DESTATIS
wissen.nutzen.

LDS NRW. Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen

ZUMA Zentrum für Umfragen, Methoden und Analysen

**Methodenverbund**

**„Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“**

**Arbeitspapier Nr.: 11**

Heike Wirth
Zentrum für Umfragen, Methoden und Analysen (ZUMA)

# Anonymisierung des Mikrozensuspanels im Kontext der Bereitstellung als Scientific-Use-File

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Deutsche Forschungsgemeinschaft
DFG

*Methodenverbund*

*"Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe"*

*Arbeitspapier Nr. 11*

# *Anonymisierung des Mikrozensuspanels im Kontext der Bereitstellung als Scientific-Use-File*

*Heike Wirth, ZUMA, Mannheim*

*März 2006*

## *1. Einleitung*

Nach §16 (Abs.6) Bundesstatistikgesetz 1987 darf die amtliche Statistik für die Durchführung wissenschaftlicher Vorhaben Einzelangaben an wissenschaftliche Einrichtungen übermitteln, sofern die Einzelangaben faktisch anonym sind, d.h. eine Reidentifikation nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft möglich ist.[1] Für den Mikrozensus wurde das Konzept der faktischen Anonymität im Rahmen des so genannten 'Anonymisierungsprojektes' (Müller et al. 1991) operationalisiert. Hierbei wurden die im Wissenschaftsbereich vorliegenden Randbedingungen (Zusatzwissen, Reidentifikationsmotive und -strategien, etc.), die für ein hypothetisches Reidentifikationsvorhaben relevant sein könnten, in umfassender Weise untersucht. Im Weiteren wurden zentrale 'Angriffsszenarien' abgeleitet, für welche detaillierte Analysen des Reidentifikationsrisikos und des damit verbundenen Aufwands vorgenommen wurden. Auf Basis dieser umfangreichen empirischen Risikoanalysen wurden Empfehlungen für die Umsetzung der faktischen Anonymisierung erarbeitet, die spezifischen Risikokonstellationen gezielt entgegenwirken bei einem zugleich möglichst schonenden Eingriff in das Analysepotenzial der Daten. Die von dem Projekt vorgeschlagenen Anonymisierungsempfehlungen beziehen sich dabei auf zwei unterschiedliche Datentypen: Mikrozensus-Grundfile und Mikrozensus-Regionalfile. Beide Datentypen enthalten jeweils nur Querschnittsdaten.

Nun ist der Mikrozensus vom Erhebungsdesign jedoch als rotierende Panel-Stichprobe angelegt und seit dem Mikrozensusgesetz 1996 darf die amtliche Statistik die für die Zusammenführung der Paneldaten benötigten Ordnungsnummern wieder speichern. Damit ist erstmals die Möglichkeit gegeben, das Potenzial des Mikrozensus für Panelanalysen zu nutzen und der Wissenschaft faktisch anonymisierte Paneldaten als Scientific-Use-File zur Verfügung zu stellen.

Im Folgenden werden konkrete Maßnahmen zur faktischen Anonymisierung des Mikrozensuspanels empfohlen. Die Empfehlungen beziehen sich sowohl auf Vier-Jahres- als auch auf Drei- und Zwei-Jahres-Längsschnitte. Als Grundlage dienen die zentralen Befunde aus dem Anonymisierungsprojekt, dessen grundsätzliche Empfehlungen zur faktischen Anonymisierung von Mikrodaten zwischenzeitlich nicht nur beim Mikrozensus und der Einkommens- und Verbrauchsstichprobe Anwendung finden, sondern bspw. auch für die Beschäftigtenstichprobe und die Zeitbudgeterhebungen übernommen wurden. Es ist an dieser Stelle nicht möglich, die umfangreichen Einzelbefunde des Anonymisierungsprojektes im Detail darzustellen. Allerdings lassen sich die zentralen Befunde wie folgt zusammenfassen (vgl. Müller et al. 1991: XIV ff.): Generell hat sich gezeigt, dass das Reidentifikationsrisiko für den Mikrozensus unter realen Bedingungen erheblich geringer ist, als aufgrund von theoretischen Erwägungen oder Simulationsrechnungen gemeinhin angenommen wurde. (1) Selbst wenn leicht verfügbares Zusatzwissen zur Verfügung steht, ist es nicht möglich eine größere Anzahl von Personen zu reidentifizieren, solange der Auswahlsatz des Mikrodatenfiles so klein ist wie bspw. beim Mikrozensus. (2) Auch bei Strategien der gezielten Suche (bspw. Wissen, dass eine

[1] Das Grundprinzip einer Reidentifikation beruht darauf, dass Einzeldatensätze einer Datei mit anonymen Mikrodaten den Einzeldatensätzen einer anderen personenbezogenen Datei (**Zusatzwissen**) in einer Eins-zu-eins-Entsprechung zugeordnet werden. Reidentifikationsversuche setzen hierbei an jenen Merkmalen an, die sowohl im Mikrodatenfile als auch im Zusatzwissen enthalten sind (**Überschneidungsmerkmale**). Durch Abgleich der Überschneidungsmerkmale auf Identität oder Ähnlichkeit wird angestrebt, jene Einzeldatensätze in den beiden Datenfiles herauszufinden, die von ein und derselben Person stammen.

bestimmte Person am Mikrozensus teilgenommen hat, bzw. **Teilnahmekenntnis**) erwiesen sich die Mikrodaten - selbst unter Berücksichtigung einer Vielzahl familien-, haushalts- und erwerbsbezogener Merkmale - ohne weitergehende Anonymisierungsmaßnahmen als faktisch anonym. Die wesentliche Erkenntnis der durchgeführten Experimente bestand darin, dass das Vorhandensein einzigartiger Ausprägungskombinationen keinesfalls eine hinreichende Bedingung für eine Reidentifikation darstellt. Die wichtigste Ursache hierfür ist die praktische Unvermeidbarkeit von Inkompatibilitäten zwischen Daten aus unterschiedlichen Generierungsprozessen. Derartige Inkompatibilitäten, die durch unterschiedliche Erhebungskontexte, Erhebungszeitpunkte, Erhebungsziele, Frageformulierungen, Erhebungsmodi, oder Erhebungs-, Übertragungs- und Kodierungsfehler etc. bedingt sind, führen zu Falsch- und Nichtzuordnungen, d.h. sie entfalten eine gewissermaßen natürliche Schutzwirkung von Reidentifikationsrisiken. Nur beim Zusammentreffen sehr spezifischer Randbedingungen und unter der Voraussetzung, dass der betreffende Forscher überhaupt ein wie immer geartetes Interesse[2] an der nachträglichen Herstellung eines Personenbezugs hat, besteht unter Umständen die Möglichkeit einer Reidentifikation (Müller et al. 1991: XVII): [3]

- Die im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezielles Merkmal eingrenzbarer Subpopulation an (z.B. herausgehobene Berufsgruppen wie etwa Professoren oder Abgeordnete);
- Das Mikrodatenfile enthält sehr tiefgegliederte Regionalinformationen, so dass in den einzelnen Regionaleinheiten nur wenige Personen der spezifischen Subpopulation leben;
- Dem Forscher ist bekannt, dass die gesuchte Person im Mikrodatenfile enthalten ist (Teilnahmekenntnis);
- Die Merkmale der gesuchten Person sind genau in der Weise im Mikrodatenfile erfasst, wie es der Forscher vermutet (Kompatibilität der Information in Zusatzwissen und Mikrodatenfile).

Wenngleich das gleichzeitige Zusammentreffen dieser vier Faktoren als ein höchst unwahrscheinliches Ereignis anzusehen ist, sehen die Empfehlungen des Anonymisierungsprojektes umfangreiche Schutzvorkehrungen vor (vgl. Abschnitt 3), um bei einer Datenübermittlung an die Wissenschaft auch einer solch unwahrscheinlichen Risikokonstellation entgegen zu wirken.

---

2 In diesem Zusammenhang sei der Hinweis erlaubt, dass die Forschung schon seit geraumer Zeit mit amtlichen Mikrodaten arbeitet und bislang kein einziger Vertrauensbruch bekannt wurde. Zugleich haben die Analysen im Rahmen des Anonymisierungsprojektes gezeigt, dass der Nutzen, den ein Wissenschaftler aus einer Deanonymisierung von Mikrozensus-Daten ziehen könnte, allenfalls marginal ist, so dass der Aufwand, der für eine Reidentifikation betrieben werden müsste, den hieraus zu erzielenden Nutzen in aller Regel bei weitem übersteigt (Müller et al. 1991: 212ff.).

3 Die damaligen Befunde haben auch in der Gegenwart Bestand. Dies zeigt eine neuere Studie von Bender et al. (2001), bei welcher gleichfalls Reidentifikationsexperimente mit realen Daten durchgeführt wurden. Die Datenkonstellation in dieser Studie war denkbar günstig, da schon bei der Datenerhebung des Zusatzwissens hohe Qualitätsmaßstäbe angelegt wurden. Weiterhin handelt es sich bei beiden Datenfiles (Zusatzwissen und Mikrodatenfile) um Ereignisdaten mit differenzierten Verlaufsangaben. Darüber hinaus lag für über 200 Personen so genannte Teilnahmekenntnis vor. Das heißt, es war bekannt, dass diese Personen auch im Mikrodatenfile enthalten sein müssen. Trotz dieses hochriskanten Angriffsszenarios gelang es nicht, auch nur einen einzigen Fall zu reidentifizieren: So war nicht nur die weitaus überwiegende Mehrheit (86%) der Zuordnungen falsch, sondern es war darüber hinaus auch nicht möglich, die falschen von den korrekten Zuordnungen zu unterscheiden.

Das Mikrozensuspanel unterscheidet sich von den im Anonymisierungsprojekt untersuchten Randbedingungen nun jedoch dadurch, dass die Daten der Befragten nicht nur im Querschnitt, sondern auch im Längsschnitt für jeweils maximal vier Erhebungszeitpunkte vorliegen. Aufgrund des damit einhergehenden höheren Informationsgehaltes der Paneldaten scheint eine Eingrenzung einzelner Personen auf den ersten Blick einfacher als bei Querschnittsdaten. Bei einer näheren Betrachtung der empirischen Randbedingungen wird - wie im Folgenden aufgezeigt werden wird - allerdings deutlich, dass das Reidentifikationsrisiko bei den Mikrozensus-Paneldaten nicht per se höher ist als bei den Querschnittsdaten. Zur Untermauerung dieser These werden nicht nur die zentralen Befunde aus dem Anonymisierungsprojekt, sondern auch die im Rahmen des Verbundprojektes „Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“[4] gewonnenen empirischen Erkenntnisse herangezogen.

## *2. Zentrale Bedingungsfaktoren des Reidentifikationsrisikos des MZ-Panels*

Wie eingangs erwähnt, besteht mit Blick auf das Reidentifikationsrisiko der wesentliche Unterschied zwischen den Mikrozensus-Paneldaten und den Querschnittsdaten darin, dass die Informationen über die Befragten im Panel nicht nur für einen, sondern für bis zu vier Erhebungszeitpunkte vorliegen. Unter idealtypischen Bedingungen würde dieser höhere Informationsgehalt der Daten auch mit einem erhöhten Reidentifikationsrisiko einhergehen. Unter empirischen Bedingungen trifft diese These hingegen nur dann zu, wenn die für eine Reidentifikation notwendigen Randbedingungen erfüllt sind. Dies betrifft zunächst die Kriterien Verfügbarkeit von Zusatzwissen sowie Kompatibilität zwischen Zusatzwissen und Mikrodatenfile (MZ-Panel).

### *Verfügbarkeit von Zusatzwissen*

Unabdingbare Voraussetzung für jegliche Reidentifikationsversuche ist die Verfügbarkeit von geeignetem **personenbezogenem Zusatzwissen**. Mit dem Begriff Zusatzwissen sind Daten gemeint, die durch spezifische Charakteristika gekennzeichnet sind: (A) Das Zusatzwissen muss personenbezogen sein, d.h. direkte Personenidentifikatoren (Namen, Adressen) enthalten. (B) Die im Zusatzwissen enthaltenen Angaben müssen Überschneidungen zum Mikrodatenfile aufweisen. Man spricht hier auch von 'Überschneidungsmerkmalen' (z.B. Alter, Geschlecht, Beruf). (C) Die im Zusatzwissen erfasste Population muss Überschneidungen zum Mikrodatenfile aufweisen. D.h. zumindest eine im Mikrozensuspanel erfasste Person muss auch im Zusatzwissen enthalten sein. Jeglicher Reidentifikationsversuch - unabhängig davon, ob es sich um ein einfaches oder aufwendiges statistisches Verfahren handelt - basiert dann auf einem Abgleich der Überschneidungsmerkmale zwischen dem Zusatzwissen und dem Mikrodatenfile auf Identität oder Ähnlichkeit der Einzeldatensätze in den Ausprägungskombinationen.

Nach derzeitigem Kenntnisstand liegen im Wissenschaftsbereich keine personenbezogenen Paneldaten vor, die als Zusatzwissen für Reidentifikationsversuche des Mikrozensus eingesetzt

[4] Siehe http://www.destatis.de/mv/mzpanel_start.htm.

werden könnten. Soweit Paneldaten für die Wissenschaft verfügbar sind, wie bspw. die IAB-Beschäftigtenstichprobe, das Sozio-oekonomische Panel (SOEP) oder der Familiensurvey des DJI, sind diese nicht personenbezogen, sondern anonymisiert und damit für Reidentifikationszwecke untauglich.

Hinsichtlich öffentlich zugänglicher Datenquellen ergibt sich ein entsprechendes Bild. Frühere Recherchen im Rahmen des Anonymisierungsprojektes haben gezeigt, dass solche Informationsquellen generell nur sehr wenige Überschneidungsmerkmale zum Mikrozensus aufweisen (vgl. Müller et al. 1991: 165ff). Es liegen keine Hinweise vor, dass sich hieran im Verlauf der letzten 15 Jahre etwas geändert hat. Kommt als weitere Bedingung hinzu, dass Überschneidungsmerkmale für mehrere Zeitpunkte vorliegen müssen, wird der Umfang des potenziell nutzbaren Zusatzwissens noch einmal erheblich eingeschränkt.

### *Kompatibilität von Zusatzwissen und Mikrodatenfile*

Weiterhin ist zu berücksichtigen, dass die Verfügbarkeit von personenbezogenem Zusatzwissen zwar eine notwendige, aber keinesfalls eine hinreichende Bedingung für eine Reidentifikation ist. Der Erfolg eines Reidentifikationsversuchs ist vielmehr wesentlich davon abhängig, inwieweit die für die Reidentifikation verfügbaren Überschneidungsmerkmale im Zusatzwissen und Mikrodatenfile kompatibel abgebildet sind. In theoretischen Überlegungen wie auch bei Simulationsstudien zur Einschätzung von Reidentifikationsrisiken spielen Dateninkompatibilitäten - sofern überhaupt berücksichtigt - eine allenfalls untergeordnete Rolle. Unter empirischen Randbedingungen zeigt sich jedoch, dass wann immer Informationen aus unterschiedlichen Quellen verglichen werden, mit zum Teil erheblichen Inkompatibilitäten zu rechnen ist. Die Ursachen hierfür sind vielfältig. Sie reichen von unterschiedlichen Frageformulierungen oder Antwortvorgaben, unterschiedlichen Erhebungszeitpunkten, Erhebungsmodi und sachlichem Bezug, über Interviewereffekte oder Missverständnissen auf Seiten der Befragten bis hin zu Verkodungsfehlern.[5]

So unangenehm derartige Dateninkompatibilitäten für die Forschung auch sein mögen, in Hinblick auf das Reidentifikationsrisiko geht von ihnen eine enorme Schutzwirkung aus, da sie die korrekte Zuordnung von Einzeldatensätzen aus Zusatzwissen und Mikrodatenfile beeinträchtigen. Dieser für die Mikrozensus-Querschnittsdaten im Kontext des Anonymisierungsprojekts dokumentierte Schutzeffekt (Müller et al. 1991: 112ff.), dessen Wirksamkeit auch durch eine weitere empirische Studie mit anderen Datengrundlagen (Bender et al. 2001; vgl. auch Fußnote 3) belegt ist, sollte sich bei den MZ-Paneldaten sogar noch verstärken. Denn im Unterschied zu den Querschnittsdaten sind bei den Paneldaten nicht nur ein Erhebungszeitpunkt, sondern bis zu vier Erhebungszeitpunkte durch das potenzielle Zusatzwissen abzudecken, sofern der erhöhte Informationsgehalt der Paneldaten für Reidentifikationsversuche genutzt werden soll. Damit potenzieren sich jedoch auch die Quellen möglicher Inkompatibilitäten zwischen Paneldaten und Zusatzwissen.

[5] Zu der allgemeinen Problematik vgl. z.B. Paull (2002), Abowed/Stinson (2003), Sala/Lynn (2004), Reimer (2004), Reimer/Künster (2004), Lynn et al. (2004), Herter-Eschweiler (2005).

Zunächst betrifft dies die zeitlichen Bezugspunkte. Die Angaben des Mikrozensus beziehen sich - von wenigen Ausnahmen abgesehen (wie z.B. den Retrospektivfragen) - auf eine bestimmte Berichtswoche. In den Jahren 1996 bis 2004 lag diese meist im April, vereinzelt fiel sie jedoch auch in den März bzw. den Mai. Seit dem Jahr 2005 wird beim Mikrozensus mit einer gleitenden Berichtswoche gearbeitet. Informationen aus anderen Datenquellen beziehen sich in der Regel entweder auf einen festgelegten Stichtag oder jeweils auf den Tag, an dem die Daten erhoben wurden. Insgesamt führt diese Konstellation dazu, dass mit jeder Einbeziehung eines weiteren Erhebungsjahres auch die Wahrscheinlichkeit dafür ansteigt, dass ein Reidentifikationsversuch an dem Sachverhalt scheitert, dass in unterschiedlichen Datenquellen erfasste Veränderungen sich auf unterschiedliche Zeitpunkte beziehen und daher in aller Regel nicht deckungsgleich abgebildet sind.

Einen zweiten relevanten Komplex bilden das erhobene Merkmalsspektrum, der Erhebungskontext, die jeweiligen Frageformulierungen und Antwortvorgaben. Diesbezügliche Variationen führen bereits im Querschnitt dazu, dass Merkmale aus unterschiedlichen Datenquellen nicht deckungsgleich abgebildet. Darüber hinaus kommt es hier bei fast allen Erhebungen im Verlauf der Jahre zu kleineren oder größeren Veränderungen. Je mehr Erhebungsjahre in eine Betrachtung einbezogen werden, desto geringer wird daher die Wahrscheinlichkeit, dass bei den entscheidenden Merkmalen des Mikrozensus-Panels und eines potenziellen Zusatzwissens Deckungsgleichheit gegeben ist.

Schließlich geht auch von inkonsistenten Angaben im Zeitverlauf eine nicht zu unterschätzende Schutzwirkung aus. Probleme dieser Art treten bei praktisch allen Erhebungen auf (vgl. z.B. Black et al. 2003 oder Lynn 2003).[6] Für den hier interessierenden Zusammenhang lässt sich dies am Beispiel der der 1984 durchgeführten ALLBUS Test-Retest-Studie und aktuellen Befunden im Kontext der Aufbereitung des Mikrozensus als Panel illustrieren. Bei der im Rahmen des ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) durchgeführten Test-Retest-Studie wurde einer Teilstichprobe (n=154) der gleiche Fragebogen im Monatsabstand insgesamt dreimal vorgelegt (Porst/Zeifang 1987a,b; Koch 1986). Wie aus Tabelle 1 hervorgeht, zeigen die soziodemografischen Merkmale zwar eine relativ hohe Antwortstabilität, mit Ausnahme des Merkmals Geschlecht sind jedoch alle betrachteten Merkmale mit einer gewissen Wahrscheinlichkeit von Abweichungen betroffen. Hiervon sind auch relativ einfache Merkmale wie Geburtsjahr, Heiratsjahr und Kinderzahl nicht ausgenommen, wobei die Abweichungen beim Geburtsjahr nicht und bei den Merkmalen Heiratsjahr und Kinderzahl nur zu einem geringen Teil auf Statusveränderungen zurückführbar sein dürften, da zwischen den verglichenen Erhebungszeitpunkten nur rund zwei Monate lagen. Darüber hinaus finden sich aber auch bei den verschiedenen Angaben zum Erwerbskontext zum Teil erhebliche Variationen, die aufgrund des relativen kurzen Zeitabstands zwischen den Erhebungen, vermutlich gleichfalls eher auf Antwortinkonsistenzen und/oder Datenfehlern als auf tatsächlichen Statusveränderungen beruhen dürften.

[6] So berichtet Lynn (2003: 65) etwa: "Overall, only 39.9% of respondents interviewed three times over a 12 month had the same three digit occupation and industry codes, even though they were with the same employer at each interview. Similar findings have been made on UK surveys such as the panel element of the Labour Force Survey."

***Tabelle 1: Anteil der Befragten der ALLBUS Test-Retest-Studie, die bei der dritten Welle andere Angaben machten als bei der ersten Welle (in Prozent)***

| Ausgewählte Merkmale*) | In % |
|---|---|
| Geschlecht | 0,0 |
| Geburtsjahr | 2,6 |
| Familienstand | 1,3 |
| Heiratsjahr | 9,1 |
| Schule derzeit | 1,3 |
| Erwerbstätig | 6,5 |
| Arbeitslosigkeit | 5,2 |
| Früher erwerbstätig | 8,4 |
| Jahr, in welchem zuletzt eine Erwerbstätigkeit ausgeübt wurde | 24,7 |
| Branche | 11,0 |
| Beruf | 7,8 |
| Stellung im Beruf | 6,5 |
| Arbeitswochenstunden | 18,2 |
| Allgemeinbildender Schulabschluss | 7,8 |
| Beruflicher Ausbildungsabschluss | 16,2 |
| Überwiegende Einkünfte | 13,0 |
| Einkommen | 44,8 |
| Kinderzahl | 7,1 |

Quelle: Auszug aus Müller et al. (1991: 125)

*) Es wurden nur solche Fragen ausgewählt, die in der amtlichen Statistik in ähnlicher Form gestellt werden.

Für das Mikrozensus-Panel stellt sich die Sachlage ähnlich dar, wobei allerdings zu beachten ist, dass die Erhebung nicht wie die Test-Retest-Studie monatlich, sondern jährlich durchgeführt wurde. Wird bspw. das Merkmal „Stellung im Beruf in der gegenwärtigen Tätigkeit“ über vier Erhebungszeitpunkte hinweg betrachtet (1996-99), so weisen insgesamt 26 Prozent der Personen im Mikrozensus-Panel Veränderungen im Zeitverlauf auf. Für das Merkmal „Familienstand“ ergibt sich ein Prozentwert von 5,2. Beim höchsten allgemeinen Schulabschluss beträgt dieser 29,4 und beim Wirtschaftszweig des Betriebes, in dem eine Person beschäftigt ist, 34,7 (vgl. Herter-Eschweiler 2003: 243ff).[7] Welchen Anteil inkonsistente Angaben im Vergleich zu tatsächlichen Veränderungen haben, lässt sich letztlich nicht ermitteln. Der vergleichsweise hohe Prozentsatz der Personen mit Veränderung beim Wirtschaftszweig deutet aber z.B. auf die Existenz von inkonsistenten bzw. fehlerhaften Angaben hin. Eine Erklärung hierfür könnte sein, dass die Interpretationsspielräume bei diesem Merkmal relativ groß sind und beim Mikrozensus auch Proxy-Interviews durchgeführt werden. Insgesamt kommt Herter-Eschweiler bei seinen Untersuchungen zu Antwortstabilitäten und Antwortinkonsistenzen für ausgewählte Merkmale im Mikrozensus-Panel zu der Schlussfolgerung, dass die Ergebnisse im Großen und Ganzen den Ergebnissen von Test-Retest-Studien und Untersuchungen zur Validität und Reliabilität von Proxy-Interviews entsprechen (vgl. Herter-Eschweiler 2003: 249).

[7] Für eine ausführliche Diskussion dieser und weiterer Befunde siehe Herter-Eschweiler (2003).

Zusammenfassend lässt sich an dieser Stelle festhalten, dass sowohl von tatsächlichen Veränderungen im Zeitverlauf als auch von Antwortinkonsistenzen eine erhebliche Schutzwirkung im Hinblick auf die faktische Anonymisierung des Mikrozensus-Panels ausgeht. Denn erstens ist davon auszugehen, dass, wie eingangs ausgeführt, zeitliche Veränderungen im Mikrozensus-Panelfile und einem potenziellen Zusatzwissen - aufgrund unterschiedlicher Zeitbezüge - nicht deckungsgleich abgebildet sind. Zweitens ist für alle Datenbestände ein gewisses 'Rauschen' (Noise) aufgrund von Inkonsistenzen anzunehmen. In der Summe führt zu dies zu Abweichungen bzw. so genannten Inkompatibilitäten zwischen Daten aus unterschiedlichen Erhebungskontexten, die zusätzlich durch unterschiedliche Frageformulierungen, Antwortvorgaben oder Erhebungsmodi verstärkt werden. Da mit jedem zusätzlich einbezogenen Erhebungszeitpunkt zugleich die Wahrscheinlichkeit einer abweichenden Abbildung von Merkmalen zwischen Zusatzwissen und Mikrozensus ansteigt, bedeutet das Mehr an Informationen, das Paneldaten bieten, damit keinesfalls, dass hierdurch eine Reidentifikation einfacher würde.

### *Substichprobe*

Abschließend ist darauf hinzuweisen, dass Mikrozensus-Panelfiles nur einen Teil der Population des Mikrozensus enthalten. Bei Zwei-Jahres-Längsschnitten handelt es sich um maximal drei Viertel, bei Drei-Jahres-Längsschnitten um maximal die Hälfte. Sofern mit einem Vier-Jahres-Längsschnitt gearbeitet wird, ist lediglich etwa ein Viertel der Auswahlbezirke enthalten. Berücksichtigt man die im Mikrozensus nicht erfasste räumliche Mobilität von jährlich circa 10 Prozent, reduzieren sich die Anteile entsprechend. Infolgedessen verringert sich im Vergleich zu den Querschnittsdaten nochmals die Wahrscheinlichkeit, dass Personen sowohl im Zusatzwissen wie auch in den Paneldaten enthalten sind.

## *3. Folgerungen für die faktische Anonymisierung des Mikrozensus-Panels*

Ausgehend von den oben angeführten Argumenten ergibt sich für die Mikrozensus-Paneldaten im Vergleich zu den Querschnittsdaten kein erhöhtes Reidentifikationsrisiko. Eher im Gegenteil ist durch die deutlich geringere Stichprobengröße sogar ein verringertes Risiko anzunehmen. Infolgedessen werden neben allgemeinen Schutzvorkehrungen für die faktische Anonymisierung der Paneldaten die gleichen spezifischen Anonymisierungsmaßnahmen wie für die Mikrozensus-Querschnittsdaten empfohlen: Im Einzelnen bedeutet dies:

### *(A) Allgemeine Schutzvorkehrungen*

(1) Vertragliche Bindung des Empfängers faktisch anonymer Daten, in der u.a. Einzelheiten der Datennutzung und Nutzungskontrolle, Datensicherung und Datenlöschung sowie eine Vertragsstrafe bei Reidentifikationsversuchen vereinbart wird.

(2) Geheimhaltung der lokalen Umsetzung der Stichprobenpläne durch die amtliche Statistik, um zu verhindern, dass das Wissen um die Teilnehmer an einer amtlichen Erhebung (Teilnahmekenntnis) für Reidentifikationszwecke eingesetzt werden kann.

### *(B) Spezifische Anonymisierungsmaßnahmen*

(1) Vergröberung der Regionalangaben (Bundesland und Siedlungsstrukturtyp bzw. vergröberte Gemeindegrößenklasse) dergestalt, dass:

- Keine einzelne Gemeinde mit weniger als 500.000 Einwohnern identifizierbar ist
- Ein Gemeindetyp, dem mehrere Gemeinden zugehören, in jedem Bundesland mindestens 400.000 Einwohner umfasst.

(2) Keine Nationalität oder Gruppe von Nationalitäten mit weniger als 50.000 Einwohnern in der Bundesrepublik darf identifizierbar sein.

(3) Ausprägungsvergröberung bei allen übrigen Merkmalen - soweit erforderlich - so, dass in univariaten Randverteilung jede ausgewiesene Merkmalsausprägung für die Bundesrepublik mindestens 5.000 Fälle umfasst.

(4) Es wird eine Substichprobe weitergegeben, deren Auswahlsatz mindestens 70 % der Haushalte und Personen betragen sollte, unabhängig davon, ob die Substichprobe auf der Basis von Auswahlbezirken oder Wohnungen gezogen wird.

Im Detail werden diese Maßnahmen an anderer Stelle dargestellt und im Hinblick auf ihre Wirksamkeit begründet (Müller et al. 1991). In der Praxis haben sie sich nicht nur beim Mikrozensus, sondern auch bei anderen Scientific-Use-Files (z.B. Beschäftigtenstichprobe, Zeitbudgeterhebung, Einkommens- und Verbrauchsstichprobe) seit geraumer Zeit bewährt. Deshalb werden die vorgeschlagenen Anonymisierungsmaßnahmen an dieser Stelle nur noch kurz begründet:

Der wesentliche Risikofaktor für die faktische Anonymität von Bevölkerungsdaten besteht in der gleichzeitigen regionalen und sachlichen Tiefengliederung. Durch die starke Vergröberung der Regionalinformationen wird dieser Risikofaktor minimiert. Werden Daten ohne kleinräumige Regionalinformationen übermittelt, sind sie in aller Regel schon durch die Entfernung der personenbezogenen Angaben faktisch anonym. Dieser Effekt wird nochmals verstärkt durch die zusätzliche Festlegung eines Minimums in den univariaten Randverteilungen. Da die Staatsangehörigkeit ein in der Regel leicht und kompatibel mit dem Mikrodatenfile erfahrbares Merkmal ist, wird darüber hinaus empfohlen, dieses Merkmal nur stark vergröbert weiterzugeben. Weiterhin wird auch die Ziehung einer Substichprobe empfohlen. Die Substichprobenziehung verhindert, dass ein potenzieller Datenangreifer mit Sicherheit weiß, ob eine bestimmte Person im übermittelten Mikrodatenfile enthalten ist und erhöht damit den Unsicherheitsfaktor bei einem Reidentifikationsversuch beträchtlich. Zugleich wird durch die Substichprobenziehung die Reidentifikationswahrscheinlichkeit prinzipiell reduziert. Die Ziehung von Substichproben ist aber auch mit Einschränkungen in der Präzision von Analyseergebnissen verbunden. Bei den Paneldaten bedeutet die Substichprobenziehung - aufgrund der bereits deutlich kleineren Ausgangsstichprobe - eine größere Analyseeinschränkung als bei den Querschnittsdaten. Es wird deshalb empfohlen, bei den Paneldaten als unterste Grenze eine Substichprobe von 70 % anzusetzen.

Darüber hinaus wird empfohlen zeitnah zur Bereitstellung des Mikrozensuspanel Scientific-Use-File auch ein entsprechendes Campusfiles (5%-Substichprobe des anonymisierten Originalmaterials) für Lehrzwecke vorzubereiten. Sofern von der Wissenschaft ein eindeutiger Bedarf nach regionalisierten MZ-Paneldaten artikuliert wird, wäre zu überprüfen, inwieweit die Anonymisierungsempfehlungen für das MZ-Regionalfile (Querschnittsdaten) auf die Paneldaten übertragbar sind.

## *Literatur*

Abowed, John/Stinson, Martha, 2003: Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data. http://instruct1.cit.cornell.edu/~jma7/abowd-stinson-SOLE-2003.pdf

Bender, Stefan/Brand, Ruth/Bacher, Johann (2001): Re-Identifying register data by survey data: An empirical study. Statistical Journal of the United Nations ECE. S. 373-381.

Black, Dan/Sanders, Seth/Taylor, Lowel, 2003: Measurement of Higher Education in the Census and CPS.

Herter-Eschweiler, Robert, 2003: Längsschnittdaten aus dem Mikrozensus: Basis für neue Analysemöglichkeiten. Dokumentationsband. Statistisches Bundesamt. Bonn.

Herter-Eschweiler, Robert, 2005: II. Zur Datenqualität retrospektiver Angaben im Mikrozensus 2000 (Teilprojekt Q). Statistisches Bundesamt. Bonn.

Koch, Achim, 1986: Wie zuverlässig lassen sich Berufs- und Bildungsangaben messen? Ergebnisse einer Test-Retest-Studie zur Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 1984. Diplomarbeit (unveröffentlicht). Universität Mannheim.

Lynn, Peter (ed.), 2003: Quality Profile: British Household Panel Survey Waves 1 to 10: 1991-2000. Institute for Social and Economic Research. University of Essex.

Lynn, Peter/Jäckle, Annette/Jenkins, Stephen P./Sala, Emanuela, 2004: The impact of interviewing methods on measurement error in panel survey measures of benefit receipt: evidence from a validation study. ISER Working Papers. Number 2004-28.

Müller, Walter/Blien, Uwe/Knoche, Peter/Wirth, Heike u.a., 1991: Die faktische Anonymität von Mikrodaten. Die faktische Anonymität von Mikrodaten. In: Statistisches Bundesamt (Hrsg.), Schriftenreihe Forum der Bundesstatistik, Band 19. Stuttgart: Metzler-Poeschel.

Paull, Gillian, 2002: Biases in the Reporting of Labour Market Dynamics. The Institute for Fiscal Studies. WP02/10.

Porst, Rolf/Zeifang, Klaus (1987a): Wie stabil sind Umfragedaten? Beschreibung und erste Ergebnisse der Test-Restest-Studie zum ALLBUS 1984. In:: ZUMA-Nachrichten 20, S.8-31.

Porst, Rolf/Zeifang, Klaus (1987b): A description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables. In: Sociological Methods & Research. Vol. 15(3), S. 177-218.

Reimer, Maike, 2004: Collecting Event History Data: About Work Careers Retrospectively: Mistakes that Occur and Ways to prevent them. Arbeitspapier Nr. 1 des Projekts Kognition und Kommunikation bei der Längsschnittdatenerhebung.

Reimer, Maike/Künster, Ralf, 2004: Linking Job Episodes from Retrospective Surveys and Social Security data: Specific Challenges, Feasibility and Quality of Outcome. Arbeitspapier Nr. 8 des Projekts Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutschland.

Sala, Emanuela/Lynn, Peter, 2004: Measuring Change in Employment Characteristics: The effects of dependent interviewing. ISER Working Papers. Number 2004-26.

Schimpl-Neimanns, Bernhard, 2005: Bildungsverläufe im Mikrozensuspanel 1996-1999: Besuch der gymnasialen Oberstufe bis zum Abitur. ZUMA-Arbeitsbericht Nr. 2005/02.

Schimpl-Neimanns, Bernhard, 2006: Zur Datenqualität der Bildungsangaben im Mikrozensus am Beispiel des Besuchs der gymnasialen Oberstufe und des allgemeinen Schulabschlusses. Erscheint in: Tagungsband zur ASI 2005 Jahrestagung.

Die erschienenen Arbeitsberichte im Überblick

| Nr. | Autor(en): Titel |
|---|---|
| 1 | Edin Basic, Ulrich Rendtel: The use of the German Microcensus as a tool for longitudinal data analysis: Methods for the control the effects of the non-coverage of residential mobility. |
| 2 | Michael Konold: Ausmaß und Folgen unbeobachteter Übergänge im Rahmen des Mikrozensus-Panels: Ergebnisse empirischer Analysen. |
| 3 | Edin Basic, Ivo Marek, Ulrich Rendtel: The German Microcensus as a tool for longitudinal data analysis: An evaluation using SOEP data. |
| 4 | Robert Herter-Eschweiler: Der Mikrozensus als Panel: Längsschnittverknüpfung und Selektivitätsanalysen im Bereich der Art der Erwerbsbeteiligung und familialen Lebensformen. |
| 5 | Edin Basic, Ivo Marek, Ulrich Rendtel: The German Microcensus as a tool for longitudinal data analysis: An evaluation using SOEP data. Erschienen in: Schmoller's Jahrbuch - Journal of Applied Social Science Studies, Vol. 125, Number 1, 2005. |
| 6 | Sandra Rohloff: Das Hochrechnungsverfahren für Längsschnittauswertungen aus dem Mikrozensus. |
| 7 | Ulrich Rendtel: Wie geeignet ist der Mikrozensus fr Lngsschnittanalysen. |
| 8 | Edin Basic, Ulrich Rendtel: Estimation strategies in the presence of non-coverage in the German Microcensus-Panel: An evaluation using SOEP data. |
| 9 | Edin Basic: Stabilität von Ergebnissen bei unterschiedlichen Arbeitsmarktabgrenzungen. |
| 10 | Ivo Marek: Weighting adjustments in the presence of non-coverage due to residential mobility in the German Microcensus-Panel. |
| 11 | Heike Wirth: Anonymisierung des Mikrozensuspanels im Kontext der Bereitstellung als Scientific-Use-File |
| 12 | Bernhard Schimpl-Neimanns: Filekonzept zum Mikrozensus |
| 13 | Bernhard Schimpl-Neimanns: Berufliche Ausbildungsverläufe bis zum Übergang ins Erwerbsleben – Analysen zur Stichprobenselektivität des Mikrozensuspanels 1996-1999 |
| 14 | Bernhard Schimpl-Neimanns: Zur Datenqualität der Bildungsangaben im Mikrozensus |


Statistisches Bundesamt
65180 Wiesbaden
Zweigstelle Bonn
Graurheindorfer Str. 198
52117 Bonn

Freie Universität Berlin
Fachbereich Wirtschaftswissenschaften
Garystr. 21
14195 Berlin

Landesamt für Datenverarbeitung und Statistik
Nordrhein-Westfalen
Postfach 101105
40002 Düsseldorf

Zentrum für Umfragen, Methoden und Analysen
Postfach 122155
68072 Mannheim