Stichprobenziehung für Migrantenpopulationen in fünf Ländern: eine Darstellung des methodischen Vorgehens im PIONEUR-Projekt

Santacreu Fernandez, Oscar; Rother, Nina; Braun, Michael

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Stichprobenziehung für Migrantenpopulationen in fünf Ländern

Eine Darstellung des methodischen Vorgehens im PIONEUR-Projekt

Sampling of Migrant Populations in Five Countries

The Approach of the PIONEUR Project

Oscar Santacreu Fernández, Nina Rother & Michael Braun[1]

Dieser Artikel stellt eine innovative Form der Stichprobenziehung für Migrantenpopulationen dar, die bei der Durchführung einer Telefonumfrage des von der Europächen Kommission im 5. Rahmenprogramm geförderten Projekts PIONEUR („Pioneers of Europe's Integration ‚from below': Mobility and the emergence of European Identity among National and Foreign Citizens in the EU") zur Anwendung kam. Das Grundprinzip besteht aus der Bestimmung der häufigsten Namen für die einzelnen Migrantengruppen über eine statistische Analyse der Telefonbücher der Herkunftsländer und der anschließenden Identifikation dieser Gruppen in den Telefonbüchern der Zielländer. Die Qualität der resultierenden Nettostichprobe wird für Deutschland im Vergleich zu Daten des Mikrozensus 2004 evaluiert.

This article describes the innovative way of sampling migrant populations that was used for a telephone survey of the PIONEUR project („Pioneers of Europe's Integration ‚from below': Mobility and the emergence of European Identity among National and Foreign Citizens in the EU"), funded by the European Commission in the 5th Framework Programme. The basic principle consists of determining the most frequent names for the different migrant groups by a statistical analysis of the telephone directories of the countries of origin and a subsequent identification of these groups in the telephone directories of the countries of residence. The quality of the resulting net sample for Germany is evaluated by a comparison with data of the Microcensus 2004.

1 Wir bedanken uns bei zwei anonymen Gutachtern der ZUMA-Nachrichten für ihre wertvollen Kommentare.

1 Einleitung

Das PIONEUR-Projekt[2] basiert auf einem *mixed-methods-Ansatz*, der qualitative und quantitative Erhebungs- und Auswertungsverfahren miteinander kombiniert und auf deren Grundlage drei unterschiedliche Gruppen von Migranten bzw. Nicht-Migranten miteinander verglichen werden: *Stayers* (EU-Bürger, die in einem Mitgliedstaat der EU wohnen, deren Staatsbürgerschaft sie auch besitzen), *internal movers* (EU-Bürger, die in einem Mitgliedstaat der EU wohnen, deren Staatsbürgerschaft sie nicht besitzen) und *external movers* (Nicht-EU-Bürger aus EU-Anwärterstaaten, die aber in einem Mitgliedstaat der EU wohnen). Kernstück des PIONEUR-Projekts ist der im Sommer 2004 durchgeführte *European Internal Movers' Social Survey (EIMSS)*. Er befragte *internal movers*, d.h. konkret diejenigen Briten, Deutsche, Franzosen, Italiener und Spanier, die von 1974 bis 2003 als Erwachsene in eines der anderen vier Länder gezogen sind und die zum Befragungszeitpunkt bereits mindestens ein Jahr dort lebten. Die Ziele dieser Umfrage waren unter anderem die Erforschung der Voraussetzungen und Motive, aber auch der Barrieren für eine Migration innerhalb der EU, der Auswirkungen der EU-internen Migration auf die Lebensqualität und die individuellen Erwartungen der Migranten sowie auf die Einstellungen gegenüber Institutionen der EU und der Identifikation mit Europa.

Der EIMSS sollte auf einer Zufallsstichprobe beruhen, wobei auch Migranten, die bereits die Staatsbürgerschaft des Aufenthaltslandes erworben hatten, zur Gruppe der *internal movers* gezählt wurden. In jedem der fünf beteiligten Länder wurden je 250 *internal mover* aus den je vier anderen Ländern, also insgesamt 5000 EU-Migranten, anhand eines standardisierten Fragebogens telefonisch befragt. Die Interviews wurden anhand eines mehrsprachigen CATI-Fragebogens durch bilinguale Interviewer durchgeführt. Die Feldphase begann Anfang Mai 2004 und sollte bis zur Europawahl im Juni 2004 abgeschlossen sein. Aufgrund von Feldproblemen konnte dieses Ziel in einigen Ländern nicht erreicht werden, in Großbritannien konnte das Feld erst Ende 2004 abgeschlossen werden.

Die besondere Schwierigkeit bei der Durchführung des EIMSS war die Stichprobenziehung. Zur Generierung einer Stichprobe wurde eine Reihe alternativer Methoden diskutiert, die sich aber alle als nicht durchführbar erwiesen: Zunächst war die Benutzung einer

2 Das Projekt PIONEUR (s. Rother 2005) wurde von der Europäischen Kommission im 5. Rahmenprojekt gefördert. Die internationale Projektleitung des PIONEUR-Projekts wurde von Ettore Recchi am Centro Interuniversitario di Sociologia Politica (CIUSPO) der Universität Florenz übernommen. Die verantwortlichen nationalen Projektleiter sind in Frankreich Anne Muxel (Centre d'Etude de la Vie Politique Française (CEVIPOF)), in Großbritannien Damian Tambini (Centre for Socio-Legal Studies (CSLS) der Universität Oxford), in Spanien Antonio Alaminos (Observatorio Europeo de Tendencias Sociales (OBETS) der Universität Alicante) und in Deutschland Michael Braun (ZUMA).

Registerstichprobe favorisiert worden, da die Qualität der auf dieser Basis erhobenen Daten in der Regel als sehr hoch angesehen werden kann. Da es für Großbritannien aber kein Ausländerregister gibt und die Register der anderen Länder nur die Staatsangehörigkeit, nicht aber das Geburtsland verzeichnen, war diese Art der Stichprobenziehung nicht möglich. Die Durchführung eines *snowball samplings* – ein Verfahren, das in vielen Migrantenbefragungen verwendet wird – wäre zwar in allen Ländern kostengünstig durchführbar gewesen, allerdings weisen so erhaltene Stichproben große Repräsentativitätsprobleme auf, so dass diese Art der Stichprobenziehung nicht ausschließlich verwendet werden konnte. Da die Umfrage ohnehin als Telefonumfrage durchgeführt werden sollte, bot sich die Orientierung an Telefonbüchern als Grundlage zur Stichprobenziehung an, auch weil ein *random digit dialing* aufgrund der geringen Größe der Zielpopulation zu viele Screening-Interviews erforderlich gemacht hätte. Einen Erfolg versprechenden Ansatz für die Stichprobenziehung stellen die Arbeiten von Humpert und Schneiderheinze (2002) dar. Mit Hilfe der Namensforschung (Onomastik) und unter Verwendung von Namensverzeichnissen wird dabei eine Zuordnung von Telefonbucheinträgen zu einer bestimmten Nationalität vorgenommen (vgl. Humpert & Schneiderheinze 2000). Die bereits vorliegende Datenbank hätte für Migranten nach Deutschland eine hohe Trefferwahrscheinlichkeit und eine niedrige Anzahl an benötigten Screeing-Interviews bedeutet. Allerdings hätten für die anderen Länder die Auswahlgrundlagen noch erstellt werden müssen, was in dem zeitlichen Rahmen nur schwer durchführbar gewesen wäre.

Vor dem Hintergrund dieser Probleme erschien es dem PIONEUR-Team am sinnvollsten, eine neue Methode zu entwickeln, die die kostengünstige Ziehung einer zufälligen und vergleichbaren Stichprobe in allen fünf Ländern erlaubte. Dabei sollte, wie bei dem Ansatz von Humpert und Schneiderheinze auch, eine Orientierung an linguistischen Gesichtspunkten erfolgen und die Telefonbücher der jeweiligen Länder als Datengrundlage benutzt werden. Im Folgenden wird nun das Vorgehen zur Stichprobenziehung genauer dargestellt, das auf der Idee, dem Algorithmus und der Software von Santacreu (vgl. Santacreu 2005) beruht.

2 Vorgehen

Um eine vergleichbare und qualitativ hochwertige Stichprobe zu erhalten, ist es zunächst notwendig, diejenigen Telefonbucheinträge zu identifizieren, die mit einer hohen Wahrscheinlichkeit der Zielnationalität entsprechen. Ausgehend von den verfügbaren Telefonbucheinträgen, die allerdings nur Festnetzanschlüsse umfassten, waren die einzigen nutzbaren Kriterien zur Identifikation des Herkunftslandes der Vor- und der Nachname (oder die Nachnamen). Dies bedeutet, dass ein Kriterium gefunden werden musste, das es erlaubte, einem bestimmten Vor- oder Nachnamen eine Wahrscheinlichkeit für die Zugehö-

rigkeit zu einer bestimmten Nationalität zuzuweisen.

Folgende Schritte waren bei der Stichprobenziehung zu befolgen:

1. Aufbereitung der Gesamtheit aller Telefonbucheinträge in Deutschland, Frankreich, Großbritannien, Italien und Spanien in ein Format, das eine statistische Auswertung erlaubt
2. Analyse der Häufigkeit der linguistischen Einheiten (Vor- und Nachnamen) in jedem Land und Berechnung der Auftretenswahrscheinlichkeit einer jeden linguistischen Einheit
3. Zuweisung einer Wahrscheinlichkeit zu jedem Telefonbucheintrag in Abhängigkeit von den Auftretenswahrscheinlichkeiten seiner jeweiligen linguistischen Einheiten
4. Ziehung der Stichprobe
5. Revision und Fertigstellung der Kontaktliste.

Diese Schritte werden nun im Einzelnen beschrieben.

2.1 Aufbereitung der Telefonbucheinträge

Nach dem Kauf einer DVD, die die Telefonbücher der fünf Länder beinhaltete, wurden die fünf Datensätze aufbereitet. Zunächst wurden nur diejenigen Einträge berücksichtigt, die Einzelpersonen oder Familien entsprechen, Einträge von Firmen wurden nicht berücksichtigt. Tabelle 1 gibt eine Übersicht über die verfügbaren Einträge für jedes Land, die Zahl der tatsächlich existierenden Haushalte und der Relation zwischen Telefonbucheinträgen und Zahl der Haushalte in den einzelnen Ländern. Es ist festzustellen, dass in Großbritannien die Zahl der Telefonbucheinträge im Vergleich zur Zahl der Haushalte relativ niedrig ist, was an den Besonderheiten des Telekommunikationsmarkts in diesem Land liegt, auf die weiter unten noch genauer eingegangen werden wird.

Tabelle 1 Telefonbucheinträge und Zahl der Haushalte in den einzelnen Ländern

Land	Einträge[1]	Haushalte (2002)[2]	Anteil
Frankreich	20.473.368	24.787.083	82,6%
Italien	16.941.708	22.105.385	76,6%
Spanien	12.411.976	13.733.333	90,4%
Großbritannien	11.451.423	25.754.087	44,5%
Deutschland	30.590.871	37.491.818	81,6%

Quelle: [1] Telefonbuch-DVD, [2] Eurostat, 2005

Anschließend wurde die Konfiguration des Felds 'Name' in jedem Land analysiert. Dieses Feld beinhaltete zum Beispiel im spanischen Fall zwei Nachnamen und den ersten Buchstaben des Vornamens, gefolgt von einem Punkt. Im französischen Fall bestanden die meisten Einträge aus einer Anrede (Mr, Mme), gefolgt von Nach- und Vornamen. Viele Einträge bezogen sich auf Ehepaare und beinhalteten den Nachnamen des Ehepaars, mit der vorausgehenden Anrede ohne die Vornamen. Im deutschen Fall bestand der Eintrag 'Name' hauptsächlich aus einem Nachnamen und einem Vornamen, obwohl in einigen Fällen zwei Einzelpersonen unter dem gleichen Eintrag erschienen. Im italienischen Fall war die häufigste Konfiguration die eines Nachnamens, gefolgt von einem oder zwei Vornamen. Die häufigste Kombination im britischen Telefonbuch war die eines Nachnamens, gefolgt von dem Vornamen und in einigen Fällen der Abkürzung des zweiten Vornamens.

2.2 Analyse der Häufigkeiten und Auftretenswahrscheinlichkeiten der Vor- und Nachnamen

Im zweiten Schritt wurden nun die Telefonbucheinträge einer statistischen Analyse unterzogen. Im Hinblick auf die länderspezifischen Besonderheiten des Felds ‚Name' wurden für jedes Land Filter-Algorithmen entwickelt, mittels derer aus jedem Eintrag zwei linguistische Einheiten extrahiert werden konnten, je nach Land zwei Nachnamen oder ein Nachname und ein Vorname. Diese linguistischen Einheiten wurden pro Land in einer Liste geordnet, aus der dann die jeweiligen Auftretenshäufigkeiten berechnet wurden.

Als zusätzliches Kontrollinstrument wurde die geografische Verteilung der Vor- und Nachnamen in jedem Land berücksichtigt. Dafür wurden als Gruppierungsvariable die ersten beiden Ziffern der Postleitzahl herangezogen und die Listen so aufgeteilt, dass die Ergebnisse in Gruppen entsprechend der Postleitzahl organisiert waren. Auf diesem Weg entstand eine Liste der häufigsten Nachnamen pro Region, wodurch die Heterogenität der Namen in den Regionen angemessen berücksichtigt werden konnte. Im britischen Fall war es aufgrund des besonderen Postleitzahlensystems notwendig, die geografische Segmentierung direkt anhand von Ortsnamen vorzunehmen.

Auf der anderen Seite wurden durch das linguistische Screening der Namen bei der Konstruktion der Stichprobe die Gruppen der Migranten auf diejenigen reduziert, die zu der dominanten Ethnie des Herkunftslandes gehören. Daher wurden zum Beispiel deutschsprachige Minderheiten in Italien und Frankreich ausgeschlossen, wie auch frühere Migranten und deren Nachkommen, die wieder in das Ursprungsland der Vorfahren zurückgekehrt waren. Die geografische Segmentierung erlaubte es darüber hinaus, alle diejenigen Zonen auszuschließen, die Probleme bei der Identifikation von nationalen Vor- und Nachnamen verursachen könnten, wie zum Beispiel Elsass-Lothringen oder die deutschsprachigen Gebiete Norditaliens.

Nach der Programmierung des aus mehreren Modulen bestehenden Systems wurden die Listen eines jeden Landes durchlaufen. Im ersten Durchgang erhielt man für jedes Land und jede geografische Zone (z) eine Gesamtmenge Ω, deren linguistische Einheiten den Vor- und Nachnamen entsprachen:

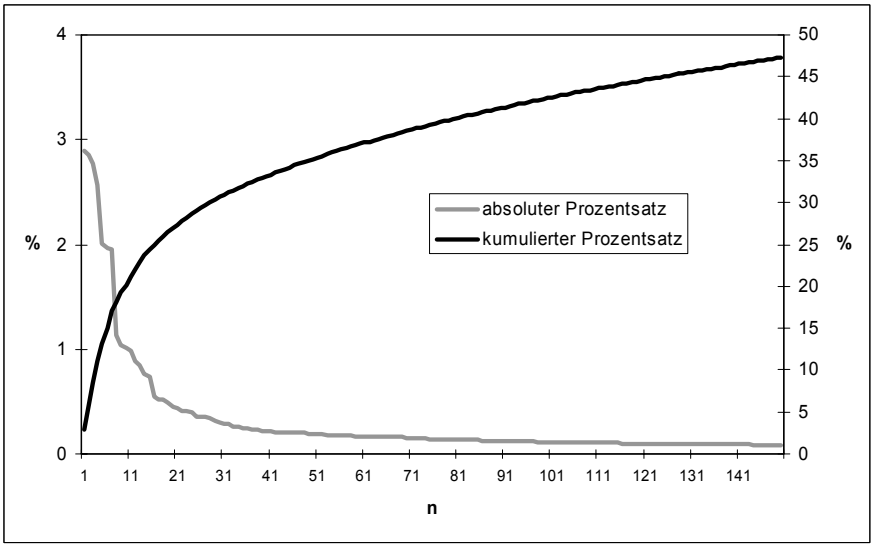
$$\Omega_z=\{\text{Einheit}1, \text{Einheit}2, \text{Einheit}3 \ldots \text{Einheit}n\}$$

Das Ziel war die Bestimmung einer Teilmenge, bestehend aus 'nationalen' linguistischen Einheiten mit der höchsten Wahrscheinlichkeit für jede geografische Zone innerhalb eines Landes:

$$A_z=\{\text{ Einheit}1, \text{Einheit}2 \ldots\}$$

Um diese Teilmenge zu bestimmen, wurden die linguistischen Einheiten der Häufigkeit entsprechend absteigend sortiert. Abbildung 1 zeigt, welchen Anteil die häufigsten Nachnamen in der Verteilung dieser linguistischen Einheiten für den konkreten Fall der Nachnamen in Alicante (Spanien) erreichen.

Abbildung 1 Absolute und kumulative Häufigkeit spanischer Nachnamen in Alicante



Quelle: eigene Berechnungen

Die absteigende Linie in der Abbildung bezieht sich auf den Prozentsatz der Population, den die häufigsten einzelnen Nachnamen ausmachen, geordnet vom häufigsten zum am wenigsten häufigen. Je mehr man sich dem rechten Ende der Verteilung nähert, desto geringer wird der Beitrag eines Namens. Auf der anderen Seite präsentiert die aufsteigende Linie den kumulativen Prozentsatz für die häufigsten Nachnamen.

Dabei wurden als Elemente der Teilmenge A_z nur diejenigen linguistischen Einheiten gewählt, die in einer geografischen Zone z über einem bestimmten Wert lagen, der dynamisch angepasst an das Ziel der Erreichung einer ausreichenden Stichprobengröße für jede geografische Zone z einzeln festgelegt wurde.

Auf der anderen Seite resultiert die Teilmenge A_p aus jedem Land aus der Vereinigung der entsprechenden Teilmengen aus jeder Zone des jeweiligen Landes. D.h. es wurden die häufigsten Vor- und Nachnamen aus jeder Zone herangezogen, um die Teilmenge der pro Land am häufigsten Vor- und Nachnamen zu konstruieren. Auf diese Art und Weise wurden Vor- und Nachnamen berücksichtigt, die in mindestens einer Zone des jeweiligen Landes sehr häufig sind.

$$A_p=\{A_{z1} \cup A_{z2} \cup A_{z3} \cup ... \cup A_{zn}\}$$

Anschließend berechnete der Algorithmus die Wahrscheinlichkeit für jedes Element der Teilmenge Ap (Vor- und Nachnamen) ausgehend von dem Konzept der relativen Häufigkeit ($fr=k/n$, wobei n die Anzahl der linguistischen Einheiten darstellt und k die Häufigkeit der jeweiligen linguistischen Einheit).

Anschließend wurden diejenigen linguistischen Einheiten mit einer höheren Wahrscheinlichkeit ausgewählt und die Liste des jeweiligen Landes an den nationalen Projekt-Partner geschickt, damit eine manuelle Revision der Namensliste stattfinden konnte. Namen, die nicht der dominanten ethnischen Gruppe des Landes entsprachen, wurden gelöscht.

2.3 Zuweisung von Wahrscheinlichkeiten zu Telefonbucheinträgen und Ziehung der Stichprobe

Das angewendete Verfahren produzierte bislang schon eine wertvolle Information: welche linguistischen Einheiten (Vor- und Nachnamen) der Nationalität eines Individuums in irgendeinem Telefonbuch entsprechen und mit welcher Wahrscheinlichkeit. Dieses Wissen erlaubte die Ziehung der Stichprobe. Daran waren folgende Elemente beteiligt:

- A_h ist die Menge aller Vor- und Nachnamen, die einem bestimmten Herkunftsland entsprechen
- A_m ist die Menge aller Vor- und Nachnamen, die allen verbleibenden Ländern außer dem Herkunftsland entsprechen

Der erste Schritt des Algorithmus war, die Menge aller Vor- und Nachnamen, die einem bestimmten Herkunftsland entsprechen (die Menge A_h), heranzuziehen und diejenigen Vor- und Nachnamen zu löschen, die ebenfalls in einem der anderen Länder auftauchten (die Menge A_m). Auf diese Weise konnte vermieden werden, dass mehrdeutigen und möglicherweise in mehr als einer Nationalität vorkommenden Vor- und Nachnamen eine Wahrscheinlichkeit zugewiesen wird.

$$A_h = A_h \setminus (A_h \cap A_m)$$

Zum Beispiel ist der Name *María* ein Indikator für eine spanische Nationalität, er wurde aber nicht als *typisch* spanischer Name berücksichtigt, da dieser Name ebenfalls in der Menge der deutschen und italienischen Vor- und Nachnamen zu finden ist.

Anschließend durchlief der Algorithmus die Telefonbuchliste jedes Landes, Eintrag für Eintrag, und arbeitete dabei die folgenden Aufgabenstellungen ab:

(a) Identifikation der verschiedenen Felder des jeweiligen Eintrags und Abspeichern eines jeden Eintrags in einer Variable.

(b) Überprüfen, ob der Eintrag einer Region entspricht, die falsche Positive beinhalten könnte (wie z.B. Elsass-Lothringen oder Südtirol) und in diesem Fall zum nächsten Eintrag vorspringen.

(c) Löschen der Anrede aus dem Feld 'Name' (z.B. Mme., Mr., Dr., etc.).

(d) Vereinfachung der in den Variablen enthaltenen Buchstaben (Löschen von Akzenten und Umlauten etc., so dass nur die Buchstaben A-Z vorkommen). Dies ist wichtig, damit beispielsweise "González" und "Gonzalez" nicht aufgrund des Akzents als unterschiedliche Wörter behandelt werden.

(e) Zur Beschleunigung des Prozesses, Zählen der Anzahl der Nachnamen als Kriterium für die Identifikation von Spaniern in anderen Ländern bzw. von Ausländern in Spanien, da nur Spanier zwei Nachnamen aufweisen.

(f) Überprüfen, ob sich der erste Nachname unter den Nachnamen aus dem Herkunftsland (A_h) befindet und in diesem Fall zum nächsten Eintrag springen. Auf diesem Weg beschleunigt sich der Prozess, da die restlichen Elemente des Felds 'Name' nicht unnötigerweise analysiert werden.

(g) Entdecken von weiteren linguistischen Einheiten aus dem Herkunftsland (Teilmenge A_h) und Berechnung der Wahrscheinlichkeit der Zugehörigkeit zum Herkunftsland für jeden dieser Einträge. Zur Berechnung dieser Wahrscheinlichkeiten werden die Wahrscheinlichkeiten aller Elemente von A_h, die sich in dem jeweili-

gen Eintrag befinden, aufsummiert. Zum Beispiel hat "Antonio Gonzales" eine höhere Wahrscheinlichkeit, spanisch zu sein als "John Gonzalez" da sich die Wahrscheinlichkeiten der linguistischen Einheiten *Gonzalez* und *Antonio* aufsummieren. Wenn sich also in einem Telefoneintrag zwei Elemente A_{h1} und A_{h2} befinden, dann wird dem jeweiligen Eintrag die folgende Wahrscheinlichkeit zugewiesen:

$$P(A_{h1} \cup A_{h2}) = P(A_{h1}) + P(A_{h2}) - P(A_{h1} \cap A_{h2})$$

(h) Vergleich der linguistischen Einheiten des Telefonbucheintrags mit den entsprechenden der anderen Länder (A_m) und Reduktion der entsprechenden Wahrscheinlichkeit dieses Eintrags. Auf diesem Weg werden die Vor- und Nachnamen, die auch zu einer anderen Nationalität gehören könnten, bestraft:

$$P(A_{h1}) = P(A_{h1}) * (1 - P(A_m))$$

(i) Wenn die Wahrscheinlichkeit des Eintrags größer ist als eine mittels eines Vortests für jedes Land vorgegebene Wahrscheinlichkeit, speichert das Programm die folgende Information des Eintrags in eine neue temporäre Datei:

- die Trefferquote des Eintrags, also die Anzahl der Elemente eines Namens, die mit hoher Wahrscheinlichkeit zu einer bestimmten Nationalität gehören
- die Wahrscheinlichkeit, zu einer bestimmten Nationalität zu gehören
- das Feld ‘Name’
- das Feld ‘Postleitzahl’
- das Feld ‘Telefonnummer’

Abbildung 2 Beispiel für einen Datensatz

```
2 0.30911  Alarcón Chavarria Vicente  Erstestr.,3   10625  Berlin     +49/30/1234567
2 0.99999  Alba-Fernandez Juan A.     Zweitestr.,21 76137  Karlsruhe  +49/721/123456
3 0.99999  Alba Garcia Jacinto.       Drittestr.,2  58644  Iserlohn   +49/2371/12345
3 0.69764  Alba Garcia Juan A.        Viertestr.,12 42653  Solingen   +49/212/12345
```

2.4 Revision und Fertigstellung der Kontaktlisten

Anschließend extrahiert das Programm eine Liste, die von der Stichprobendatei ausschließlich die Vor- und Nachnamen enthält, damit eine visuelle Revision stattfinden kann und inkorrekte Einträge nachträglich gelöscht werden können. In dieser manuellen Revision wurden durchschnittlich knapp 13% der Namen gelöscht. Diese Vorgehensweise hat

sich als außerordentlich effizient erwiesen, um eine hohe Qualität der Stichprobe zu garantieren.

In einem letzten Schritt verwendet das Programm die revidierte Datei, ordnet sie von der höchsten zur niedrigsten Trefferquote und Wahrscheinlichkeit und extrahiert pro Interviewland zwei endgültige Dateien mit einer bestimmten Anzahl an Einträgen, die mit hoher Wahrscheinlichkeit Telefonbucheinträgen der Zielnationalität angehören.

- **Datei 1** beinhaltet vier Felder: 'Trefferquote', 'Wahrscheinlichkeit', 'Telefonnummer' und 'Postleitzahl'. Die Einträge der Datei werden vor der Benutzung durch CATI-Programme zufällig geordnet.
- **Datei 2** ist eine Datei zur Kontrolle der Qualität der Stichprobe. Sie beinhaltet die auf eine Nachkommastelle beschnittene Wahrscheinlichkeit und das Feld 'Name'. Die Einträge dieser Datei werden ebenfalls zufällig sortiert, damit eine Identifikation mit dem entsprechenden Eintrag aus Datei 1 nicht möglich ist, so dass die Anonymität des Befragten garantiert ist.

Der Prozess lieferte in Form einer zufälligen Auswahl für jedes mit der Feldarbeit beauftragte Institut eine Liste mit zwischen 5000 und 10000 Kontaktdaten pro jeweils benötigter Nationalität. Diese Liste war in allen Ländern außer Großbritannien ausreichend, um die benötigten 250 Interviews pro Nationalität durchzuführen. Traf das Auswahlkriterium auf mehrere Personen im Haushalt zu, so wurde die zu befragende Person mittels der *last-birthday*-Methode ermittelt.

3 Probleme dieser Methode der Stichprobenziehung

Obwohl sich das vorgestellte Vorgehen insgesamt als sehr erfolgreich erwiesen hat, traten im Lauf der Datenerhebung drei größere Probleme auf. Diese und die von uns gewählten Lösungsmöglichkeiten werden im Folgenden kurz dargestellt.

Ein Problem, das alle Länder betrifft, ist die mangelnde Abdeckung von Migranten ohne Telefonbucheintrag. Da dies vor allem weibliche Migranten betrifft, die mit einem Mann aus dem Zielland verheiratet sind, wurde eine kleine Netzwerkstichprobe in das Design einbezogen, bei dem Befragte am Ende des Interviews nach einer Telefonnummer von Frauen, die mit einem Mann des Ziellandes verheiratet sind, gefragt wurden. Diese Prozedur ist als alternativer Stichprobenrahmen berechtigt, da alle anderen alternativen Möglichkeiten größere Probleme verursacht hätten.

In Großbritannien trat ein schwerwiegenderes Problem dahingehend auf, dass die gelieferte Liste zu viele falsche Positive produziert hatte. Allerdings scheint dieses Problem nicht

an der Extraktionsmethode, sondern an der Qualität des Telefonbuchs in diesem Land und insbesondere an den besonderen Charakteristiken des Telekommunikationsmarktes zu liegen. Die einzig mögliche Lösung war die Optimierung der Listen (weniger Nummern, aber mit höherer Wahrscheinlichkeit) und der Einsatz eines *snowball samplings*, um die Telefonnummern weiterer Zielpersonen zu erhalten. Dazu wurden alle Befragten am Ende gebeten, den Interviewern auf freiwilliger Basis Telefonnummern von weiteren, dem Profil entsprechenden Migranten zu nennen, die dann wiederum befragt werden konnten.

Die Langsamkeit des Algorithmus war das wichtigste technische Problem, da in vielen Fällen in jedem Eintrag tausende Namen miteinander verglichen werden mussten. Die Lösung wurde durch zahlreiche Revisionen und Optimierungen des Programms erreicht, beispielsweise indem die Anzahl der Nachnamen berücksichtigt wurde oder temporäre Listen für Initialen generiert wurden.

4 Überprüfung der Qualität der Stichprobenziehung

Eine Quantifizierung des aufgetretenen *non-response bias* ist unmöglich, da im Gegensatz zu Bevölkerungsumfragen nicht angenommen werden kann, dass die große Mehrheit derjenigen, die nicht erreicht werden konnten oder die die Teilnahme verweigerten, zu der Zielpopulation gehörten. Im Fall des EIMSS war die Zielstichprobe viel kleiner als die Bruttostichprobe, selbst wenn das linguistische Screening perfekt funktionieren würde. Der Grund liegt darin, dass die Zielpopulation auf diejenige Gruppe beschränkt worden war, die als Erwachsene nach 1973 in das jeweilige Land kamen und die bereits seit einem Jahr dort wohnen. Aus diesem Grund ist es wahrscheinlich, dass viele Fälle scheinbarer Nicht-Erreichbarkeit oder Verweigerung tatsächlich stichprobenneutrale Ausfälle waren. Eine Angabe von Antwortraten ist daher bei dieser Studie nicht möglich.

Eine Analyse der Qualität der Stichprobe anhand des Berichts des deutschen Feldinstituts (SUZ, Duisburg) zeigt, dass die Teilstichprobe der Italiener am hochwertigsten war. Von allen verwendeten Telefonnummern waren nur 6,0% nicht Telefonnummern von Italienern. Eine niedrigere Qualität wies allerdings die französische und britische Teilstichprobe auf. Hier war der Prozentsatz der Nummern mit falscher Nationalität höher (35,9% bzw. 27,9% aller Telefonnummern).

Auf der anderen Seite kann zur Überprüfung der Qualität der EIMSS-Stichprobe ein Vergleich mit offiziellen Daten vorgenommen werden.3 Dies ist allerdings nur für

3 Eine Designgewichtung mit der reduzierten Haushaltsgröße zum Ausgleich der unterschiedlichen Inklusionswahrscheinlichkeiten ist leider bei den EIMSS-Daten nicht möglich. Aus befra-

Deutschland möglich, wo Zahlen des Mikrozensus zum Vergleich herangezogen werden können. Benutzt wird im Folgenden die anonymisierte 70% Unterstichprobe des Mikrozensus 2004 (ZUMA-File). Da Angaben zum Migrationszeitpunkt freiwillig sind, kann nur mit einer eingeschränkten Fallzahl gerechnet werden. Insgesamt machten 19,8% derjenigen mit britischer, französischer, italienischer und spanischer Staatsangehörigkeit zum Migrationszeitpunkt keine Angabe. Allerdings ist darauf hinzuweisen, dass auch der Mikrozensus selbst bei Migrantenbefragungen kein völlig zutreffendes Bild der Realität zeichnet. Hierzu trägt sicher auch bei, dass die Befragung ausschließlich mit einem deutschsprachigen Instrument erfolgt.

Die zum Vergleich herangezogenen Variablen zur Bewertung der Qualität der EIMSS-Stichprobe sind Geschlecht, Familienstand, Alter, Alter zum Zeitpunkt der Migration, Migrationsperiode, Bildung und Erwerbstätigkeit.

4.1 Geschlecht

Tabelle 2 erlaubt einen Vergleich der Geschlechtsstruktur von EIMSS und Mikrozensus 2004. Die Tabelle präsentiert pro Nationalität drei Spalten. Die erste Spalte stellt den Prozentsatz an Männern und Frauen im EIMSS dar, während die zweite Spalte sich auf die offiziellen Daten des Mikrozensus bezieht. Die dritte Spalte bildet die Differenz zwischen der EIMSS-Stichprobe und dem Mikrozensus. Es zeigt sich, dass die Unterschiede in der Geschlechtsstruktur bis auf den Fall der Spanier in Deutschland gering sind. Hier sind Männer mit 13,9 Prozentpunkten überrepräsentiert.

Tabelle 2 Vergleich zwischen EIMSS und Mikrozensus – Geschlecht

	Nationalität											
	Französisch			**Englisch**			**Italienisch**			**Spanisch**		
	EIMSS	**MZ**	**Diff.**	**EIMSS**	**MZ**	**Diff.**	**EIMSS**	**MZ**	**Diff.**	**EIMSS**	**MZ**	**Diff.**
Männlich	45,9%	42,8%	*3,1%*	63,4%	59,5%	*3,9%*	53,9%	59,0%	*-5,1%*	52,6%	38,7%	*13,9%*
Weiblich	54,1%	57,2%	*-3,1%*	36,6%	40,5%	*-3,9%*	46,1%	41,0%	*5,1%*	47,4%	61,3%	*-13,9%*

Es ist offensichtlich, dass die Extraktionsmethoden (wie z.B. die Benutzung von Telefonbüchern, die Uhrzeit des Interviews oder die Sachkenntnis der Interviewer) die Geschlechtsverteilung kaum beeinflusst hat.

gungstechnischen Gründen wurde die *last-birthday*-Methode ohne explizite Aufzählung der zur Stichprobe gehörigen Haushaltsmitglieder durchgeführt.

4.2 Familienstand

Die zweite Variable, die zur Untersuchung der Qualität der Stichprobe herangezogen wurde, ist der Familienstand. In diesem Fall ist eine allgemeine Überrepräsentation der Verheirateten im Vergleich zu den Alleinstehenden in der EIMSS-Stichprobe zu beobachten. Diese Abweichung ist vor allem bei Spaniern (12,7 Prozentpunkte) und Italienern (9,1 Prozentpunkte) bedeutsam (Tabelle 3).

Bei Betrachtung der mittleren Abweichung pro Nationalität kann festgestellt werden, dass die beste Übereinstimmung zwischen EIMSS und Mikrozensus bei den Engländern liegt, gefolgt von den Franzosen, Italienern und Spaniern. Die mittlere Abweichung der absoluten Differenzen pro Kategorie und Nationalität können als ein globaler Indikator der Übereinstimmung zwischen beiden Datensätzen herangezogen werden. Bei der Variable Familienstand zeigt sich eine recht niedrige mittlere Differenz (3,7 Prozentpunkten) mit einer Standardabweichung von 2,5.

Tabelle 3 Vergleich zwischen EIMSS und Mikrozensus – Familienstand

	Nationalität											
	Französisch			Englisch			Italienisch			Spanisch		
	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.
Verheiratet	57,2%	52,0%	*5,2%*	67,3%	67,0%	*0,3%*	84,3%	75,2%	*9,1%*	69,7%	57,0%	*12,7%*
Geschieden	7,2%	9,6%	*-2,4%*	9,8%	8,4%	*1,4%*	2,8%	6,2%	*-3,4%*	5,2%	7,5%	*-2,3%*
Verwitwet	1,6%	1,7%	*-0,1%*	1,6%	3,5%	*-1,9%*	1,2%	2,3%	*-1,1%*	2,0%	1,1%	*0,9%*
Nie verheiratet	34,0%	36,7%	*-2,7%*	21,3%	21,1%	*0,2%*	11,8%	16,3%	*-4,5%*	23,1%	34,4%	*-11,3%*

	Französisch	Englisch	Italienisch	Spanisch
Mittlere Abweichung	2,6%	1,0%	4,5%	6,8%
Standard-abweichung	2,1%	0,8%	3,4%	6,1%

4.3 Alter und Migrationsalter

Der Vergleich der Altersverteilungen für die EIMSS-Umfrage und den Mikrozensus zeigt für die Italiener und Spanier eine Überrepräsentation in der Altersgruppe der 40- bis 59-Jährigen (12 Prozentpunkte für die Italiener und 20,3 für die Spanier) sowie eine Unterrepräsentation für die 18- bis 39-jährigen in vergleichbarer Größenordnung (Tabelle 4). Bei Briten und Franzosen ergeben sich demgegenüber keine nennenswerten Abweichungen zum Mikrozensus.

Tabelle 4 Vergleich zwischen EIMSS und Mikrozensus – Alter

	Nationalität											
	Französisch			Englisch			Italienisch			Spanisch		
	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.
18-39	54,5%	55,0%	*-0,5%*	34,6%	34,4%	*0,2%*	31,1%	42,4%	*-11,3%*	38,3%	59,1%	*-20,8%*
40-59	43,1%	43,2%	*-0,1%*	59,1%	58,1%	*1,0%*	63,8%	51,8%	*12,0%*	56,9%	36,6%	*20,3%*
60+	2,4%	1,7%	*0,7%*	6,3%	7,5%	*-1,2%*	5,1%	5,8%	*-0,7%*	4,7%	4,3%	*0,4%*
	Französisch			Englisch			Italienisch			Spanisch		
Mittlere Abweichung	0,4%			0,8%			8,0%			13,8%		
Standard-abweichung	0,3%			0,5%			6,3%			11,6%		

Hinsichtlich des Alters zum Zeitpunkt der Migration ist die Übereinstimmung der EIMSS-Stichproben mit dem Mikrozensus gut, besonders für Briten und Franzosen (Tabelle 5). Die Abweichungen für Italiener und Spanier betragen etwa 5 Prozentpunkte, wobei die Jüngeren (unter 29 Jahre zum Zeitpunkt der Migration) leicht überrepräsentiert sind.

Tabelle 5 Vergleich zwischen EIMSS und Mikrozensus – Migrationsalter

	Nationalität											
	Französisch			Englisch			Italienisch			Spanisch		
	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.
< 29	69,0%	69,9%	*-0,9%*	52,8%	53,7%	*-0,9%*	78,0%	69,1%	*8,9%*	71,5%	62,4%	*9,1%*
30-39	24,3%	24,5%	*-0,2%*	33,1%	31,7%	*1,4%*	15,7%	21,0%	*-5,3%*	23,3%	32,3%	*-9,0%*
40-49	5,1%	3,9%	*1,2%*	9,8%	9,7%	*0,1%*	3,9%	6,8%	*-2,9%*	4,3%	2,2%	*2,1%*
50+	1,6%	1,7%	*-0,1%*	4,3%	4,8%	*-0,5%*	2,4%	3,0%	*-0,6%*	0,8%	3,2%	*-2,4%*
	Französisch			Englisch			Italienisch			Spanisch		
Mittlere Abweichung	0,6%			0,7%			4,4%			5,7%		
Standard-abweichung	0,5%			0,6%			3,5%			3,9%		

4.4 Migrationsperiode

Bezüglich der Migrationsperiode zeigt Tabelle 6, dass die Franzosen und Briten der EIMSS-Stichprobe der Realität des Mikrozensus sehr gut entsprechen. Allerdings zeigt sich eine größere Überrepräsentation von Italienern und Spaniern, die zwischen 1974 und 1983 migriert sind, sowie eine Unterrepräsentation dieser Gruppen in der jüngsten Periode (1993-2003).

Tabelle 6 Vergleich zwischen EIMSS und Mikrozensus – Migrationsperiode

	Nationalität											
	Französisch			Englisch			Italienisch			Spanisch		
	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.
1974-1983	26,3%	19,2%	*7,1%*	30,7%	30,0%	*0,7%*	46,5%	29,9%	*16,6%*	41,9%	24,7%	*17,2%*
1984-1993	23,1%	29,3%	*-6,2%*	28,7%	31,3%	*-2,6%*	32,3%	37,5%	*-5,2%*	22,9%	18,3%	*4,6%*
1994-2003	50,6%	51,5%	*-0,9%*	40,6%	38,8%	*1,8%*	21,3%	32,5%	*-11,2%*	35,2%	57,0%	*-21,8%*

	Französisch	Englisch	Italienisch	Spanisch
Mittlere Abweichung	4,7%	1,7%	11,0%	14,5%
Standard-abweichung	3,4%	1,0%	5,7%	8,9%

4.5 Bildung

Die Variable Bildung zeigt im EIMSS eine allgemeine leichte Überrepräsentation der Bevölkerung mit einer höheren Bildung, vor allem bei den Franzosen in Deutschland (Tabelle 7).

Tabelle 7 Vergleich zwischen EIMSS und Mikrozensus – Bildung

	Nationalität											
	Französisch			Englisch			Italienisch			Spanisch		
	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.
<Fachabitur	18,2%	27,6%	*-9,4%*	31,5%	38,0%	*-6,5%*	72,6%	79,1%	*-6,5%*	50,8%	43,5%	*7,3%*
(Fach-)Abitur	13,8%	24,3%	*-10,5%*	17,1%	17,8%	*-0,7%*	21,0%	11,8%	*9,2%*	17,5%	16,5%	*1,0%*
(Fach-) Hochschule	68,0%	48,1%	*19,9%*	51,4%	44,1%	*7,3%*	6,3%	9,1%	*-2,8%*	31,7%	40,0%	*-8,3%*

	Französisch	Englisch	Italienisch	Spanisch
Mittlere Abweichung	13,3%	4,8%	6,2%	5,5%
Standard-abweichung	5,8%	3,6%	3,2%	4,0%

4.6 Erwerbstätigkeit

Hinsichtlich der momentanen beruflichen Situation zeigt Tabelle 8 eine gute Übereinstimmung für die Franzosen und Briten in Deutschland und eine mittlere Übereinstimmung für Italiener und Spanier mit einer Überrepräsentation von arbeitenden Befragten im EIMSS (7,6% bzw. 11,7%). Dieser Zusammenhang wird erklärbar, wenn man berücksichtigt, dass im EIMSS ebenfalls eine leichte Überrepräsentation von Spaniern und Italienern vorliegt, die in jungen Jahren und vor 1983 nach Deutschland kamen und die jetzt zwischen 40 und 59 Jahre alt sind.

Tabelle 8 Vergleich zwischen EIMSS und Mikrozensus – berufliche Situation

	Nationalität											
	Französisch			Englisch			Italienisch			Spanisch		
	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.	EIMSS	MZ	Diff.
erwerbstätig	71,4%	72,9%	*-1,5%*	78,7%	73,6%	*5,1%*	73,6%	66,0%	*7,6%*	75,1%	63,4%	*11,7%*
nicht erwerbstätig	28,6%	27,1%	*1,5%*	21,3%	26,4%	*-5,1%*	26,4%	34,0%	*-7,6%*	24,9%	36,6%	*-11,7%*

Auf Grund dieser Daten lässt sich für die betrachteten Variablen sagen, dass die durch das Selektionsverfahren erhaltenen Stichproben hinreichend gut an die tatsächliche Struktur der untersuchten Populationen angepasst sind. Die Unterschiede zum Mikrozensus können darüber hinaus nicht nur auf die Methode zur Extraktion von Telefonnummern zurückgeführt werden, sondern liegen auch an der allgemeinen Beschränkung auf Festnetzanschlüsse. Der Anteil derjenigen, die nur über einen Mobilfunkzugang und keinen Festnetzanschluss mehr verfügen, steigt ständig und damit sinkt der Teil der Population, der auf der Grundlage von Festnetz-Stichproben erreichbar ist. Dennoch kann festgehalten werden, dass zumindest in Deutschland die Stichprobenziehung noch nicht zu einer erheblichen Beeinträchtigung der Validität der Daten geführt hat.

5 Zusammenfassung und Ausblick

Der methodische Ansatz, der im PIONEUR-Projekt verfolgt wurde, kann aus verschiedenen Gründen als innovativ gelten. Erstens wurde die Studie in unterschiedlichen Ländern in einer vergleichbaren Weise durchgeführt, während die frühere Forschung meist auf ein einziges Zielland beschränkt ist. Zweitens berücksichtigt die Studie durch die Ziehung von Zufallsstichproben Repräsentativitätsgesichtspunkte, während die überwiegende Zahl von Migrantenbefragungen qualitativ angelegt ist. Schließlich wurde in allen Ländern auch der gleiche Fragebogen und die gleiche Art von Interviewern eingesetzt, die sowohl die Sprache des Herkunfts- als auch des Ziellandes perfekt beherrschten.

Zum Erfolg der Studie hat das verwendete Verfahren zur Ziehung der Telefonstichproben aber nicht unwesentlich beigetragen. Wie wir hier gezeigt haben, wird die Qualität der Stichprobe durch den Vergleich der Verteilungen des Geschlechts, des Familienstands, des Alters bei der Befragung und zum Zeitpunkt der Migration, der Migrationsperiode, der Bildung und des Erwerbsstatus mit dem Mikrozensus weitgehend bestätigt, wenn es auch einige Abweichungen gibt. Diese Abweichungen dürfen aber ohnehin nicht ausschließlich dem bei der Ziehung der Stichprobe verwendeten Verfahren angelastet werden, sondern liegen zumindest auch an der Qualität der für die jeweiligen Zielländer vorhandenen Telefonbücher sowie an Prozessen der Nicht-Erreichbarkeit und Teilnahmeverweigerung durch die Befragten, wie sie für Telefonumfragen insgesamt zu berücksichtigen sind. Auch die zusätzliche Berücksichtigung einer kleinen Netzwerkstichprobe im Design, bei dem zusätzlich Telefonnummern von Frauen gesammelt wurden, die mit einem Mann des Ziellandes verheiratet sind, dürfte die Ergebnisse kaum beeinflusst haben, zumal nur 26 realisierte Interviews aus diesem Rekrutierungsvorgang stammen.

Literatur

Eurostat (2005). *Europe in Figures – Eurostat Yearbook 2005*. Luxembourg: Office for Official Publications of the European Communities.

Humpert, A., & Schneiderheinze, K. (2000). Stichprobenziehung für telefonische Zuwandererumfragen. Einsatzmöglichkeiten der Namensforschung. *ZUMA-Nachrichten 47,* 36-64.

Humpert, A., & Schneiderheinze, K. (2002). Stichprobenziehung für telefonische Zuwandererumfragen. Praktische Erfahrungen und Erweiterung der Auswahlgrundlage. In S. Gabler, & S. Häder (Hrsg.), *Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland* (S. 187-214). München: Waxmann.

Rother, N. (2005). Wer zieht innerhalb der EU wohin und warum? Das PIONEUR-Projekt. *ZUMA-Nachrichten, 56,* 94-97.

Santacreu Fernández, O. A. (2005). Diseño muestral para una encuesta telefónica a nivel Europeo. In J. Andreu, J. L. Padilla, & M. de Mar Rueda (Hrsg.). *Libro de Actas del III Congreso de Metodología de Encuestas* (S. 272-279). Sevilla: Sociedad Internacional de Profesionales de la Investigación.

Korrespondenzadressen

Oscar Santacreu Fernández
Universidad de Alicante
Dpto. Sociología II
Apdo. Correos, 99
E-03080 Alicante/Spain
E-Mail: oscar.santacreu@ua.es

Nina Rother,
PD Dr. Michael Braun
ZUMA
Postfach 12 21 55
68072 Mannheim
E-Mail: rother@zuma-mannheim.de
E-Mail: braun@zuma-mannheim.de