hi} + n^{(2)}_{b2hi}}{n_{b2hi}}\ \bar{y}^{(1)}_{b2hi}\,.$$

$\bar{y}^{(1)}_{b2hi}$ ist der Stichprobenmittelwert der Merkmalssummen in allen respondierenden Wohnungen der i-ten PSU in der Schicht $b2h$ und $\bar{y}_{b2hi}$ der Schätzer für den Gesamtmittelwert dieser PSU. $s^2_{\hat{t}b2h}$ in (9) ist die Stichprobenvarianz der Merkmalssummenschätzer bei Nonresponse zwischen den Stichproben-PSUs innerhalb der Schicht $b2h$. Sie berücksichtigt den Nonresponse in $\bar{y}_{b2hi}$. Innerhalb der PSUs wird die Varianz eines Horvitz-Thompson-Schätzers für z.B. die Anzahl der Erwerbstätigen unter Berücksichtigung des angesprochenen Nonresponse-Problems erwartungstreu geschätzt durch

$$s^2_{b2hi} = \frac{n_{b2hi}}{n_{b2hi}-1}\left[\frac{1}{n^{(1)}_{b2hi}}\sum_{j=1}^{n^{(1)}_{b2hi}}\left(y_{b2hij} - \bar{y}^{(1)}_{b2hi}\right)^2\ \frac{n^{(1)}_{b2hi} + n^{(2)}_{b2hi}}{n_{b2hi}} + \right.$$
$$\left. + \bar{y}^2_{b2hi}\ \frac{n_{b2hi} - n^{(1)}_{b2hi} - n^{(2)}_{b2hi}}{n_{b2hi}} + \left(\bar{y}_{b2hi} - \bar{y}^{(1)}_{b2hi}\right)^2\ \frac{n^{(1)}_{b2hi} + n^{(2)}_{b2hi}}{n_{b2hi}}\right].$$

Der Schätzer $\hat{V}_{HT}(\hat{t}_b)$ für die Varianz von $\hat{t}_b$ ergibt sich als Summe der beiden Bereichsvarianzschätzer des b-ten Bundeslandes. Der Gesamtvarianzschätzer $\hat{V}_{HT}(\hat{t})$ errechnet sich schließlich aus der Summe der neun Bundesländer-Varianzschätzer $\hat{V}_{HT}(\hat{t}_b)$. Bei vollem Response gilt $\hat{V}_{HT}(\hat{t}) = \hat{V}^*_{HT}(\hat{t})$.

In Tabelle 5 werden die Mittelwerte der Varianzschätzungen durch den unkorrigierten Varianzschätzers $\hat{V}^*_{HT}(\hat{t}_b)$ eines Horvitz-Thompson-Schätzers und den korrigierten Schätzer $\hat{V}_{HT}(\hat{t}_b)$ nach (8) und (9) mit dem Referenzwert $V_{SIM}(\hat{t})$ für die tatsächliche Varianz des verwendeten Schätzers $\hat{t}$ bei Nonresponse verglichen.

In der dritten Spalte sieht man für den unkorrigierten Varianzschätzer $\hat{V}^*_{HT}(\hat{t})$ nach (6) die mit steigender Nonresponse-Rate zunehmende Unterschätzung der Varianz eines Horvitz-Thompson-Schätzers. Schon bei 11% Nonresponse liegt der Mittelwert der Schätzungen auch unter der tatsächlichen Varianz $V_{SIM}(\hat{t})$ des genaueren Schätzers nach (1). Die vierte Spalte zeigt jedoch wieder das schon aus Tabelle 3 bekannte Bild einer leichten Überschätzung der Varianz unseres Horvitz-Thompson-*nahen*-Schätzers $\hat{t}$ durch die Varianz des Horvitz-Thompson-Schätzers. Die Verwendung von $\hat{V}_{HT}(\hat{t})$ an Stelle von $\hat{V}^*_{HT}(\hat{t})$ erhöht selbstverständlich die Überdeckungshäufigkeiten der approximativen 95%-Konfidenzintervalle, die jedoch wegen der in 3.2 beschriebenen zunehmenden Verzerrung von $\hat{t}$ mit wachsender Nonresponse-Rate dennoch immer deutlicher unter 95% liegen.

3.5 Imputation statt Ausgewichtung

Bei der Imputation von Daten werden fehlende Werte y_k durch Schätzungen $\hat{y}_k$ derselben ersetzt, wobei es unterschiedlichste Imputationstechniken zur Bestimmung des Ersatzwertes $\hat{y}_k$ gibt. Durch die Imputation fehlender Daten entsteht in jedem Fall ein vollständiger Datensatz für die interessierende Variable y. Dieser besteht dann aus den Werten z_k, für die gilt

$$z_k = \begin{cases} y_k\,, & \text{wenn } k \text{ ein Respondent ist,} \\ \hat{y}_k\,, & \text{wenn } k \text{ ein Nonrespondent ist.} \end{cases}$$

Der Horvitz-Thompson-Schätzer $\hat{t}_{HT}$ nach (2) für die Merkmalssumme von y wird zu

$$\hat{t}_{imp} = \sum_{s} d_k \, z_k \tag{10}$$

In diesem Schätzer wird also bei Imputation der Wert $\hat{y}_k$ wie der richtige Wert y_k behandelt. Auch für den AMC-Schätzer $\hat{t}$ nach (1), (3) und (4) wird deshalb dieselbe Vorgangsweise wie bei vollem Response gewählt.

Bei der hier beschriebenen Simulationsstudie wurden sechs mehr oder weniger verschiedene Imputationstechniken angewendet und die Ergebnisse dieser Anwendungen miteinander verglichen (vgl. für eine genauere Beschreibung mit Laaksonen et al., 2004, S.12ff). Fünf der sechs verwendeten Methoden bedienten sich zur Bestimmung der Ersatzwerte $\hat{y}_k$ für fehlende Werte y_k einer logistischen Regression. Bei diesen Methoden werden unter Verwendung von Hilfsvariablen innerhalb vorgegebener Teilgesamtheiten die Wahrscheinlichkeiten für die beiden Merkmalsausprägungen eines binären Merkmals bei einer Gruppe von Erhebungseinheiten geschätzt. Bei den ersten vier Methoden betraf dies die Schätzung der Wahrscheinlichkeiten für das Merkmal Erwerbstätigkeit (ja/nein) bei den Nichtrespondierenden auf Basis der Antwortenden in jedem Bundesland durch Verwendung der Hilfsvariablen Alter, Geschlecht und Haushaltsgröße. Den Nichtrespondierenden wurde schließlich auf Basis dieser Wahrscheinlichkeiten mit Hilfe einer auf dem Intervall $[0, 1]$ gleichverteilten Zufallszahl eine der beiden Ausprägungen von y zugewiesen (vgl. Rubin, 1987, S.168ff). Mit dieser modellbasierten Methode wurden in der Studie durch Wahl von vier verschiedenen Anzahlen m an imputierten Ersatzwerten ($m = 1, 5, 10, 15$) eine Single Imputationstechnik (LR1) und drei Methoden der Multiplen Imputation (LR5, LR10, LR15) simuliert.

Die fünfte der in dieser Simulationsstudie verwendeten Imputationstechniken, die sich ebenfalls bundesländerweise einer logistischen Regression bedient hat, unterschied sich von den vier anderen dadurch, dass dabei auf Basis aller Erhebungseinheiten durch die Hilfsvariablen Alter, Geschlecht und Haushaltsgröße deren Wahrscheinlichkeiten für die Ausprägungen des binären Merkmals Responseverhalten geschätzt wurden. Mit diesen Schätzungen wurden die Erhebungseinheiten in 10 gleich große Gruppen aufgeteilt, wobei sich in der ersten Gruppe die 10% der Personen mit den niedrigsten geschätzten Responsewahrscheinlichkeiten, in der zweiten Gruppe diejenigen 10% mit den nächsthöheren Responsewahrscheinlichkeiten usf. befanden. In jeder dieser so gebildeten Imputationsgruppen wurde für jeden Nichtrespondenten zufällig ein Donor aus den Antwortenden ausgewählt und dessen Merkmalsausprägung beim Merkmal y für den beim Nichtrespondenten fehlenden Wert imputiert. Diese Mischung aus modell- und donorbasierter Single Imputationstechnik wird Hot-Deck-Propensity-Score-Matching genannt (HDPS). Die Begründung für die bundesländerweise Anwendung dieser Methoden liegt in dem Umstand, dass man für vernünftige Schätzungen der jeweiligen interessierenden Wahrscheinlichkeiten eine genügend große Anzahl an Erhebungseinheiten zur Verfügung haben muss und in der notwendigen Beschränkung der Simulationszeiten.

Für die sechste verwendete Imputationstechnik wurden in jedem Bundesland innerhalb der Schichten, die aus regional zusammenhängenden Erhebungseinheiten bestanden, für jeden fehlenden Wert der Wert eines zufällig aus den Antwortenden dieser Schicht ausgewählten Donors imputiert. Diese donorbasierte Methode wird als Hot Deck within cells bezeichnet (HD).

Tabelle 6: Vergleich von Mittelwerten, Varianzen und mittleren quadratischen Fehlern des Merkmalssummenschätzers $\hat{t}$ bei verschiedenen Imputationstechniken

	LR1	LR5	LR10	LR15	HD	HDPS
$M_{SIM}(\hat{t})$	105560	105570	105570	105570	106680	106270
$V_{SIM}(\hat{t})$	$8.07 \cdot 10^6$	$6.95 \cdot 10^6$	$6.83 \cdot 10^6$	$6.79 \cdot 10^6$	$6.62 \cdot 10^6$	$6.64 \cdot 10^6$
$\text{MSE}_{SIM}(\hat{t})$	$8.58 \cdot 10^6$	$7.44 \cdot 10^6$	$7.32 \cdot 10^6$	$7.28 \cdot 10^6$	$6.78 \cdot 10^6$	$6.64 \cdot 10^6$

Tabelle 6 enthält die Simulationsresultate bei einer Nonresponse-Rate von 40% im Bundesland Burgenland. Zum Vergleich dienen folgende Parameter und Statistiken: $t = 106273$, $M_{SIM}(\hat{t}) = 109310$, $V_{SIM}(\hat{t}) = 7.45 \cdot 10^6$ und $\text{MSE}_{SIM}(\hat{t}) = 17.67 \cdot 10^6$. Beim angenommenen Nonresponse-Mechanismus (siehe Abschnitt 2) gelingt es ausnahmslos allen verwendeten Imputationstechniken, die Verzerrung des Merkmalssummenschätzers im Vergleich zu seiner Ausgewichtung bei 40% Nonresponse deutlich zu verringern oder gänzlich zu vermeiden. Dies beruht auf der geeigneten Wahl von mit dem Nonrespoonse-Verhalten zusammenhängenden Hilfsinformationen im Imputationsprozess. Innerhalb der vier Methoden, die sich nur hinsichtlich der Imputationsanzahlen m unterscheiden, nimmt die Streuung des Schätzers $\hat{t}_m$ der Merkmalssumme, der der Mittelwert der Schätzer $\hat{t}$ in den m Imputationsdatensätzen ist, mit zunehmender Anzahl m klarerweise ab. Der Gewinn an Genauigkeit ist jedoch für $m = 10$ und 15 im Vergleich zu $m = 5$ sehr gering, so dass sich hier durchaus die Frage stellt, ob sich der zusätzliche Rechenaufwand lohnt. Deutlich hingegen fällt der Unterschied zwischen LR1 und LR5 aus, also zwischen Single und Multipler Imputation.

Auch an den Varianzen $V_{SIM}(\hat{t})$ erkennt man die positive Auswirkung der Verwendung von über die Nichtantwortenden vorliegenden Informationen als Hilfsvariablen im Vergleich zur kompletten Tilgung dieses Personenkreises aus der Erhebung (vgl. Abschnitt 3.3). Die Varianzen sind natürlich allesamt höher als jene bei vollem Response. Sie nähern sich Letzterer jedoch je nach erhebungsspezifischer Eignung der verwendeten Imputationstechnik mehr oder weniger an.

Sowohl hinsichtlich der Verzerrung als auch hinsichtlich des mittleren quadratischen Fehlers stellte sich HDPS als beste Methode heraus. Die Miteinbeziehung der Teilnahmebereitschaft in den Imputationsprozess wirkt sich also positiv auf die Schätzergenauigkeit aus und HDPS überbietet hinsichtlich der Genauigkeit die multiplen Imputationen, die diese Variable nicht miteinbezogen. Beinahe so gut wie HDPS funktioniert das reine Hot Deck Verfahren HD, wofür alleine die im Vergleich zu den anderen fünf Verfahren engere regionale Definition der Imputationszellen verantwortlich zeichnet. Das Merkmal Erwerbstätigkeit hängt in unserer Pseudo-Grundgesamtheit wie auch in der Realität mit den Regionen des Landes zusammen.

3.6 Varianzschätzung bei imputierten Daten

Bei Verwendung einer Single Imputationstechnik wird durch den Varianzschätzer (6), der bei vollem Response geeignete Schätzungen der Schätzervarianz liefert, die Varianz des „Imputationsschätzers“ $\hat{t}$ klarerweise unterschätzt, da die Unsicherheit der Imputationen

Tabelle 7: Vergleich von Varianzschätzungen bei Single Imputation mit den diesbezüglichen Referenzwerten

	LR1	HD	HDPS
$V_{SIM}(\hat{t})$	$8.07 \cdot 10^6$	$6.62 \cdot 10^6$	$6.64 \cdot 10^6$
$M_{SIM}[\hat{V}^*_{HT}(\hat{t})]$	$5.09 \cdot 10^6$	$5.37 \cdot 10^6$	$5.35 \cdot 10^6$
Überdeckung (in %)	86.5	91.9	92.1
Zeitaufwand (in sec)	3.8	4.5	6.0
$M_{SIM}[\hat{V}_{B25}(\hat{t})]$	$8.09 \cdot 10^6$	$6.35 \cdot 10^6$	$6.35 \cdot 10^6$
Überdeckung (in %)	92.8	93.3	93.2
Zeitaufwand (in sec)	97.5	110.7	150.3
$M_{SIM}[\hat{V}_{B50}(\hat{t})]$	$\mathbf{8.06 \cdot 10^6}$	$\mathbf{6.41 \cdot 10^6}$	$\mathbf{6.33 \cdot 10^6}$
Überdeckung (in %)	**93.4**	**93.8**	**93.7**
Zeitaufwand (in sec)	195.0	221.4	300.5

unberücksichtigt bleibt (siehe $M_{SIM}[\hat{V}^*_{HT}(\hat{t})]$ in Tabelle 7). Eine Alternative dazu bietet die Verwendung von Resampling-Verfahren. In dieser Studie wurde das Bootstrapverfahren nach Shao and Sitter (1996) bundesländerweise angewendet. Eine korrekte schichten- bzw. schichten- und PSU-weise Anwendung war aus Gründen der Beschränkung der Simulationszeiten nicht durchführbar. Bei diesem Verfahren wird aus einer Stichprobe vom Umfang n eine Anzahl r an ebenso großen unabhängigen Bootstrapstichproben mit Zurücklegen gezogen. In diesen Bootstrapstichproben werden Daten, die schon in der ursprünglichen Stichprobe imputiert waren, nach demselben Verfahren wie ursprünglich aus der Bootstrapstichprobe neu imputiert. Auf diese Weise wird beim Resamplen das Imputationsverfahren mitberücksichtigt. Die Berechnung des Schätzers $\hat{t}$ nach (1) in jeder der Bootstrapstichproben ermöglicht die Verwendung der Monte-Carlo Approximation für die Varianz des Schätzers (vgl. Shao and Sitter, 1996, S.1278ff). In den Simulationen wurden $r = 25$ bzw. $r = 50$ Bootstrapstichproben (B25 bzw. B50) für jede der simulierten Stichproben gezogen. In Tabelle 7 sind die Mittelwerte der Varianz(unter)schätzung $\hat{V}^*_{HT}(\hat{t})$ und der Bootstrapvarianzschätzungen für eine Nonresponse-Rate von 40% im Bundesland Burgenland angegeben. Diese Schätzer werden verglichen mit der tatsächlichen Varianz $V_{SIM}(\hat{t})$ von $\hat{t}$.

Die Bootstrapmethode von Shao and Sitter (1996) zeigt sowohl eine starke Übereinstimmung des Mittelwerts mit den Referenzwerten als auch eine dem angepeilten 95%-Niveau nahe kommende Überdeckungshäufigkeit der mit den Bootstrap-Varianzschätzern und den dazugehörenden Schätzern $\hat{t}$ (siehe Tabelle 6) errechneten approximativen Konfidenzintervalle, was wegen der aus Zeitgründen ohne Berücksichtigung der PSUs im Bereich 2 schichtweise erzeugten Bootstrapstichproben ein wenig überrascht. Offenbar ist der Genauigkeitsverlust durch die zweistufige Auswahl der Wohnungen innerhalb der Schichten dieses Bereichs im Vergleich zu einer uneingeschränkten Zufallsauswahl gering. Zu erkennen ist ferner, dass die Verdopplung der Anzahl der Bootstrapstichproben von 25 auf 50 eine nur mehr geringe Ergebnisverbesserung hervorgebracht hat. Diese – wenn auch geringe – Verbesserung ist dadurch zu erklären, dass die Varianz des Varianzschätzers $\hat{V}_{B50}(\hat{t})$ nur halb so groß ist wie jene von $\hat{V}_{B25}(\hat{t})$. Gleichzeitig verdoppel-

Tabelle 8: Vergleich der Varianzschätzer bei Multipler Imputation mit Referenzwerten

	LR5	LR10	LR15
$V_{SIM}(\hat{t})$	$6.95 \cdot 10^6$	$6.83 \cdot 10^6$	$6.79 \cdot 10^6$
$M_{SIM}[\hat{V}_{MI}(\hat{t})]$	$6.74 \cdot 10^6$	$6.62 \cdot 10^6$	$6.56 \cdot 10^6$
Überdeckung (in %)	93.4	93.5	93.6

te sich die durchschnittliche Rechenzeit eines Schätzdurchganges für die interessierende Größe im betrachteten Bundesland von ca. 2.5 auf 5 Minuten.

Für die drei angewendeten Verfahren der Multiplen Imputation, die sich lediglich in der Anzahl m der Imputationen pro fehlendem Wert unterschieden haben, wurde für die Varianzschätzung $\hat{V}_{MI}(\hat{t}_m)$ die Varianz in einen Teil innerhalb, der die Stichprobenstreuung, und einen zwischen den m vollständigen Datensätzen, der die Imputationsstreuung beschreibt, zerlegt und jeder dieser Teile geschätzt. Es ergibt sich daraus als Varianzschätzer

$$\hat{V}_{MI}(\hat{t}_m) = \frac{1}{m} \sum_m \hat{V}^*_{HT}(\hat{t}) + \left(1 + \frac{1}{m}\right) \hat{V}_m(\hat{t}) \tag{11}$$

(vgl. hierzu Little and Rubin, 2002, S.85ff). Der Varianzschätzer $\hat{V}_{MI}(\hat{t}_m)$ ergibt sich nach (11) aus dem Mittelwert der Varianzschätzer $\hat{V}^*_{HT}(\hat{t})$ in jedem der m Datensätze und der geschätzten Varianz $\hat{V}_m(\hat{t})$ der m verschiedenen Merkmalssummenschätzer. Diese Möglichkeit zur Bestimmung eines Varianzschätzers ist der große Vorteil der Multiplen gegenüber der Single Imputation. Tabelle 8 enthält die diesbezüglichen Ergebnisse wiederum im Vergleich mit den Referenzwerten.

Die deutliche Unterschätzung der Varianz $V_{SIM}(\hat{t})$ des imputierten Schätzers durch $\hat{V}^*_{HT}(\hat{t})$ (siehe Tabelle 7) wird durch die Miteinbeziehung der Varianz zwischen den m imputierten Datensätzen in (11) in den Simulationen nicht ganz kompensiert. Die Varianzschätzungen werden mit zunehmender Anzahl m an Imputationen gemessen an den mit ihnen erzielten Überdeckungen besser. Diese Verbesserung ist jedoch geringfügig, während sich der Zeitaufwand für einen Schätzvorgang von durchschnittlich 22.5 sec (für LR5) auf 44.9 sec (für LR15) erhöht. Dies unterstreicht einmal mehr die diesbezügliche Aussage Rubins, dass man schon bei sehr geringen Anzahlen m gute Überdeckungshäufigkeiten erzielen kann (vgl. Rubin, 1987, S.114f).

4 Zusammenfassung

Das Auftreten von Nonresponse kann die Qualität von Stichprobenergebnissen in zweierlei Hinsicht beeinträchtigen. Zum einen sind sie trotz (z.B. regionaler) Ausgewichtung verzerrt, sobald sich die Teilnahmebereitschaft nicht nur durch die für die Ausgewichtung verwendeten Merkmale erklären lässt. Zum anderen erhöht der niedrigere Stichprobenumfang jedenfalls die Varianz der Stichprobenergebnisse. Liegen von den Nichtrespondierenden jedoch Informationen zu Hilfsvariablen vor, dann lässt sich durch Verwendung dieser Informationen für die Imputation fehlender Daten die Qualität der Stichprobenergebnisse erhöhen.

Für die vorgestellte Simulationsstudie wurde der Nonresponse auf Basis eines Mechanismus in die Stichproben eingebaut, der sich an diesbezüglichen Informationen aus dem deutschen Mikrozensus orientiert hat. In diesen Simulationen stellte sich beim Vergleich der Tauglichkeit von sechs unterschiedlichen Imputationstechniken beim vorliegenden Nonresponsemechanismus sowohl hinsichtlich der Verzerrung eines Anzahlschätzers als auch in Bezug auf seine Varianz die Methode des „Hot-Deck-Propensity-Score-Matching" als beste heraus. Sie bediente sich bundesländerweise berechneter logistischen Regressionen, durch die auf Basis aller Erhebungseinheiten mit den Hilfsvariablen Alter, Geschlecht und Haushaltsgröße die Wahrscheinlichkeiten für die Ausprägungen des dichotomen Merkmals Responseverhalten geschätzt wurden. Nach diesen Wahrscheinlichkeitsschätzungen wurden die Erhebungseinheiten in 10 gleich große Gruppen aufgeteilt. In jeder dieser Gruppen wurde für jeden Nichtrespondenten zufällig ein Donor aus den Antwortenden ausgewählt und dessen Merkmalsausprägung beim Merkmal y für den beim Nichtrespondenten fehlenden Wert imputiert.

Unter den vier verglichenen Methoden, die sich nur durch die Anzahlen an Imputationen unterscheiden, ergab sich auf Basis des Vergleichs der damit erzielten Schätzervarianzen eine deutliche Präferenz für die Multiplen Techniken, wobei der Genauigkeitsgewinn mit zunehmender Anzahl an Imputationen abnimmt und schon bei einer Erhöhung der Imputationsanzahlen von 5 auf 10 bzw. 15 im Vergleich zum zusätzlichen Zeitaufwand vernachlässigbar erscheint (siehe Tabelle 6).

Neben der Betrachtung der Varianz der verschiedenen Imputations-Schätzer empfiehlt es sich, auch die z.B. für Intervallschätzungen nötigen Schätzungen dieser Varianz zu betrachten. Am unproblematischsten in dieser Hinsicht erweisen sich natürlich die Multiplen Imputationstechniken mit ihrer impliziten Möglichkeit der Varianzschätzung. Bei den drei Single Imputationsmethoden zeigt sich, dass die direkte Varianzschätzung, also die Behandlung der imputierten Daten als „wahre Werte", in allen Fällen zu einer deutlichen Unterschätzung der Schätzervarianz führt, wodurch der große Zeitvorteil dieser Vorgangsweise irrelevant wird. Die Varianzschätzung mittels eines Boostrapverfahrens nach Shao and Sitter (1996), das die durch die Imputationen zusätzlich erzeugte Ungenauigkeit mitberücksichtigt, funktioniert dagegen gut, ist jedoch sehr zeitaufwändig (siehe Tabellen 7 und 8).

Anerkennung

Ich möchte mich bei meinen Kolleginnen am IFAS Doris Eckmair und Helga Wagner für die erfolgreiche Zusammenarbeit beim EU-Projekt DACSEIS bedanken. Ferner möchte ich besonders hervorstreichen, dass der Projektkoordinator Ralf Münnich immer mit Rat und Tat zur Seite gestanden ist und dass für die Simulationsresultate in den Abschnitten 3.5 und 3.6 R-Prozeduren verwendet wurden, die an der Abteilung Statistik, Ökonometrie und Unternehmensforschung der Eberhard Karls Universität in Tübingen entwickelt und zusammen mit dem gesamten Rahmen für die Simulationen für das Projekt bereitgestellt wurden. Weitere Ergebnisse des DACSEIS-Projekts sind unter http://laplace.wiwi.uni-tuebingen.de/dacseis/start.html bereit gestellt.

Literatur

Eckmair, D. (2004). *Simulation eines Stichprobendesigns aus generierten Pseudogrundgesamtheiten zur Qualitätsbeurteilung unterschiedlicher Varianzschätzer.* Unpublished doctoral dissertation, Institut für Angewandte Statistik der Johannes Kepler Universität Linz.

Haslinger, A. (1996). Stichprobenplan des Mikrozensus ab 1994. *Statistische Nachrichten*, 312-324.

Kytir, J., and Stadler, B. (2004). Die kontinuierliche Arbeitskräfteerhebung im Rahmen des neuen Mikrozensus. *Statistische Nachrichten*, 511-518.

Laaksonen, S., Rässler, S., and Skinner, C. (2004). *Imputation and Non-Response.* Deliverable 11.2 of the DACSEIS project.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken: Wiley.

Lundström, S., and Särndal, C.-E. (2002). *Estimation in the Presence of Nonresponse* (2nd ed.). Statistics Sweden.

Münnich, R. (2003). *Data Quality in Complex Surveys.* Deliverable 1.1 of the DACSEIS project.

Münnich, R., and Wiegert, R. (2001). *The DACSEIS Project.* DACSEIS Research Paper Series No. 1. (http://www.dacseis.de/research.html)

Quatember, A. (2002). *Analysis of National Surveys.* Deliverables 2.1 and 2.2 of the DACSEIS project.

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

Shao, J., and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, *91*, 1278-1288.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Wagner, H. (2003). *Monte-Carlo Simulation Study of European Surveys.* Deliverables 3.1 and 3.2 of the DACSEIS project.

Wagner, H., and Eckmair, D. (2004). *Simulation Studies of the Austrian Microcensus* (Tech. Rep. No. 5). Institut für Angewandte Statistik der Johannes Kepler Universität Linz.

Adresse des Autors:

Andreas Quatember
IFAS – Institut für Angewandte Statistik
Johannes Kepler Universität Linz
Altenbergerstraße 69
A-4040 Linz
Österreich

E-mail: andreas.quatember@jku.at
Internet: www.ifas.jku.at