Faktische Anonymisierung ILO-Piloterhebung

Inhaltsverzeichnis

0 Einleitung

Die ILO-Telefonerhebung des Statistischen Bundesamtes ist eine als rotierendes Panel konzipierte und ausschließlich telefonisch durchgeführte Stichprobenerhebung zur Erwerbsbeteiligung der Erwerbsbevölkerung in (15 bis 74 Jahre) Privathaushalten. Um die internationale Vergleichbarkeit der Daten zu gewährleisten, orientiert sich die Erhebung bei der Messung des Erwerbsstatus am Labour-Force-Konzept der Internationalen Arbeitsorganisation (ILO). Die Pilotphase der Telefonerhebung mit einem Stichprobenumfang von 10.000 realisierten Interviews pro Monat lief über 18 Wellen von April 2003 bis einschließlich September 2004.

Die Zulässigkeit der Übermittlung von Mikrodaten aus der amtlichen Statistik an die Wissenschaft ist in § 16 Abs. 6 des Bundesstatistikgesetzes (BStatG) geregelt. Danach dürfen Mikrodaten an die Wissenschaft übermittelt werden, wenn die Einzelangaben nur mit unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft eindeutig und richtig zugeordnet werden können (Konzept der faktischen Anonymität). Das BStatG verlangt demnach keine absolute Anonymität. Faktische Anonymität bedeutet, dass eine Re-Identifikation von Einzelangaben nicht grundsätzlich ausgeschlossen werden kann, aber der Aufwand einer Re-Identifikation den Nutzen übersteigt.

Bei der faktischen Anonymisierung von Mikrodaten muss auf der einen Seite der gesetzlichen Auflage eines möglichst weitgehenden Schutzes vor Re-Identifikationen Rechnung getragen werden. Auf der anderen Seite soll aber ein möglichst umfassendes Analysepotenzial der Daten gewährleistet sein. Diese beiden Anforderungen in Einklang zu bringen, ist die zentrale Herausforderung von Projekten zur faktischen Anonymisierung von Mikrodaten der amtlichen Statistik.

Bereits in der Konzeptionsphase für die ILO-Telefonerhebung war vereinbart worden, dass die Mikrodaten der ILO-Piloterhebung der Wissenschaft als faktisch anonymisierter Mikrodatensatz zur Verfügung gestellt werden sollen. In Zusammenarbeit mit dem Forschungsdatenzentrum des Statistischen Bundesamtes (FDZ) sowie dem „Zentrum für Umfragen, Methoden und Analysen“ (ZUMA) (Mannheim) wurde ein Anonymisierungskonzept entwickelt mit dem Ziel, sowohl den Ansprüchen des BStatG als auch den Wünschen der externen Wissenschaftler hinsichtlich eines möglichst großen Analysepotenzials der Daten gerecht zu werden. Dieses Konzept wird im Folgenden vorgestellt.

1 Messkonzept

Die Messung des Erwerbsstatus erfolgt nach den Kriterien des Labour-Force-Konzeptes der Internationalen Arbeitsorganisation (ILO).

Das Labour-Force-Konzept unterscheidet zwischen Erwerbstätigen, Erwerbslosen und Nichterwerbspersonen.

Abbildung 1: Das Labour-Force-Konzept der ILO

Erwerbspersonen		**Nichterwerbspersonen**
Erwerbstätige • Personen ab 15 Jahre *und* • in einem Arbeitsverhältnis mit mindestens einer Stunde je Woche geleisteter Arbeitszeit *oder* • Selbständige oder Freiberufler *oder* • Soldaten/Zivildienstleistende oder • unbezahlt mithelfende Familienangehörige *oder* • Auszubildende	**Erwerbslose** • Personen ab 15 Jahren *und* • ohne Beschäftigungsverhältnis bzw. nicht selbständig und nicht freiberuflich tätig *und* • gegenwärtig für eine Beschäftigung verfügbar *und* • aktiv Arbeit suchend	• weder erwerbstätig noch erwerbslos
Erwerbstätige	**Nichterwerbstätige**	

2 Auswahlverfahren, Auswahlwahrscheinlichkeit und Auswahlsatz

2.1 Auswahlverfahren

Die Ziehung der Stichprobe erfolgt nach dem Gabler-Häder-Verfahren. Die Basis für die Erzeugung des Auswahlrahmens bildet das Verzeichnis der Festnetzrufnummern der Deutschen Telekom AG. Im nächsten Schritt werden von jeder der in diesem Verzeichnis vorhandenen Rufnummern die beiden letzten Ziffern entfernt und der so entstandene „Rufnummern-Stamm“ mit allen zweistelligen Ziffern von „00“ bis „99“ ergänzt.

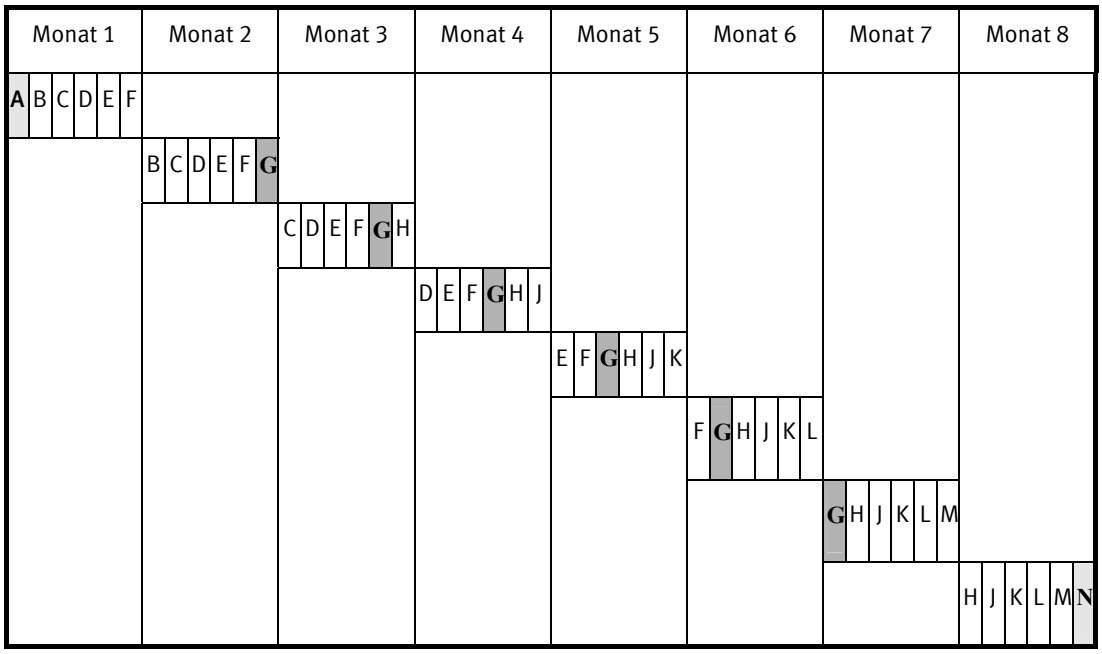
Abbildung 2: Das Grundprinzip des Gabler-Häder-Verfahrens

Auf diese Weise entsteht aus jeder eingetragenen Rufnummer ein so genannter „Block“ von 100 aufeinander folgenden potenziellen Festnetztelefonnummern. Die Menge aller erzeugten Blöcke – reduziert um die Blöcke mit identischen Nummernstämmen sowie erkennbare geschäftliche Telefonnummern, FAX-Anschlüsse etc. – bildet die Auswahlgrundlage, aus der anschließend zufällig einzelne Nummern gezogen werden. Über die Vorwahlnummern können Regionalinformationen entnommen werden, sodass eine regionale Schichtung möglich wird.

2.2 Rotierendes Panel mit sechsmaliger Befragung

Zentrales Ziel der Erhebung sind neben der Ermittlung der absoluten Zahl der Erwerbslosen und der Erwerbslosenquote auch deren monatliche Veränderungen.

Abbildung 3: Rotierendes Panel mit sechsmaliger Befragung



Um die monatlichen Veränderungen mit einem möglichst geringen Standardfehler messen zu können, wurde die Erhebung als rotierendes Panel mit sechsmaliger Befragung und einer monatlichen Nettoüberlappung von mindestens 70% konzipiert.

2.3 Auswahlwahrscheinlichkeiten

Die Wahrscheinlichkeit, in die Stichprobe der Piloterhebung zum ILO-Erwerbsstatus zu gelangen, ist auf Grund des geringen Stichprobenumfangs sowie des Panelcharakters (rotierendes Panel) der Erhebung sehr gering.

Tabelle 1: Auswahlwahrscheinlichkeiten für Erstbefragte

Welle	Jahr	Monat	Erstbefragte	Bevölkerung 15-74 Jahre	Wahrscheinlichkeit für die Auswahl	Auswahlsatz in % der Bevölkerung 15 74 Jahre
1	2003	April	10.000	63.666.007	0,000157	0,015707
2		Mai	1.736	63.672.602	0,000027	0,002726
3		Juni	2.363	63.680.603	0,000037	0,003711
4		Juli	2.440	63.685.279	0,000038	0,003831
5		August	2.026	63.680.795	0,000032	0,003181
6		September	2.410	63.691.919	0,000038	0,003784
7		Oktober	2.320	63.699.982	0,000036	0,003642
8		November	1.915	63.697.382	0,000030	0,003006
9		Dezember	2.370	63.689.192	0,000037	0,003721
10	2004	Januar	2.249	63.674.933	0,000035	0,003532
11		Februar	2.110	63.664.446	0,000033	0,003314
12		März	2.367	63.655.376	0,000037	0,003718
13		April	2.393	63.651.588	0,000038	0,003760
14		Mai	2.214	63.656.269	0,000035	0,003478
15		Juni	2.476	63.661.889	0,000039	0,003889
16		Juli	2.403	63.661.889	0,000038	0,003775
17		August	2.386	63.661.889	0,000037	0,003748
18		September	2.245	63.661.889	0,000035	0,003526

Die größte Auswahlwahrscheinlichkeit besteht am Beginn der Panelerhebung in Welle 1 (April 2003). Hier mussten für die Ausgangsstichprobe zunächst 10.000 Erstbefragte rekrutiert werden, wodurch sich eine Wahrscheinlichkeit von $\pi = 0{,}000157$ ergab. Der Auswahlsatz beträgt damit 0,015707% der Bevölkerung im Alter von 15 bis 74 Jahren. In den Folgewellen werden jeweils das Rotationssechstel sowie die Ausfälle neu gezogen. Hierdurch ergeben sich deutlich geringere Auswahlwahrscheinlichkeiten.

2.4 Einstieg in die Wiederholungsinterviews – Teilgruppen A - G

Da die Erhebung als Panel mit sechsmaliger Befragung konzipiert ist, wird besonderer Wert darauf gelegt, die zeitliche Inanspruchnahme der Befragungspersonen möglichst gering zu halten.

Hierzu wurde ein Erhebungsinstrument entwickelt, das zwischen Erst- und Wiederholungsinterview unterscheidet und systematisch Informationen aus dem vorangegangenen Interview nutzt. Bei den Wiederholungsbefragungen wird am Beginn des Interviews geprüft, ob sich der Erwerbsstatus seit der letzten Befragung geändert hat. Je nachdem, ob sich der Status geändert hat oder nicht, steigen die Wiederholungsbefragten an jeweils definierten Stellen im Fragebogen ein.

Es wurden folgende Teilgruppen gebildet:

- „Normal-Erwerbstätige“ (Teilgruppe A): Befragte, die im Vormonat einer Erwerbstätigkeit im Sinne des Labour-Force-Konzeptes nachgegangen sind.
- „Unterbrecher“ (Teilgruppe B): Befragte, die im Vormonat angegeben haben, ihre Erwerbstätigkeit für einen Zeitraum von weniger als drei Monaten unterbrochen zu haben und deshalb nach den Kriterien des Labour-Force-Konzeptes als Erwerbstätige gelten.
- „Mithelfende“ (Teilgruppe C): Befragte, die im Vormonat angegeben haben, regelmäßig oder gelegentlich einer Tätigkeit als mithelfendes Familienmitglied nachgegangen zu sein.
- „Hinzuverdiener“ (Teilgruppe D): Befragte, die im Vormonat angegeben haben, regelmäßig oder gelegentlich einem Hinzuverdienst nachgegangen zu sein.
- „Nicht-Interessierte unter 65 Jahren“ (Teilgruppe E): Befragte unter 65 Jahren, die im Vormonat angegeben haben, nicht erwerbstätig zu sein und auch in den kommenden Monaten nicht an der Aufnahme einer Erwerbstätigkeit interessiert sind.
- „Nicht-Interessierte ab 65 Jahren“ (Teilgruppe F): Befragte ab 65 Jahre, die im Vormonat angegeben haben, nicht erwerbstätig zu sein und auch in den kommenden Monaten nicht an der Aufnahme einer Erwerbstätigkeit interessiert sind.
- „Residualkategorie“ (Teilgruppe G): Befragte, die nicht den Teilgruppe A – F zuordenbar sind.

2.5 Fortgeschriebene Fälle

Befragte der Teilgruppe F (s.o.) werden grundsätzlich nur einmal interviewt. Ihre Daten werden für die folgenden 5 Erhebungswellen fortgeschrieben.

2.6 Grundzüge des Hochrechnungsverfahrens

Die Hochrechnung dient dem Ziel, von der Stichprobe auf die Grundgesamtheit zu schließen. Um einerseits die durch Antwortverweigerungen entstehenden systematischen Verzerrungen abzuschwächen und andererseits den Standardfehler zu vermindern, werden die Ergebnisse der ILO-Telefonerhebung an bekannte Totalwerte zusätzlich erhobener Merkmale (Eckwerte) angepasst. Kombinationen der folgenden Merkmale wurden für die Eckwerteanpassung ausgewählt: Alter, Bundesland (zum Teil klassiert nach West/Ost oder so genannten Nielsengebieten), Geschlecht, Staatsangehörigkeit, Berufs- und Schulabschluss sowie Arbeitslosmel-

dung bei einer Arbeitsagentur oder einer diese Aufgabe wahrnehmenden Kommune (ja/nein). Für diese Merkmale existieren Ergebnisse aus der Statistik der Bundesagentur für Arbeit (registrierte Arbeitslose), dem Mikrozensus (Schul- und Berufsabschluss) und der laufenden Bevölkerungsfortschreibung (übrige Merkmale). Für die Panelfälle wird als zusätzlicher Eckwert der (geschätzte) Erwerbsstatus des Vormonats verwendet, um insbesondere den Zufallsfehler für monatliche Veränderungen zu reduzieren.

3 Komponenten der Anonymisierung im Überblick

Das Re-Identifikationsrisiko ist auf Grund des sehr geringen Stichprobenumfangs und der daraus resultierenden kleinen Auswahlwahrscheinlichkeit als gering zu bezeichnen. Der Auswahlsatz liegt noch weit unter dem des Mikrozensus (Auswahlsatz etwa 1% der Bevölkerung!). Außerdem liegen in der Wissenschaft keine personenbezogenen und mit dem Datensatz kompatiblen Zusatzinformationen vor, die geeignet wären, gewinnbringende Re-Identifikationsversuche (Angriffsversuche) durchzuführen. Aus diesen Gründen wurden im Rahmen der Anonymisierung keine Re-Identifikationsversuche (Angriffsszenarien) simuliert.

Komponenten bzw. Maßnahmen zum Erreichen der faktischen Anonymität:

- *Rechtliche Maßnahmen*: Alle Personen, die faktisch anonymisierte Mikrodaten erhalten, werden entsprechend dem Verpflichtungsgesetz vom 2. März 1974 (BGBl. I S. 469, 547) das durch Gesetz vom 15. August 1974 (BGBl. I S. 1942) geändert worden ist, verpflichtet. Darüber hinaus müssen die Datennutzer einen Datennutzungsvertrag mit dem Statistischen Bundesamt abschließen. Die im Nutzungsvertrag gestellten Bedingungen entsprechen denen anderer Datennutzungsverträge für faktisch anonymisierte Datensätze der amtlichen Statistik (z. B. Mikrozensus, Zeitbudgeterhebung, Einkommens- und Verbrauchsstichprobe). Wesentliche Inhalte des Datennutzungsvertrags sind:
  - Benennung des Forschungsvorhabens, für welches die Daten ausschließlich verwendet werden.
  - Verbot von Deanonymisierungsmaßnahmen. Im Falle einer (unbeabsichtigten) Deanonymisierung ist das Statistische Bundesamt unmittelbar und unverzüglich zu unterrichten.
  - Veröffentlichungen, die auf den faktisch anonymisierten Daten beruhen, sind dem Statistischen Bundesamt zu melden, Belegexemplare sind abzuliefern.

- Die faktisch anonymisierten Daten dürfen vom Datennutzer nur Personen zugänglich gemacht werden, die mit dem angegebenen Forschungsvorhaben betraut sind. Diese Personen müssen Amtsträger, für den öffentlichen Dienst besonders Verpflichtete oder Verpflichtete nach dem BStatG sein.
- Die Daten müssen vom Datennutzer zu dem vertraglich vereinbarten Zeitpunkt gelöscht werden.

- *Technische Maßnahmen/Speichermedium*: Die Daten werden ausschließlich auf dem Speichermedium CD/ROM vertrieben. Damit ist gewährleistet, dass der Datensatz auf der CD/ROM von Seiten der Datennutzer nicht verändert werden kann. Außerdem wird die CD/ROM mit einem Passwort versehen.
- *Datenseitige Maßnahmen*: Durch datenseitige Maßnahmen der Anonymisierung wird der Informationsgehalt der Daten reduziert.

4 Datenseitige Anonymisierungsmaßnahmen

4.1 Ziehung einer 95%-Zufallsstichprobe

An die Datennutzer wird eine 95%-Zufallsstichprobe aus dem Gesamtdatensatz weitergegeben. Die Weitergabe einer Stichprobe der Mikrodaten aus dem kompletten Datenmaterial an die Nutzer stellt einen sehr wichtigen Schutz vor Re-Identifikationsversuchen dar. Der Datenangreifer kann sich nicht sicher sein, dass ein spezieller Mikrodatensatz tatsächlich im Datensatz enthalten ist, auch wenn er weiß, dass die Person im Rahmen der Erhebung befragt wurde.

Während der Laufzeit der Erhebung wurden 48.423 Personen mindestens einmal befragt. Aus diesem Personenkreis wurde eine Stichprobe von 95% gezogen, sodass der Scientific-Use-File Informationen zu 46.004 Personen enthält. Die Zufallsauswahl wurde geschichtet nach der Verweildauer im Panel (1 bis max. 6 Wellen) gezogen (Tabelle 2).

Die Zufallsauswahl wurde mit Hilfe des Statistikprogramms SAS (Prozedur Surveyselect, Startzahl=9999, Methode=SRS) gezogen.46.004 Personen (95% aus 48.423 Personen) geschichtet nach der Dauer der Panelteilnahme (1 bis maximal 6 Wellen) ausgewählt (siehe Tabelle 1).

Tabelle 2: Vergleich Stichprobenstrukturen hinsichtlich Dauer der Panelteilnahme

Teilnahmedauer (Wellen)	Personen Originaldatensatz		Personen Faktisch anonymisierter Datensatz	
	Absolut	%	Absolut	%
1	11.313	23,4	10.748	23,4
2	6.312	13,0	5.997	13,0
3	5.165	10,7	4.907	10,7
4	4.247	8,8	4.035	8,8
5	3.720	7,7	3.534	7,7
6	17.666	36,4	16.783	36,4
Summe	48.423	100,0	46.004	100,0

Die Quer- und Längsschnitthochrechnungsfaktoren wurden neu gerechnet und an den 95%-Datensatz gespielt.

Um die Qualität der 95%-Stichprobe beurteilen zu können, wurden für die zentrale Variable (Erwerbsstatus nach dem Labour-Force-Konzept) Vergleichsrechnungen durchgeführt. Dazu wurde über alle 18 Wellen hinweg die Ergebnisse aus der 95%-Stichprobe mit denen aus der 100%-Stichprobe verglichen und die prozentualen Abweichungen zwischen den Ergebnissen festgestellt (Tabelle 3).

Die Vergleichsrechnungen zeigen, dass die aus dem 95%-Datensatz resultierenden Abweichungen in den Schätzergebnissen für die Variable Erwerbsstatus in einem vertretbaren Rahmen liegen.

Tabelle 3: Prozentuale Abweichung der Originalergebnisse und der Ergebnisse aus der 95%-Zufallsstichprobe in Prozent

Erwerbsstatus/Quoten	Durchschnitt über alle 18 Wellen	Minimum	Maximum
Erwerbslose	1,03	0,02	2,44
Erwerbstätige	0,23	0,01	0,60
Nichterwerbspersonen	0,45	0,08	1,56
Bevölkerung im Alter von 15-74 Jahre	0,00	0,00	0,01
Erwerbslosenquote	0,62	0,01	1,87
Erwerbstätigenquote	0,24	0,04	0,40

In absoluten Zahlen liegt die größte Abweichung für die Schätzung Zahl der Erwerbslosen bei -99,9 Tsd., die geringste bei -5,9 Tsd. Die größte Abweichung für die geschätzte Zahl der Erwerbstätigen liegt bei 239,0 Tsd., die geringste bei -4,5 Tsd. Für die Nichterwerbspersonen differieren die Schätzungen maximal mit -316,4 Tsd. und minimal mit -17,4 Tsd. Für die Bevölkerung im Alter zwischen 15 und 74 Jahren liegen die Werte bei 3,5 Tsd. (Maximum) und 2,8 Tsd. (Minimum).

4.2 Vergröberungen

4.2.1 Vergröberung von Regionalangaben

Als Regionalangabe wird eine Variable „Alte/Neue Bundesländer“ (v003) geliefert. Berlin wird zu den neuen Bundesländern gezählt. Alle anderen Regionalangaben wurden gelöscht.

4.2.2 Vergröberung des ISCO

Die Berufsangabe wird im Interview bzw. in einem anschließenden Verfahren nach der 4-stelligen Internationalen Standardklassifikation der Berufe (International Standard Classification of Occupations, ISCO) verschlüsselt. Für den Scientific-Use-File wurden die 4-stelligen Angaben auf 3-stellige Angaben (Ebene der Untergruppen) umcodiert.

5 Fazit

Mit dem vorliegenden Anonymisierungskonzept ist es möglich geworden, die Daten aus der Piloterhebung zum ILO-Erwerbsstatus der Bevölkerung als faktisch anonymisierten Mikrodatensatz der Wissenschaft für Analysezwecke zu geringen Kosten zur Verfügung zu stellen. Trotz des Anspruchs, den Analysegehalt der Daten möglichst weitgehend zu erhalten, konnten Vergröberungen der Ausgangsdaten aus Geheimhaltungsgründen nicht gänzlich vermieden werden. Dies hat zur Folge, dass die Schätzungen aus dem faktisch anonymisierten Datensatz geringfügig weniger exakt sind als die aus dem Originaldatensatz. Dennoch kann davon ausgegangen werden, dass der faktisch anonymisierte Mikrodatensatz eine Vielzahl von Auswertungsmöglichkeiten eröffnet.