Imputation fehlender Werte im Labour Force Survey

Thomas Burg
Österreichisches Statistisches Zentralamt, Wien

Zusammenfassung: Im Zuge eines Sonderprogramms zum Mikrozensus wurde im 1. Quartal 1995 erstmals eine Arbeitskräfteerhebung (LFS - Labour Force Survey) durchgeführt.
Nachdem dieses Sonderprogramm auf freiwilliger Basis erhoben wurde, kam es einerseits zu Verweigerungen bezüglich des kompletten Sonderprogramms, sowie andererseits zu fehlenden Werten bei einzelnen Merkmalen.
Um komplette Datensätze zu erhalten, wurde eine Methode entwickelt, die mittels einer auf soziodemographischen Merkmalen basierenden Distanzfunktion den "ähnlichsten" Spender (Donor) zu einem vorgegebenen Datensatz aufsucht und sämtliche fehlende Merkmale gleichzeitig imputiert.

Abstract: Together with the Austrian Microcensus, which contains a voluntery part the first Labour Force Survey was carried out.
Since this part of the survey was not mandatoty there turned out to be not only refusals of the whole LFS (unit non-response) but also mising values at special variables (item non-response).
To come to complete records a method which uses a distance measure based on sociodemographic items was developed. This method searches for the most similar donor for a record and imputes all the missing values simultanously.

1 Ausgangslage und Problemstellung

1.1 Durchführung des 1. Labour Force Survey

Im Zuge des Sonderprogramms zum Mikrozensus wurde im 1. Quartal 1995 das erste mal die Arbeitskräfteerhebung (im Englischen Labour Force Survey - im weiteren sei die Erhebung mit LFS bezeichnet) durchgeführt.

Der Respondent wurde vom gleichen Interviewer zum LFS befragt, der ihn auch über das Gundprogramm befragt hatte. Zum Unterschied zum Grundprogramm beruhte das Sonderprogramm jedoch auf freiwilliger Basis. Die Stichprobe des Mikrozensus umfaßt ca. 60.000 Personen.

Dadurch, daß die Personen zu verschieden Themenkreisen bezüglich ihrer momentanen Arbeitssituation befragt wurden, zerfiel das Interview gewissermaßen in etliche kleine Teilstücke. Somit ergab sich eine relativ komplexe Fragebogenstruktur und detaillierte Richtlinien für den Interviewer waren notwendig.

1.2 Antwortausfälle

Durch die Freiwilligkeit des LFS lag es in der Natur der Sache, daß es zu Antwortausfällen kam. Wie im klassischen Sinne üblich, muß man zwischen zwei Arten des Antwortausfall unterscheiden:

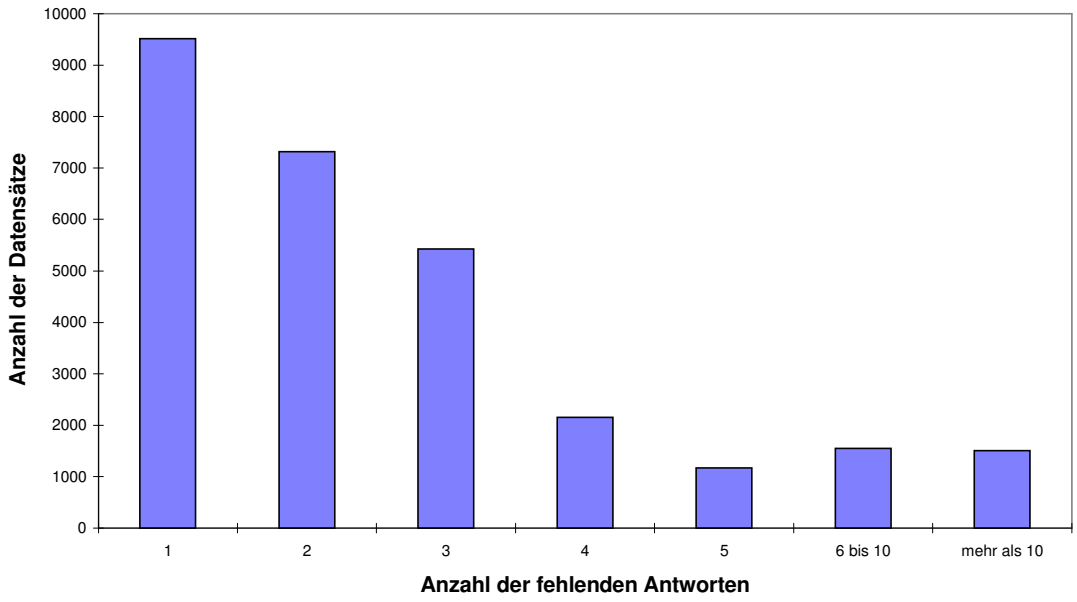
1. Der Respondent verweigert die Beantwortung des LFS komplett ("Unit Non-Response").
2. Es fehlen nur die Antworten zu einer oder mehrerer Fragen, die eigentlich beantwortet hätte werden müssen ("Item Non Response").

Hier sind einige typische Gründe für das Auftreten von Nonresponse angeführt:

- Mangelnde Kooperationsbereitschaft des Respondenten
- Zu geringes Engagement des Interviewers
- Komplexität des Fragebogens
- Sensibilität der Fragestellung
- Respondent weiß die Antwort nicht

Die Non-Response-Rate bei Unit-Nonresponse im LFS betrug ca. 10%. Für das Verhalten bei Item-Non-Response gibt Abbildung 1 Aufschluß.

Abbildung 1: Verteilung der fehlenden Antworten pro Datensatz



Insgesamt waren ca. 53% der Datensätze mit Item-Nonresponse behaftet. Zum größten Teil sind ein bis drei Fragen unbeantwortet geblieben. Sind mehr als 10 Fragen ohne Wert so kann man den Datensatz wohl als obsolet bezeichnen und ihn als Totalausfall (d.h. als Unit-Nonresponse) rechnen.

1.3 Warum müssen Daten imputiert werden?

Warum entstand nun die Notwendigkeit, die Antwortausfälle (bei "Item Non-Response") durch Imputation zu korrigieren? Die Hauptargumentation amtsintern dafür war, daß EUROSTAT auf die Lieferung kompletter Datensätze seitens des ÖSTAT bestand. Außerdem können durch jede Art der Non-Response auch statistische Verzerrungen entstehen, die durch Imputationen ausgeglichen werden.

Auch ist es eventuell möglich (abhängig von der Imputationsmethode), die Non-Respondenten durch Analyse der beim Imputationsvorgang entstehenden Outputstatistiken näher zu charakterisieren.

2 Methode der Imputation

2.1 Das Imputationsprofil

Der LFS besteht aus insgesamt 64 Fragen, die - abhängig von den Antworten des Respondenten - beantwortet werden müssen. In den meisten Fällen mußten nicht alle 64 Antworten gegeben werden. Jedenfalls kann man die 64 Fragen als Bitkette der Länge 64 auffassen, die dergestalt aufgebaut ist, daß an der Stelle, wo sich ein fehlender Wert befindet, aber eine Antwort erwartet wird, das Bit gesetzt wird, ansonsten aber nicht. Daraus ergibt sich, daß die Gesamtzahl der möglichen Imputationsprofile mit 2^{64} sehr hoch ist. Bei der Untersuchung der Imputationsprofile im LFS fiel auf, daß die gesetzten Bits zumeist geklumpt waren, d.h. meist hintereinander auftraten, was darauf hindeutet, daß die Fragen, die nicht beantwortet wurden in Gruppen auftraten, die sich mit einem bestimmten Problemkomplex befassen.
Abbildung 2 zeigt typische Beispiele für die Gestalt von Imputationsprofilen im LFS.

Abbildung 2: Beispiele für die Gestalt von Imputationsprofilen im LFS

```
0000100000000000110001100000000000000000000000000000000000000000
0010100000000011110000000100000000000011000000000111111111100000
0000100000000000110000000100000000000111000001000011111111100000
0000101000000000110000000100000000000011000000000111111111100000
0000100000000011110001100000000000000000000000000000000000000000
0000100001000000111111100000000000000000000000000000000000000000
```

2.2 Das Antwortprofil

Durch die gültigen Antworten des LFS und durch einige Merkmale des Grundprogramms wurde festgelegt, welche Fragen des LFS beantwortet werden mußten. Ähnlich wie beim Imputationsprofil wird nun eine Bitkette der Länge 64 gebildet. Überall dort wo eine Antwort erwartet wird, ist das Bit gesetzt. Damit ergibt sich, daß ein Bit im Imputationsprofils eines Datensatzes nur dann gesetzt sein kann, wenn das entsprechende Antwortbit gesetzt ist.

2.3 Die Distanzfunktion

Um Merkmale imputieren zu können, muß man wissen woher man die Werte nimmt, die man letztendlich imputiert. Dazu ist es notwendig, zu wissen, wie "nahe" ein bestimmter Datensatz einem anderen ist. Mathematisch gesehen benötigt man also eine Distanzfunktion. Einerseits ist zu überlegen, welche Merkmale in die Distanzfunktion eingehen sollen, andererseits muß man sich über die algebraische Gestalt der Distanzfunktion klar werden. Zum einen ist zu sagen, daß zwei Datensätze dann als nahe (oder "ähnlich") anzusehen sind, wenn die dahinterstehenden Personen in gewissen soziodemographischen Merkmalen übereinstimmen (d. h. Alter, Geschlecht, Familienstand usw.). Auf der anderen Seite ist der Unterschied in qualitativen bzw. quantitativen Merkmalen sinnvoll zu evaluieren. Nehmen wir an in eine Distanzfunktion gehen insgesamt $I = I_q + I_n$ soziodemographische Merkmale ein, wobei I_q die Anzahl der qualitativen Merkmale darstellt und I_n die Anzahl der quanitativen. Eine allgemeine Form einer solchen Distanzfunktion zwischen zwei Datensätzen *R* und *S* hat dann folgende Gestalt:

$$D(R,S) = \sum_{i=1}^{I_q} w_{qi} D_{qi}(R,S) + \sum_{i=1}^{I_n} w_{ni} D_{ni}(R,S)$$

Für den LFS sei nun $D_{qi}(R,S)$ gleich 1, wenn der Eintrag in der *i*-ten qualitativen Variable von *R* ungleich dem Eintrag in der *i*-ten qualitativen Variable von *S* ist und 0 ansonsten.

Für quantitative Variable ist die Funktion etwas komplizierter. Es wird ein Parameter maxdiff definiert. Wenn für das *i*-te numerische Merkmal $|R_{ni} - S_{ni}| > maxdiff$ gilt, so folgt $D_{ni}(R,S)=1$. Ansonsten gilt $D_{ni}(R,S) = 1 - \left(1 - \frac{|R_{ni} - S_{ni}|}{maxdiff}\right)$. Die w_{qi} bzw w_{ni} stellen Gewichte dar mit denen die Wichtigkeit der einzelnen Merkmale gesteuert werden kann.

Bei der Imputation für den LFS kamen mit "Geschlecht", "Familienstand", "Stellung im Haushalt" und "Erwerbstätigkeit" 4 qualitative und mit "Alter" ein quantitatives Merkmal zum Zug. "Alter" war insgesamt die am stärksten gewichtete Variable, die anderen Merkmale waren alle gleichrangig. Der Parameter *maxdiff* bei "Alter" lag bei 8 Jahren.

2.4 Aufsuchen des ähnlichsten Donor

Der Vorgang der Imputation von Merkmalen in einem fehlerhaften Datensatz reduziert sich damit auf das Auffinden jenes Datensatzes, welcher alle fehlenden Merkmale korrekt eingesetzt hat und zu dem gegebenen eine minimale Distanz hat.

Durch das Imputationsprofil eines fehlerhaften Datensatzes wird a priori eine Menge möglicher Donor-Datensätze festgelegt. Mögliche Donor sind durch deren Antwort- bzw. Imputationsprofil charakterisiert. Es gilt nämlich: Ein Datensatz ist dann und nur dann möglicher Donor, wenn er an jenen Stellen, an denen der zu korrigierende Datensatz im Imputationsprofil ein Bit gesetzt hat, im eigenen Antwortprofil ein Bit gesetzt hat und das entsprechende Bit im Imputationsprofil nicht gesetzt ist. Die Mächtigkeit der Menge der möglichen Donor-Datensätze hängt damit vor allem davon ab, wie viele Bits im eigenen Imputationsprofile gesetzt sind.

Nachdem die Menge der möglichen Spender festgelegt worden ist, gilt es jenen Datensatz darin zu finden, der die zuvor erläuterte Distanzfunktion minimiert. Hierbei gibt es zunächst das Problem, daß es zumeist nicht möglich ist, die Distanzfunktion zu jedem möglichen Donor zu berechnen, da dies zu enormen Rechenzeiten führen würde. Man muß sich also auf eine Teilmenge der Menge der möglichen Donor beschränken. Natürlich ist es wünschenswert, in dieser Teilmenge jenen Donor mit der minimalen Distanz zu finden (zumindest in den meisten Fällen). Man kann diese Teilmenge so auswählen, daß die möglichen Donor-Datensätze in einer oder mehreren (evtl. auch allen) soziodemographischen qualitativen Merkmalen, die in die Distanzfunktion eingehen mit denen des zu korrigierenden Satzes übereinstimmen. Durch die Übereinstimmung liefern diese Variable keinen Beitrag zur Distanzfunktion. Verwendet man aber zu viele Variable für die Auswahl dieser Teilmenge, so besteht die Gefahr, daß nur sehr wenige mögliche Donor übrig bleiben, unter welchen man dann einen passenden finden muß. Dadurch kann es passieren, daß verschiedene Datensätze unverhältnismäßig oft als Donor fungieren.

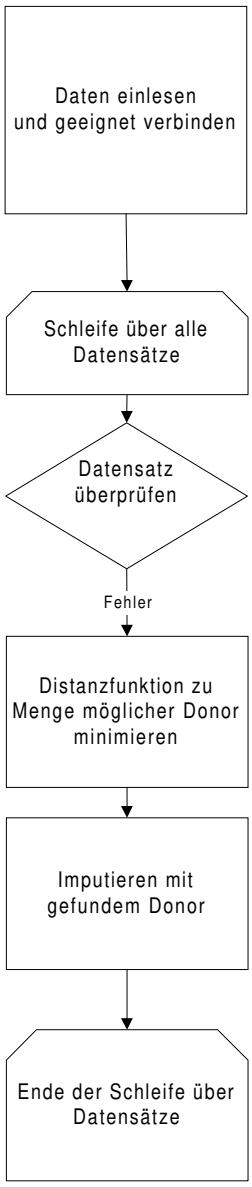
Ein weiteres Problem liegt darin, daß das Minimum der Distanzfunktion nicht eindeutig sein muß. Findet man zum Beispiel einen Datensatz, der in allen Merkmalen der Distanzfunktion mit dem zu korrigierenden Satz übereinstimmt, so ist die Distanz 0. Man kann nun die Suche abbrechen und mit dem gefunden Donor imputieren. Würde man nun einen anderen Satz mit gleichen soziodemographischen Merkmalen und gleichem Informationsprofil korrigieren wollen, so käme man unweigerlich zum gleichen Donor. Dadurch käme es zu unerwünschten Klumpungseffekten. Daher ist es notwendig mitzuführen, wie oft ein Datensatz schon als Donor gedient hat.

3 Algorithmische Betrachtungen

3.1 Der Algorithmus

Die Methode, die vorab beschrieben wurde, wurde implementiert und dahingehend durchgeführt, daß die Menge der möglichen Donor derart eingeschränkt wurde, daß für jeden zu korrigierenden Datensatz Übereinstimmung in jenem qualitativen Merkmal verlangt wurde, welches die geringste Auftrittshäufigkeit hat. Abbildung 3 gibt eine simplifizierte Darstellung des verwendeten Algorithmus wieder.

Abbildung 3: Simplifizierte Darstellung des verwendeten Algorithmus



Zunächst werden alle Datensätze in eine Datenstruktur eingelesen. Die Datensätze werden nun sequentiell abgearbeitet. Wird bei einem Datensatz ein fehlender Wert festgestellt, so wird sowohl sein Imputations- als auch sein Antwortprofil ermittelt. Nun werden die möglichen Donor-Datensätze sequentiell abgearbeitet. Wird ein Datensatz, der noch nie als Donor fungiert hat mit Distanz 0 gefunden, so wird die Suche abgebrochen. Wird kein Donor mit Distanz 0 gefunden, so wird jener Datensatz als Donor genommen, der die minimale Distanz aufweist.

Für ca. 60.000 Datensätze ergab sich mit diesem Algorithmus eine Rechenzeit von ca. 7 CPU-Stunden. Daraus ersieht man die Notwendigkeit weiterer algorithmischer Verbesserungen, will man sich Projekten zuwenden, die mit mehr Datensätzen arbeiten.

Die Gesamtrechenzeit T zerfällt in T_0, der Zeit zum Einlesen der Datensätze und Aufbau der Struktur und $N \cdot T_R$, wobei N die Anzahl der zu korrigierenden Datensätze ist. Nachdem die Zeit T_0 nur einmal während des ganzen Programmlaufs anfällt, ist es sinnvoll, die Datenstruktur mit einem hohen Maß an Informationen anzureichern, die dann die Zeit T_R für den jeweiligen Datensatz verkürzen.

3.2 Aufbau einer sinnvollen Datenstruktur

Eine Hauptaufgabe bei der Implementierung von Algorithmen, die mit großen Datenmengen arbeiten, liegt darin, Datenstrukturen zu schaffen, die auf geeignete Verbindungen unter den einzelnen Datensätzen zurückgreifen können.

Der Aufbau dieser Verbindungen geschieht mittels Zeigervariablen. Die Programmiersprache PLI bietet dafür geeignete Implementationsmöglichkeiten. Abbildung 4 gibt vereinfacht den Source der verwendeten PLI-Struktur wieder.

Abbildung 4: Vereinfachter Source der verwendeten PLI-Struktur

```
1 RECO
  2 . RECO_DATA
  2  DONOR BIN FIXED(31,0),
  2  DON_F(64) BIN FIXED(31,0),
  2  IMP BIT(1),
  2  IMP_PRF(64) BIT(1),
  2  ANT_PRF(64) BIT(1),
  2  Q1_P(2) POINTER,
  2  Q2_P(4) POINTER,
  2  Q3_P(5) POINTER,
  2  Q4_P(6) POINTER,
  2  NEXT_PAS(64)  POINTER,
  2  RECO_BEF  POINTER,
  2  RECO_NEXT POINTER;
```

RECO_DATA enthält die Daten des eingelesenen Datensatzes. DONOR und DONOR_F(64) enthalten Informationen, wie oft dieser Datensatz schon als Donor fungiert hat, bzw. wie oft er als Donor für jedes spezifische Feld gedient hat. Das Bit IMP wird gesetzt wenn auf dem Datensatz selbst Imputationen vorgenommen wurden. In den Bitfeldern IMP_PRF und ANT_PRF werden Imputations- bzw. Antwortprofil gespei-

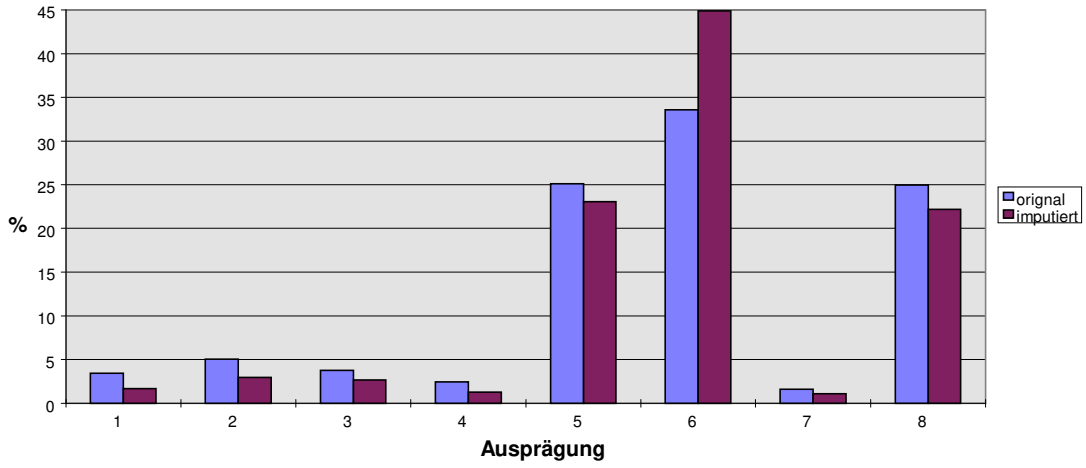
chert. Für die Funktionalität des Algorithmus entscheidend sind die Zeigerfelder Q1-Q4 und NEXT_PAS. Zum Beispiel zeigt der Pointer Q1(1) auf den nächsten Datensatz, der in der qualitativen Variable Q1 ("Geschlecht") den Wert 1 ("männlich") hat. Die Pointerfelder Q1-Q4 stehen für jene 4 qualitativen Merkmale, die in die Distanzfunktion eingehen. Die Zeiger des Feldes NEXT_PAS(64) zeigen auf jenen Datensatz, wo das entsprechende Bit im Antwortprofil gesetzt und im Imputationsprofil nicht gesetzt ist. RECO_BEF und RECO_NEXT zeigen schließlich auf den zuvor bzw. danach eingelesenen Datensatz.

4 Ergebnisse

4.1 Abweichungen der Verteilung

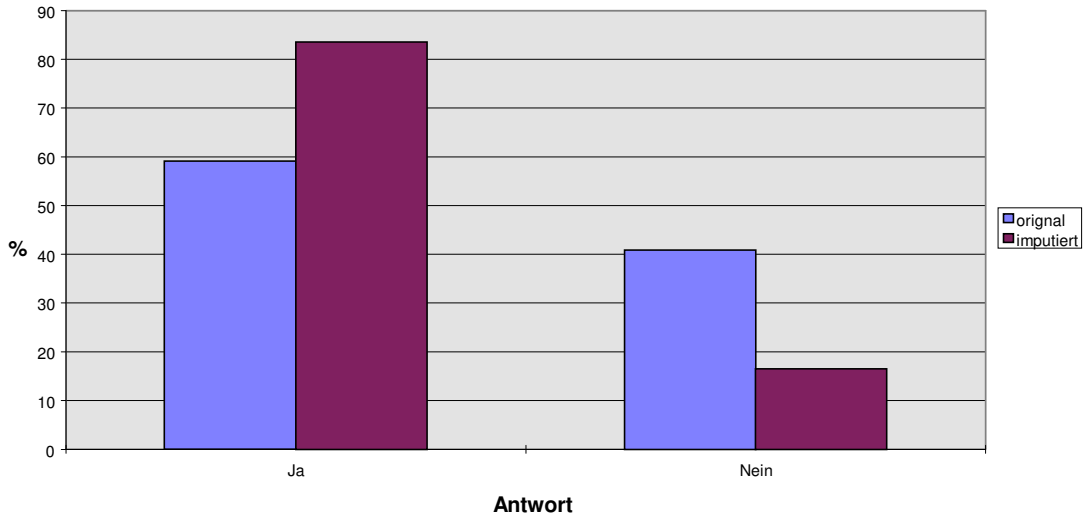
Eine interessante Frage war, ob nach erfolgter Imputation die Verteilung unter den imputierten Werten annähernd die gleiche war, wie unter den nicht-imputierten. Zum Großteil ist diese Frage zu bejahen. Es gab einzelne Ausreißer bei solchen Fragen, die bedingt durch die Fragebogenstruktur eine geringe Besetzungszahl aufwiesen. Abbildung 5 zeigt eine typische prozentuale Verteilung anhand des Merkmals "Grund für das Ende der letzten "Erwerbstätigkeit". Insgesamt gab es bei diesem Merkmal 9220 nicht imputierte und 2993 imputierte Werte.

Abbildung 5: Verteilung des Merkmals “Gund für Ende der letzten Erwerbstätigkeit”



Zum Unterschied dazu zeigt Abbildung 6 eine Ausreißerverteilung beim Merkmal "Empfänger von Arbeitslosengeld bzw. Notstandshilfe" (1780 originale und 291 imputierte Werte).

Abbildung 6: Verteilung des Merkmals “Bezieher von Arbeitslosengeld (Ja/Nein)”



4.2 Möglichkeiten zur Weiterverarbeitung

Durch die vorliegenden Ergebnisse ergibt sich die Möglichkeit, die Antwortverweigerer auf bestimmte Fragen näher zu charakterisieren. Obige Abbildung legt z.B. den Schluß nahe, daß Leute, die Arbeitslosengeld (bzw. Notstandshilfe) beziehen häufiger die Antwort verweigerten.

Ebenso wäre es denkbar, durch Anwendung bestimmter statistischer multivariater Verfahren (z.B. Diskriminanzanlyse, logistische Regression) zu tieferen Erkenntnissen über die Antwortverweigerer auf bestimmte Fragen zu gelangen.

Literatur

BANKIER, M. (1994). Imputing Numeric and Qualitative Variables Simultaneously. *Statistics Canada.*

HASLINGER, A. (1996). Stichprobenplan für den Mikrozenus ab 1994. *Statistische Nachrichten,* Heft 4/1996.

EUROSTAT (1993). Erhebung über Arbeitskräfte, Methodik und Definitionen.

KALTON, G., and KASPRZYK, D. (1982). Imputing for missing Survey Responses. *Proceedings of the Section on Survey Research Methods*. American Statistical Association.

Adresse des Autors:

Dipl.-Ing. Thomas Burg
Österreichisches Statistisches Zentralamt
Technisch-Methodische Abteilung
Hintere Zollamtsstraße 2b
A-1033 Wien