Messung und Analyse des sozialen Wandel anhand der Vergabe von Vornamen: Aufbereitung und Auswertung des SOEP

Dokumentation der Datenbasis und der Vercodung

von

Denis Huschka[1)], Jürgen Gerhards[2)] und Gert G. Wagner[3)]

[1)] Institut für Soziologie der Freien Universität Berlin &
Sozio-oekonomisches Panel (SOEP) am DIW Berlin (Deutschen Institut für Wirtschaftsforschung)
[2)] Institut für Soziologie der Freien Universität Berlin
[3)] Sozio-oekonomisches Panel (SOEP) am DIW Berlin (Deutschen Institut für Wirtschaftsforschung) &
Technische Universität Berlin

August 2005

Wir bedanken uns bei folgenden Personen, die insbesondere zur Aufbereitung der Datensätze und der Codes, beigetragen haben: Gabriele Rodriguez, Namensberatungsstelle der Universität Leipzig; Ulrich Kohler, Wissenschaftszentrum Berlin für Sozialforschung; Joachim R. Frick, SOEP/Deutsches Institut für Wirtschaftsforschung Berlin und Peter Krause, SOEP/Deutsches Institut für Wirtschaftsforschung Berlin.

1Einführung
2 Die Datenbasis: das Sozio-oekonomische Panel (SOEP)
3 Die Bereinigung und Vercodung der Vornamen
3. 1 Die Bereinigung und Standardisierung von Vornamen
3. 2 Die Regionale Klassifikation von Vornamen
3. 3 Die kulturhistorische Klassifikation der Vornamen Besonderheiten
3. 4 Besonderheiten
4 Die Generierung eines Arbeitsdatensatzes
4. 1 Datensatzdesign und Spezifikationen
4. 2 Die Generierung von geburtszeitpunktbezogenen Elterninformationen in „neuen“ Variablen
4. 3 Probleme, Limitationen und Besonderheiten im Umgang mit umstrukturierten SOEP Daten
5 Abschließende Bemerkungen
Literatur

1 Einführung

Vornamen sind, im Gegensatz zu Nachnamen, wählbare Attribute. Und es sind in der Regel die Eltern des Kindes, die - manchmal in Abstimmung mit Verwandten und Freunden - aus der endlichen Zahl an Vornamen eine Auswahl treffen. Jeder wird im Freundeskreis oder selbst erfahren haben, wie sich der Prozeß der Namenswahl abspielt: Vorschläge werden gemacht und verworfen. Bücher werden zu Rate gezogen, um sich über die Bedeutung der verschiedenen Namen zu informieren. Es werden Hitlisten angesehen, man hört sich um, welche Namen wie häufig im eigenen Umfeld auftauchen. Manche geben aus Tradition die Vornamen der Eltern oder Großeltern weiter, einige entwickeln eine große Kreativität und versuchen Neuschöpfungen anerkennen zu lassen.

In Deutschland greifen Eltern bei der Auswahl eines Vornamens auf einen Pool an rechtlich zugelassenen Vornamen zurück. Die Vornamensexpertin Gabriele Rodriguez der Namensberatungsstelle der Universität Leipzig beziffert die derzeit tatsächlich benutzten Vornamen auf mindestens 50.000. Dazu kommen Vornamen, die außer Mode sind und das Reservoir der benutzbaren Vornamen auf mehrere Hunderttausend Vornamen anschwellen lassen. Eine Antwort auf die Frage, wie viele verschiedene Vornamen es in Deutschland tatsächlich gibt, kann nicht gegeben werden, da es an einer zentralen standesamtlichen Statistik fehlt.

Da die Auswahl eines Vornamens eine soziale Handlung im Sinne Max Webers ist, kann man Vornamen auch als „soziale Tatsache“ begreifen und nach den gesellschaftlichen Rahmenbedingungen fragen, die die Selektion von Vornamen prägen. Gerhards (2003, 2005) hat in einer abgeschlossenen Studie den Einfluß verschiedener, sich über die Zeit wandelnder gesellschaftlicher Einflußfaktoren auf Vornamenspräferenzen nachgewiesen. Die von ihm benutzen Daten waren aber in zwei Punkten für tiefer gehende Analysen begrenzt. Zum einen handelt es sich um Fallstudien über zwei Gemeinden, so dass man nicht weiß, ob die gefundenen Ergebnisse repräsentativ für Deutschland sind. Zum anderen beruhen die Analysen auf Standesamtsdaten. Diese enthalten, außer Angaben zur Religionszugehörigkeit und zum Beruf, keine Informationen über soziodemographische Merkmale der Eltern. Eine Analyse der Daten des **S**ozio-**oe**konomischen **P**anels (SOEP) ermöglicht u.a. genau diese beiden Beschränkungen zu überwinden: Die anonymen Daten sind repräsentativ für Deutschland und enthalten eine Fülle von Informationen über die Eltern und die Kinder.

Wir werden in Projekt[1] „Messung und Analyse des sozialen Wandel anhand der Vergabe von Vornamen - Aufbereitung und Auswertung des SOEP“ eine Vielzahl von Hypothesen überprüfen. Aus der Vielzahl an Fragestellungen greifen wir hier zu Illustrationszwecken diejenigen heraus, die sich auf den sozialen Wandel beziehen und erläutern diese kurz:[2]

a. Veränderung familiärer Bindungen. Bindungen an die Traditionen der eigenen Familie finden in der Vergabe von Vornamen in der Weitergabe der Namen der Eltern an die Kinder ihren Ausdruck. Ein Generationsvergleich der im SOEP Befragten ermöglicht es zu prüfen, ob man im Zeitverlauf ein Nachlassen familiärer Bindungen beobachten kann. (*Operationalisierung: Entwicklung des Verhältnisses von weitergegebenen Namen zu nicht weitergegebenen Namen*).

b. Veränderung religiöser Bindungen/Säkularisierung: Der Jenseitsbezug in der Vergabe von Vornamen erfolgte traditionell durch die Bezugnahme auf die Namen christlicher Heiliger. Wir sprechen im Hinblick auf die Analyse der Entwicklung von Vornamen dann von einem Säkularisierungsprozess, wenn der Anteil der Namen christlichen Ursprungs zurückgeht. Ob dies der Fall ist, können wir mit Hilfe eines Generationsvergleichs prüfen und spezifizieren. (*Operationalisierung: Entwicklung des Verhältnisses von religiösen zu nicht-religiösen Namen*).

c. Veränderung der Bindung an nationale Traditionsbestände. Neben christlichen Namen sind es die deutschen Namen, die die traditionellen Bezugspunkte in der Vergabe von Vornamen darstellen. Wir wollen den Zusammenhang zwischen der Entwicklung der deutschen Namen und dem Wechsel politischer Regime und dem damit verbundenen Aufstieg und Niedergang des Nationalismus (Weimarer Republik, Nationalsozialismus, BRD bzw. DDR) durch einen Generationsvergleich analysieren. Welchen Einfluss hat der Aufstieg des Nationalismus und seine Übersteigerung im Nationalsozialismus auf die Vergabe von Vornamen, und wie wirkt sich die Delegitimierung des Nationalismus in Deutschland nach 1945 auf die Vergabepraxis

[1] Es handelt sich um ein im Normalverfahren von der DFG finanziertes Projekt (DFG Sachbeihilfe Az. GE 991/7-1). Vgl. dazu den Projektantrag Gerhards und Wagner 2002, auf http://userpage.fu-berlin.de/~gerhards/proj_vornamen.html.

[2] Neben den genannten Fragestellungen bearbeiten wir weitere soziologische Fragestellungen, aber auch solche, die sich beispielsweise auf die Muster der Verteilungen von Vornamen und weitere linguistische und statistische Aspekte beziehen. Im Rahmen des Projektes werden derzeit grundlegende Artikel zur Methodik von Namensverteilungsvergleichen publiziert, welche für das Projekt im Hinblick auf die statistisch valide Messung von Differenzierungsprozessen als Vorarbeit relevant sind.

der Vornamen aus? (*Operationalisierung: Entwicklung des Verhältnisses von deutschen zu nicht-deutschen Namen*).

d. Transnationalisierung von Vornamen. Die Öffnung in der Vergabe von Vornamen gegenüber vormals fremden Kulturkreisen kann man als Transnationalisierungsprozess beschreiben; man findet damit Anschluss an die Globalisierungsdebatte; ob und in welchem Maße welche vormals fremde Namen Eingang in das Namensrepertoire finden, kann man ebenfalls mit einem Generationsvergleich prüfen. Von besonderem Interesse ist diese Entwicklung im Vergleich BRD und DDR; anhand der Unterschiede kann man bestimmen, ob und wann sich die BRD- und DDR-Gesellschaften von der ursprünglichen „gesamtdeutschen" Gesellschaft abgelöst haben. (*Operationalisierung: Entwicklung des Verhältnisses von Namen aus fremden Kulturkreisen zu Namen aus dem deutschen und christlichen Kulturkreis*).

e. Individualisierungsprozesse in der Vergaben von Vornamen: Je unähnlicher die Lebensbedingungen der Menschen bedingt durch die Zunahme an sozialen Kreisen, in denen sie verkehren und durch die Zunahme der Arbeitsteilung in einer Gesellschaft, desto individueller werden ihre Präferenzen und Geschmacksorientierungen sein. In der Vornamensvergabe führen Prozesse der Individualisierung dazu, daß immer weniger Menschen den gleichen Namen teilen. (*Operationalisierung: Entwicklung der Streuung der Vornamen*).

All diese Prozesse werden im Hinblick auf Klassen-, Stadt/Land-, West/Ost-, und ethnische Zugehörigkeit genauer spezifiziert und analysiert. Wir wollen an dieser Stelle die einzelnen Hypothesen und das analytische Potential des Projekts insgesamt nicht genauer explizieren. Dies bleibt anderen Papieren vorbehalten. Der Zweck des hier vorgelegten Papiers ist ein anderer: Es soll vor dem Hintergrund unserer spezifischen Fragestellungen einen Überblick über die Datensatzgenerierung sowie die vorgenommenen Codierungen geben. Dies ist nicht nur im Sinne einer methodischen Selbstreflexion und zur Nachvollziehbarkeit der Ergebnisse notwendig, sondern auch insofern sinnvoll, weil die vorgenommenen Codes nach Abschluß des Projektes allen interessierten Nutzern des SOEP- Datensatzes zur Verfügung gestellt werden.

Die Art und Weise, wie die Daten aufbereitet und die Namencodes vergeben wurden ist natürlich durch die Fragestellungen der Studie gesteuert. Wer mit anderen Fragestellungen

die Vornamen des SOEP analysieren möchte, muss neue Codes entwickeln und entsprechende Datensätze generieren. Diese Arbeit muß in datenschutzrechtlich einwandfreier Weise im DIW Berlin (Deutschen Institut für Wirtschaftsforschung) erledigt werden.

2 Die Datenbasis: das Sozio-oekonomische Panel (SOEP)

Das **S**ozio-**oe**konomische **P**anel (kurz: SOEP) ist eine der weltweit größten prospektiven Längsschnittstudien, welche systematisch für die Bevölkerung eines Landes repräsentative Daten erhebt.[3] Das Erhebungsinstrument hat den Charakter einer „face to face" durchgeführten Mehrthemen-Befragung. Die aufbereiteten Daten sind Teil der weltweiten Dateninfrastruktur der Sozial- und Wirtschaftswissenschaften (vgl. zusammenfassend: Wagner et al 1993).

Das Grunddesign des SOEP ist das eines Längsschnittes auf Personenebene, wobei jedes Stichproben-Mitglied innerhalb seines jeweiligen Haushaltes befragt wird. Die Haushalts- und Individualinformationen werden über die Zeit zum großen Teil immer wieder erhoben, so dass sich Veränderungen analysieren lassen. Diese Longitudinaldaten stehen aber nicht explizit im Zentrum unserer Auswertungen, da Vornamen nur einmalig vergeben werden. Wie wir später ausführen werden, erleichtert das Longitudinaldesign jedoch die zeitlich korrekte Zuordnung von Elterninformationen zu vergebenen Vornamen von Kindern.

Die Ersterhebung des SOEP erfolgte 1984. Damit enthalten die aktuellen Teildatensätze des SOEP für die ursprünglichen Stichprobenmitglieder, die noch befragt werden, Informationen, die derzeit einen Zeitraum von 20 Jahren umfassen. Die Samplegröße umfaßte im Jahr 1984 5.921 Haushalte. Dabei wurden auch die fünf großen Gruppen nichtdeutscher Arbeitsmigranten erfaßt. Das SOEP wurde jedoch im Laufe der Zeit mehrfach erweitert und deckt seit 1990 Ostdeutschland (2.179 Haushalte) und seit 1994/5 Zuwanderer, darunter auch Aussiedler (402 Haushalten) repräsentativ ab (vgl. Burkhauser/Kreyenfeld/Wagner 1997).

Um auf der Basis einer großen Fallzahl bessere Analysen kleiner Teilgruppen der Bevölkerung zu ermöglichen, wurde das SOEP im Jahre 1998 um eine Ergänzungsstichprobe (1.056 Haushalte) sowie im Jahr 2000 um eine Innovationsstichprobe (4.586 Haushalte) erweitert (vgl. Schupp/Wagner 2002). Im Jahre 2002 konnte schließlich eine Zufalls-

[3] Für detailliertere Informationen über das SOEP kann man auf die umfassenden Informationsangebote der SOEP-Gruppe am Deutschen Instituts für Wirtschaftsforschung unter der Webadresse www.diw.de/soep zurückgreifen. Hier findet sich neben umfänglichen Informationen zu den Daten auch die Recherchemöglichkeiten „SOEPinfo" und „SOEPlit", in welchen sich neben einem Online-Zugriff auf alle verwendeten Indikatoren auch nahezu alle auf dem SOEP basierenden Publikationen (auch strukturiert nach „Keywords") befinden.

Stichprobe G für Haushalte von Hocheinkommensbeziehern realisiert werden. Sie umfaßt 1.224 Haushalte (vgl. insgesamt für den Stand 2005 Haisken-DeNew/Frick 2005).

Insgesamt liegen uns für Auswertungen im Rahmen unseres Projektes Informationen und Vornamen von 50.454 Personen vor (Stand 2004). Damit ist das SOEP für nahezu alle relevanten Bevölkerungsgruppen, aber auch für viele Regionalanalysen, repräsentativ auswertbar.

Die SOEP-Daten wurden immer wieder für Analysen genutzt, an die zu Beginn der Erhebung im Jahre 1984 *nicht* gedacht wurde, so z. B. für die Analyse der Re-Migration von Arbeitsmigranten zurück in ihre Heimatländer (vgl. z. B. Merkle und Zimmermann 1992, Schmidt 1994) und für die Analyse der sozio-ökonomischen Differenzen der Lebenserwartung, (vgl. zuerst Schepers und Wagner 1989 sowie später z. B. Voges und Schmidt 1996). In der Reihe "überraschender" SOEP-Analysen steht auch die Analyse der Vornamen der Befragten.

Bei allen Erhebungen, die in Deutschland durchgeführt werden, dürfen Adressen aus Datenschutzgründen nicht analysiert werden (Trennungsgebot von Adressen und erhobenen Daten). Im SOEP werden die Vornamen jedoch erhoben (d. h. erfragt), ursprünglich nur, um eine zusätzliche Sicherheit (und Prüfmöglichkeiten) für die richtige Verknüpfung von Informationen mit den Personen innerhalb der Haushalte und über der Zeit zu erhalten.

Für alle Teilnehmer des SOEP werden vom Interviewer die Vornamen ermittelt, d. h. sie werden von Befragten angegeben und für Kindern im Befragungshaushalt von den Eltern erfragt. Die Vornamen sind also nicht nur Bestandteil der Adressen der Befragungshaushalte und sie sind deswegen grundsätzlich analysierbar. Nachnamen sind nicht nur aus juristischen Gründen nicht analysierbar, sondern sie werden bereits beim Erhebungsinstitut „Infratest Sozialforschung" zusammen mit den Adressen von den Befragungsdaten getrennt und sind deswegen auch technisch nicht verfügbar.

Man kann davon ausgehen, dass es sich bei den im SOEP erhobenen Vornamen um „Rufnamen" handelt. Erkennbar ist das z.B. am geringen Anteil von Bindestrich-Vornamen (vgl. auch Abschnitt 3.4 unten).

Da Vornamen die De-Anonymisierung von Befragten erheblich erleichtern können, werden sie auch innerhalb der SOEP-Gruppe im DIW Berlin nicht in der dort allgemein verfügbaren SOEP-Datenbank gehalten, sondern sie sind, wie einige weitere datenschutzrechtlich sensitive Informationen, z. B. über die regionale Verortung von Befragungshaushalten, gesondert gespeichert. Sie können, unter Beachtung umfangreicher und restriktiver Datenschutzvorkehrungen, temporär und auf einem speziell gesicherten Computer zu Zwecken der Auswertung den normalen SOEP-Daten zugespielt werden.. Um die Datensicherheit zu erhöhen, werden die Befragungsdaten und die Vornamen faktisch nur sehr sparsam zusammengeführt, nämlich Geburtsjahr, Geschlecht und Nationalität zum Zwecke der Vercodung der Daten. Die eigentlichen Befragungsdaten werden also nur mit den Codes zusammengespielt.

Es werden keine Analyse-Ergebnisse mit Klartext-Vornamen veröffentlicht werden. Dies wäre zwar nur bei seltenen Namen verboten, könnte aber auch bei häufigeren Namen zu Irritationen führen. Ebenso aus Datenschutzgründen verwenden wir auch im Folgenden als Beispiele für die Vercodung nur sehr häufig vorkommende Vornamen und Schreibweisen, oder aber „fiktive“, nicht im SOEP vorkommende Namen. Vornamen die seltener als 10 Mal im Datensatz vorkommen werden, werden auch als Beispiele keinesfalls genannt.

3 Die Bereinigung und Vercodung der Vornamen

Die Voramen müssen bereinigt, standardisiert und vercodet werde. Die Benutzung von Codes anstatt der Namen selbst erhöht zusätzlich den Schutz vor Aufdeckung der Identitäten der SOEP Befragten.

Anders als die meisten im SOEP erhobenen Informationen werden die Vornamen datentechnisch betrachtet als sogenannte Strings abgespeichert. Ein „String“ ist eine Folge von Buchstaben (oder auch Zahlen bzw. Kombinationen), also ein Wort, welcher ein Name ist. Dieses Format bringt einige Probleme mit sich. Die meisten Fragen des SOEP beantworten die Teilnehmer, indem sie eine der vorgegebenen Antwortkategorien ankreuzen. Diese eineindeutigen Antworten lassen sich recht einfach in eine Zahl transferieren. Mit Stringvariablen verhält sich dies anders. Sie sind nicht skaliert, weil es kein Klassifikationssystem gibt, das man mit den Stringangaben verknüpfen kann. In unserem Fall gibt es zu viele Namen, viele davon auch jeweils nur einmal, als daß sich ein sinnvolles Zahlencodesystem benutzen ließe.

Die erste Aufgabe im Rahmen des Vornamenprojektes war deshalb die Entwicklung und Implementation eines geeigneten Klassifikationsschemas für alle im Datensatz vorhandenen Namen. Es mußte ein Modus gefunden werden, der die Vornamen in eine handhabbare Anzahl von Kategorien „codiert“. Ziel einer solchen Vercodung ist es, in einer stringenten, nachvollziehbaren und auf gleichen Annahmen beruhenden Art und Weise alle Namen im Datensatz „einzufangen“. Gegebenenfalls müssen die Kategorien erweiterbar und spezifizierbar sein, um zukünftig neue Namen integrieren zu können. Vor allem aber müssen die Codes das klassifizieren, was wir im Rahmen des Projektes auswerten wollen.

Entlang der Fragestellungen des Projektes nutzen wir die „Information“ Vornamen auf zwei verschiedenen Weisen:

a) Für Analysen der Streuung, Verbreitung, und Entwicklung der Vornamen aber auch bei der Beantwortung der Frage, ob eine Individualisierung oder Transnationalisierung der Vornamen zu beobachten ist, benötigen wir vor allem die Vornamen selbst. Um einer künstlich aufgeblähten Variabilität vorzubeugen, müssen die Vornamen im ersten Schritt bereinigt und standardisiert werden. Weiterhin wird ein Code nötig, der die jeweilige „geographische Herkunft“ eines Namens angibt. Daneben sind

Informationen über das Geburtsjahr und das Geschlecht der *Namensträger* (sowie weitere Informationen) unerläßlich. Eine Verknüpfung der Vornamen und ihren Charakteristika mit den Informationen, die uns jeweils über die *Namensgeber* vorliegen, ist für diese Art der Analysen noch nicht notwendig.

b) Eine schwierige Aufgabe ist es, angesichts der Vielzahl der Namensmerkmale ein Aggregierungssystem zu entwickeln, welches, strukturiert nach den uns interessierenden Merkmalen, eine Handhabung der Merkmalsunterschiede möglich macht. In unserem Falle sind die interessierenden Merkmale die „kulturhistorischen Herkünfte" der einzelnen Namen. Für soziologische Analysen auf der Individualebene ist es weiterhin notwendig, die Namen bzw. die Namenscodes mit den Hintergrundsvariablen der jeweiligen Eltern zu verknüpfen um beispielsweise den Einfluß der praktizierten Religiosität der Eltern auf die Charakteristika des Namens des Kindes (z. B. Heiligenbezug im Namen) zu untersuchen. Die Erstellung dieses Arbeitsdatensatzes (der je nach Auswertungsfrage variieren kann) ist eine etwas komplexere Angelegenheit, die unten (Abschnitt 4) exemplarisch beschrieben wird.

Je tiefer man sich in die Namensforschung (Onomastik) hineinarbeitet, um so deutlicher lernt man, daß es nicht einfach ist, eine stringente und eineindeutige Verfahrensanweisung zur Vercodung von Vornamen zu finden. Es stellen sich bei der Genese der Namenscodes Fragen wie: Ab wann gilt ein Name als „eingedeutscht"? Wie geht man mit verschiedenen, zeitlich gestaffelten kulturhistorischen Bezügen (etwa wenn ein Vorname altgriechischen Ursprungs ist, später aber als christlicher oder Heiligenname betrachtet wird) um? Die folgenden Absätze 3.a) und 3.b) beschreiben detailliert die für dieses Projekt gewählte technische als auch inhaltliche Herangehensweise in der Bereinigung und Standardisierung der Vornamen und der Generierung der Namenscodes.

3.1 Bereinigung und Standardisierung von Vornamen

Die erste neu gebildete Variable besteht in einer bereinigten/korrigierten Schreibweise der im SOEP vorkommenden Namen. Dabei wurden ähnlich geschriebene Namen in die gebräuchlichste Schreibweise transferiert. (Beispiel: „Claus" wird als „Klaus"; codiert.) Auf diese Weise gleicht man Interviewerfehler aus die dadurch auftreten dürften, weil die Namen bei der Erhebung von Mitarbeitern des Erhebungsinstitutes aufgeschrieben wurden. Somit

sind Schreibfehler oder nicht zutreffende Schreibvarianten der tatsächlich erhobenen Namen, besonders bei ausländischen Namen, nicht auszuschließen.

Sonderzeichen konnten nicht immer berücksichtigt werden. Es wurde auf eine vereinfachte „deutsche“ Schreibform zurückgegriffen. So steht z.B. C, G, S, Z für ursprüngliches Č, Ç, Ğ, Š, Ş, Ž, auch wenn sich dann die Aussprache der Namen ändert. Die Vereinheitlichung beugt einer artifiziellen Varianz der Namen vor und ermöglicht eine robustere Berechnung der Streuung und Häufigkeit von Namen(-sgruppen).[4] Wenn der real erhobene Name der gebräuchlichsten Schreibweise entspricht, musste eine Vereinheitlichung natürlich nicht durchgeführt werden. Die vergebenen, auf die jeweiligen „regionalen“ und „kulturhistorischen Herkünfte“ der Namen abzielenden Codes wurden ausschließlich auf der Basis der unbereinigten Schreibweisen der Namen vergeben.

3. 2 Regionale Klassifikation von Vornamen

Die zweite neue Variable gibt die „regionenbezogene“ Herkunft eines Namens an. Der Name Andy wurde beispielsweise als „englisch“ klassifiziert, Maximilian, Peter und Maria als deutsch, auch wenn es sich streng genommen nicht um deutsche, sondern um christliche Namen handelt. Die Vornamen wurden nach Ihrer Herkunft aus „deutscher Sicht“ unter Berücksichtigung des Geburtsjahres vercodet. „Deutsche Sicht“ heißt: Wenn ein Vorname ein zum Zeitpunkt der Geburt des Namensträgers in Deutschland gebräuchlicher Vorname war, wurde er als „deutscher Vorname“ codiert. So versteht man beispielsweise die Namen[5] Nikolaus/Klaus/Claus, Johann(es)/Hans, Alexander, Paul, Matthias, Adam, Michael, Michaela, Anna/Anne, Maria/Marie, Eva, Elisabeth/Elisa und Franziska im deutschen Sprachraum heute als deutsche Vornamen, obwohl sie nicht deutscher, sondern hebräischen, lateinischen bzw. griechischen Ursprungs sind. Seit etwa dem 12. Jahrhundert haben sich diese Namen im Zuge der Christianisierung und Heiligenverehrung im deutschen Raum durchgesetzt und wurden als Tauf- und Rufnamen sehr gebräuchlich und beliebt. Es bildeten sich zudem noch rein deutsche Kurzformen zu den ‚fremden‘ Namen heraus, so z.B. Klaus zu

[4] Die original erhobenen Vornamen sind uns dennoch zugänglich, können also Gegenstand von eigenen Analysen sein (beispielsweise um die Vergabe von Modeschreibweisen oder von „Oberschichtsschreibweisen“ betrachten zu können), oder zu Überprüfungszwecken erneut zu Rate gezogen werden.
[5] Die im folgenden aufgeführten Namen sind teilweise „fiktive“ Beispielnamen, um das Codesystem zu verdeutlichen. Sie kommen nicht unbedingt im SOEP vor.

Nikolaus oder Hans zu Johannes. Der eigentliche Ursprung dieser Namen ist in der Regel den Namensgebern nicht bekannt, sie werden als deutsche Vornamen verstanden.

Vornamen, die erst in der Neuzeit (d. h. seit Beginn des 20. Jahrhunderts) in den deutschen Sprachraum gekommen sind, werden dagegen häufig noch als ‚fremde' (d.h. nichtdeutsche) Namen verstanden, und auch der Ursprung ist hier sehr wahrscheinlich bekannt. So erkennt man die Namen Michel, Michelle noch als französische Vornamen; Ricarda, Carmen als spanische Vornamen; Steve(n), Mike, Kevin, John, , als englische bzw. amerikanische Vornamen oder Natascha, Nadja, Tanja, Sascha als russische Vornamen.
Diese Art der Vercodung kann mit recht von Onomastikern und Namensexperten kritisiert werden. Die Krux besteht allein im Fehlen belastbarer und anwendbarer objektiver Entscheidungshilfen und Kriterien, *ab wann* ein Name beispielsweise als „deutsch" (im Sinne von „in Deutschland gebräuchlich") gelten kann. Wir glauben dennoch eine hinreichend objektive und vor allem „in sich" stringente Vercodung vorgenommen zu haben. Die Codes der Variable „Regionenbezug" spiegelt die Erfahrung bzw. das Wissen wieder, das die Eltern über die Vornamen zum Zeitpunkt der Geburt ihres Kindes *höchstwahrscheinlich* gehabt haben. Und da wir auf die Präferenzen der Eltern rückschließen wollen, müssen wir uns bei der Codierung auf *deren* Wissen über Vornamen zum Zeitpunkt ihrer Entscheidung beziehen.

Um sich in die Entscheidungssituation der Eltern versetzen zu können, benötigt man eine enorme historische Namenskompetenz. Diese stand uns zur Verfügung. Die Vornamen wurden alle von Gabriele Rodriguez codiert. Sie ist für die Vornamenberatung in der Beratungsstelle an der Universität Leipzig zuständig und ist eine *der* Experten in dem Feld.[6]

Als Referenz für die vergebenen Codes nutzte Frau Rodriguez neben ihrem Expertenwissen folgende Literatur: Seibicke (1996-2001), Salahuddin (1999), Schimmel (1992) Olivart (1993), Ilčev (1996) sowie die Personennamen-Datei der Personnamen-Beratungsstelle der Universität Leipzig, die ständig aktualisiert wird.

In der Datenauswertung können wir mit Bezug auf die Variable „Regionenbezug" z. B. Transnationalisierungsprozesse von Vornamen im Zeitverlauf analysieren (gemessen durch

[6] Gabriele Rodriguez bewältigte die umfassenden Bereinigungs- und Vercodungsaufgaben mit außerordentlichem Einsatz, Geduld und Ausdauer. Ihr gilt unser besonderer Dank!

den Anteil der Namen, die nicht in Deutschland gebräuchlich waren). Folgende Codes wurden festgelegt:

Tabelle 1: Regionenbezogene Codes der Vornamen im SOEP

Die Spalte % bezieht sich auf die ungewichtete Verteilung der Vornamen im SOEP auf der Basis von 50.454 Fällen.

Oberkategorien	%	Untergruppen
01 – Deutsche Vornamen bzw. in Deutschland seit Beginn des 20. Jh. gebräuchliche Vornamen	66.1	011 – Friesisch-deutsch 012 Süddeutsch
02 – Slawische Vornamen	4.2	021 – Russische Vornamen 022 – Polnische Vornamen 023 – Tschechische Vornamen 024 – Südslawische (bulgarische, serbo-kroatische u.ä.) 025 – Ukrainische Vornamen
03 – Romanische Vornamen	10.7	031 – Italienische Vornamen 032 – Spanische Vornamen 033 – Französische Vornamen 034 – Portugiesische Vornamen 035 – Rumänische Vornamen
04 – Englische Vornamen	5.0	041 – Angloamerikanische Vornamen 042 – Afroamerikanische Vornamen
05 – Türkische Vornamen	5.5	(nicht gesondert betrachtet wurden kurdische Vornamen)
06 – Griechische bzw. neu-griechische Vornamen	1.2	
07 – Arabische Vornamen	1.3	
08 – Persische bzw. neu-persische Vornamen	0.2	
09 – Asiatische Vornamen	0.1	091 – Chinesische Vornamen 092 – Japanische Vornamen 093 – Vietnamesische Vornamen 094 – Koreanische Vornamen 095 – Thailändische Vornamen
10 – Afrikanische Vornamen	0.1	101 Westafrikanische Vornamen (v. a. Yoruba und Igbo) 102 – Ostafrikanische Vornamen (bes. Swahili/Kiswahili) 103 – Südafrikanische Vornamen (Zulu u.ä.)
11 – Indische bzw. Hindu-Namen	0.1	
12 – Albanische Vornamen	0.2	
13 – Nordische (skandinavische) Vornamen	5.1	131 – Niederländische Vornamen 132 – Finnische Vornamen 133 – Schwedische Vornamen 134 – Dänische Vornamen 135 – Norwegische Vornamen 136 – Isländische Vornamen
14 – Mongolische Vornamen	*	
15 – Ungarische Vornamen	0.2	
16 – Jüdische bzw. israelische Vornamen	0.1	
18 – Baltische Vornamen		181 – Litauische Vornamen, 182 – Lettische Vornamen

*) Fallzahl < 10.

Untergruppen des Regionencodes wurden nur dann gebildet, wenn der Name auch nur in dieser speziellen Region gebräuchlich ist oder war. Kommen Namen in verschiedenen Sprachräumen gleichzeitig vor, wurde der gebräuchlichere codiert. So wird Anja als slawischer Vorname codiert, auch wenn der Name im nordischen und friesischen Raum als Vorname nachweisbar ist. Dagegen geht man in Deutschland beim Namen Jan in der Regel von einem niederdeutsch-friesischen Namen aus (011), obwohl er auch im slawischen Raum recht gebräuchlich ist.

Auch das Geburtsjahr und das Geschlecht der Namensträger spielte bei der Vercodung eine Rolle, da dies für eine unterschiedliche Zuordnung der Vornamen von Bedeutung sein kann. Aus Datenschutzgründen geben wir keine Beispiele.

Unterschiedliche Schreibformen von Vornamen können unterschiedliche Herkünfte des Namens anzeigen. In diesen Fällen wurden die Schreibform nicht bereinigt, sondern unterschiedlich vercodet: Denis ist ein französischer männlicher Vorname (033) (die häufige russische Form Denis sowie eine türkische Form blieben unberücksichtigt).

Es kann Fälle geben, bei denen sich die regionale Verortung über die Zeit verändert hat. Im Falle von Luca ist es offensichtlich so, daß der Name slawischen Ursprungs ist, jedoch Eingang in die englische Sphäre gefunden hat. Wir haben versucht, die Regionsvariable „zeitpunktbezogen“ zu codieren. Dieser Code (beispielsweise slawisch vs. englisch) kann nicht in die (im folgenden beschriebenen) Variablen 3 bis 5 (kulturhistorischer Bezug) übertragen werden, da diese dann in sich inkonsistent, was den Vercodungsbezug angeht, würden.

3. 3. Die kulturhistorische Klassifikation von Vornamen

Die dritte bis fünfte neu gebildete Variable beschreibt die jeweilige kulturhistorische Herkunftsverortung eines Vornamens. Das eigens entwickelte Vercodungsschema lehnt sich in seiner Konzeption an die von Gerhards (2003) verwendete Methode an, klassifiziert die Vornamen jedoch detaillierter. Zusätzlich wurde durch die Vergabe von bis zu drei Codes pro Name die zeitliche Abfolge verschiedener kulturhistorischer Zugehörigkeiten beachtet. Die erste kulturhistorische Variable ist die zeitlich nächste, die dritte Variable die zeitlich am

weitesten entfernte (falls vorhanden – hier können 1-2 Variablen unbesetzt bleiben). Vergeben wurden folgende Codes (mit Untergruppen):

Tabelle 2: Kulturkreisbezogene Codes der Vornamen im SOEP

Die Spalte % bezieht sich auf die ungewichtete Verteilung der Vornamen im SOEP auf der Basis von 50.454 Fällen.
Die hier angegebenen Prozentwerte beziehen sich auf die erste von drei möglichen Kulturkreiseinordnungen.

Oberkategorien	%	Untergruppen
10 – Christliche Namen	83.6	11 – Aus dem Alten Testament 12 – Aus dem Neuen Testament 13 – Ursprünglich ein Heiligenname
20 – Moslemischer Kulturraum	5.7	21 – Namen aus dem Koran
30 – Jüdisch-hebräischer Kulturraum	*	31 – Name aus dem aramäischen Raum und 32 – Name aus dem ägyptischen Kulturraum
40 – Antike Namen	1.1	41 – Altrömischer Kulturraum (lateinisch) 42 – Altgriechischer Kulturraum 43 – Babylonischer Raum
50 – Traditionelle Namen (nicht christliche oder islamische Namen)	2.4	51 – Aus dem friesischen Raum, 52 – Aus dem türkischen Kulturraum
60 – Germanischer Kulturraum	1.8	61 – Nordgermanischer Raum 62 – Angelsächsischer Raum 63 – Westgermanischer Raum
80 – Persischer Kulturraum	*	
90 – Türkischer Kulturraum	0.1	
100 – Romanischer Kulturraum	2.1	101 – Französischer Raum 102 – Spanischer Raum 103 – Italienischer Raum
110 – Afrikanischer Kulturraum	0.1	111 – Westafrikanischer Raum 112 – Ostafrikanischer Raum, 113 – Südafrikanischer Raum
120 – Slawischer Kulturraum	0.1	121 – Westslawischer Raum 122 – Ostslawischer Raum 123 – Südslawischer Raum
130 – Albanischer Sprachraum	0.2	131 – Illyrischer Raum
140 – Baskischer Kulturraum	*	
150 – Keltischer Kulturraum	0.2	
160 – Asiatischer Kulturraum	0.1	161 – Japanischer Raum, 162 – Vietnamesischer Raum, 163 – Chinesischer Raum, 164 – Innerasiatischer (mongolischer) Raum, 165 – Koreanischer Raum, 166 – Thailand, 167 – Kambodscha
170 – Irischer Kulturraum	0.2	171 – Irisch-gälischer Raum
180 – Englischer Kulturraum	0.7	181 – Schottischer Raum 182 – Walisischer Raum 183 – Kornischer Raum
190 – Nordisch-baltischer Raum	1.3	191 – Litauischer Raum 192 – Schwedischer Raum 193 – Grönländischer Raum 194 – Finnischer Raum
200 – Altindische (Sanskrit) Namen	0.03	
210 – Ungarische Namen	0.38	
220 – Indianische Namen (Nord- und Südamerika)	*	
230 – Polynesischer Kulturraum	*	231 – Hawaii 232 – Maori

*) Fallzahl < 10.

3. 4. Besonderheiten

Doppelnamen.

Ein sehr geringer Anteil von Personen, die im SOEP befragt wurden, haben einen Doppelnamen. Daß der Anteil so gering ist, dürfte daran liegen, dass die Interviewer nach dem „Rufnamen“ gefragt bzw. die Befragten allein den Rufnamen angegeben haben. Damit werden Analysen von Doppelnamen mit unserem Datensatz schwierig bzw. unmöglich. Wir können keine verläßlichen Aussagen z. B. über den Zusammenhang zwischen Bildung einerseits und der Vergabe mehrerer Vornamen andererseits machen. da die Nennung von Doppelnamen höchst selektiv erfolgt ist.

Doppelnamen wurden jeweils nach Ihren Bestandteilen vercodet, also nach dem ersten und dann nach dem zweiten Bestandteil. Historisch gewachsene Doppelnamen wie z.B. Annemarie, Lieselotte, Marianne u.ä., die heute als ein Name verstanden werden, wurden jedoch als ein Name vercodet, auch wenn die Bestandteile gemischten Ursprungs sind. Es wurde hier von dem bekannteren Namen ausgegangen. So wurde z. B. der Name Lieselotte (gebildet aus Liese, einer Kurzform des hebräischen Namen Elisabeth und Lotte, einer Kurzform des französischen Namens Charlotte) als christlicher Name (Charlotte hat auch einen christlichen Ursprung) mit hebräischem Ursprung vercodet.

Heiligennamen.

Tatsächlich als Heiligennamen nachgewiesene Vornamen wurden mit einem eigenen Code (13) versehen. Alle Neben-, Kurz- und Koseformen bzw. weibliche Bildungen zu männlichen Heiligennamen wurden als christlich motiviert (10) vercodet.

Verschiedene Herkunftsmöglichkeiten.

Bei einigen Namen kommen verschiedene Herkunftsmöglichkeiten in Frage. Wir haben dann die am ehesten zutreffende Herkunft codiert. Ein im Datensatz vorkommender heutiger weiblicher Vorname ist zum Beispiel ein ursprünglich afroamerikanischer weiblicher Vorname, der in den 90iger Jahren des 20. Jahrhunderts (durch eine afroamerikanische Sängerin) nach Deutschland gekommen ist. Afroamerikanische Namen gehen wiederum häufig auf arabisch-moslemische Namen zurück. Und manchmal haben diese Namen auch noch einen hebräischen (biblischen) Ursprung; dieser ist unberücksichtigt geblieben, da er für diesen Namen wohl eher unwahrscheinlich als Herkunft angenommen werden kann.

Tabelle 3: Beispiele für die Vercodung:

Name	Variable 1 Regionenbezug	Variable 3 Kulturelle Herkunft 1	Variable 2 Kulturelle Herkunft 2	Variable 4 Kulturelle Herkunft 3
Adolf	01 Deutsch	13 Heiligenname	60 Germanisch	
Angelika	01 Deutsch	100 Romanisch	10 Christlich	40 Lateinisch/antik
Arne	13 Nordisch	10 Christlich	60 Germanisch	
Birgit	01 Deutsch	190 Nord./baltisch	10 Christlich	150 Keltisch
Denis	03 Romanisch	10 Christlich	40 Lateinisch/antik	
Esther	01 Deutsch	10 Christlich	30 Hebräisch/jüdisch	80 Persisch
Gert	01 Deutsch	10 Christlich	60 Germanisch	
Hans	01 Deutsch	10 Christlich	30 Hebräisch/jüdisch	
Jürgen	01 Deutsch	10 Christlich	40 Lateinisch/antik	
Lena	01 Deutsch	10 Christlich	40 Lateinisch/antik	
Lisa	01 Deutsch	10 Christlich	30 Hebräisch/jüdisch	
Maria	01 Deutsch	10 Christlich	30 Hebräisch/jüdisch	
Nadine	03 Romanisch	10 Christlich	120 Slawisch	

4 Generierung eines Arbeitsdatensatzes

Die Vornamencodes wurden mit Hintergrundinformationen über die „Namensvergeber“ (Eltern) verknüpft. Welche Informationen aus der Vielzahl der im SOEP erhobenen Fragestellungen relevant sind, hängt von der jeweiligen Forschungsfrage ab. Insofern gibt es auch nicht „den“ Datensatz, der dann für die Analyse zur Verfügung steht. Wie bei jeder Analyse auf Grundlage des SOEP sieht sich der Nutzer der Aufgabe der individuellen Arbeitsdatensatzgenerierung entgegen. Aus der großen (und komplizierten) Datenbankstruktur SOEP gilt es die relevanten Informationen herauszufiltern und mit den richtigen Personen und Zeitpunkten zu verknüpfen.

Einen „ready to use“ Datenfile gibt es aufgrund des Umfanges des SOEP, seiner inneren und äußeren Struktur, und insbesondere wegen der Erhaltung des gewaltigen Analysepotentials nicht. Jede Art von Analyse erfordert eine spezielle, jeweils gesondert aus der SOEP Datenbank (derzeit ca. 245 Einzeldateien) zu generierende Datensatzstruktur. Böte man eine Art „fertigen“ SOEP-Datensatz an, wäre dieser für verschiedene Analysen unbrauchbar. Genau deshalb geht das SOEP diesen Schritt nicht, sondern setzt auf die Möglichkeiten, die das SOEP in seinem bewährten Weitergabeformat bereitstellt, und die umfangreiche Betreuung der jeweiligen Nutzer:[7]

Bei Analysen von Vornamen gibt es jedoch wichtige Besonderheiten. Vornamen werden von den Eltern vergeben. Dies impliziert einen Datensatzaufbau, der Elterninformationen mit den Vornamen ihrer Kinder verknüpft. Des weiteren werden Vornamen nur einmalig vergeben. Um die Einflüsse des gesellschaftlichen Umfeldes sowie individueller Elterncharakteristika in Bezug auf die Vornamenswahl für ihre Kinder zu analysieren, müssen diese zum richtigen Zeitpunkt, d.h. genau oder möglichst genau den Zeitpunkt der tatsächlichen Namenswahl treffend, mit den Vornamen der jeweiligen Kinder verknüpft werden. Dies ist nötig, da sich sowohl gesellschaftliche Einflußsphären, als auch individuelle Charakteristika ändern.

[7] Eine technische Einführung in das SOEP: siehe „Desk Top Compendium“ (http://www.diw.de/deutsch/sop/service/index.html) sowie SOEPinfo und SOEPlit (auf der gleichen Website zu finden).

4. 1 Datensatzdesign und Spezifikationen

Generell müssen zwei Bedingungen erfüllt sein, um aus einem Fall einen - bezüglich unseres engeren Projektthemas - interessanten Fall zu machen:

1. Der Vorname einer Beobachtung muß bekannt sein.
2. Es sollten so viele Informationen wie möglich über die für die Namenswahl Verantwortlichen (Eltern) zum Zeitpunkt der Namensvergabe vorhanden sein.

Insgesamt liegen im SOEP nach dem Erhebungsjahr 2004 valide Vornamensinformationen für 57.323 Personen vor. Das Geburtsjahr, das Geschlecht und die Nationalität sind im von den Befragten erhobenen SOEP-Haushaltsprotokoll enthalten. Damit erfüllt jeder Fall (von dem wir auch Namensinformationen haben) die Minimalvoraussetzungen für deskriptive Analysen (Verteilung und Streuung der Vornamen etc.). Verschiedene Fragestellungen benötigen jedoch weit darüber hinausgehende Informationen über die Eltern zum Zeitpunkt der Namensvergabe. Die Qualität und Quantität dieser Informationen ist für bestimmte Gruppen von Befragten unterschiedlich. Technisch kann man von einer Dreierklassifizierung der SOEP Analysepopulation ausgehen.

Typ 1 Fälle:

Zur Bearbeitung vieler Fragestellungen sind wir an den Charakteristika der Namensgeber (Eltern) interessiert, welche mit den Vornamen ihrer Kinder und dadurch mit den speziell generierten Charakteristika (Codes) dieser Vornamen verknüpft werden müssen. Die Nutzung eines Datensatzes, der die Untersuchungseinheit „Haushalt“ unterstützt, und die nutzerfreundliche Anlage des Vornamensfiles machen diese Verknüpfungen einfacher, da innerhalb der Haushalte alle (erwachsenen) Personen befragt werden und diese Informationen leicht mit den für die jeweiligen Kinder gewählten Vornamen zu verknüpfen sind. Das SOEP (2003) enthält seit seiner Ersterhebung Informationen über 16.967 Haushalte. In diesen Haushalten wurden 4.687 Kinder geboren. Für die Eltern dieser Kinder sind im SOEP alle relevanten Informationen direkt zum Zeitpunkt der Vornamenswahl für ihr Kind zugänglich. Diese Gruppe von befragten Eltern(teilen), die mindestens ein Kind seit 1984 bekamen, stellt die im Folgenden „Typ1“ genannte Befragungspopulation dar. Theoretisch sind nahezu alle Informationen, soweit sie zum SOEP-Repertoire gehören, erhältlich. Es können Tiefenanalysen zu den familiären, sozioökonomischen und sozialstrukturellen Charakteristika

der Eltern, welche, so die Theorie, einen Einfluß auf die Vergabe der Vornamen haben, durchgeführt werden. Der abgedeckte Analysezeitraum beträgt bei der Verwendung dieser Teilanalysepopulation ca. 20 Jahre (1984 bis heute).

Jedoch sind auch Analysen über wesentlich mehr Personen und größere Untersuchungszeiträume möglich. Der Analysezeitraum kann, retrospektiv, über das Jahr 1984 hinaus, bis zurück ins Jahr 1888, dem Jahr der Geburt des ältesten SOEP Teilnehmers erweitert werden. Um diese Informationen aus dem SOEP zu extrahieren sind umfangreiche Umstrukturierungen nötig. Das Haushaltsdesign und die Längsschnittperspektive des SOEP kommen uns hierbei zu Gute.

Typ 2 Fälle:

Das SOEP wurde entwickelt, um Aussagen über die „heutige" Zeit zu treffen. Da wir u.a. an der Namensvergabe und deren Determinanten im Zeitverlauf interessiert sind, kann man durch Umstrukturierungen die Längsschnittdatenfiles in einen Querschnittsdatensatz transferieren, so daß jeder jemals Befragte einen „Fall" darstellt, und die zuordenbaren Längsschnittinformationen die Variablen. Durch das Geburtsjahr der Personen (Fälle) wird der Zeitpunkt definiert, zu welchem die Informationen relevant sind. „Perfekte" Elterninformationen liegen uns lediglich für die während der SOEP Teilnahme geborenen Kinder vor (Typ 1). Zu einer zweiten Befragungspopulation, welche für das Projekt durchaus relevant wird, gehören jene Fälle, die als „erwachsene Kinder im Haushalt der Eltern" befragt wurden. Um den Namen der Kinder, die bereits über 16 Jahre alt und somit selbst Befragungspersonen sind, mit den Informationen über ihre Eltern verknüpfen zu können, müssen also (mindestens) zwei Generationen in einem Haushalt leben. Der Zeitpunkt der Namenswahl (Geburt des Kindes) liegt dabei *vor* dem ersten Erhebungsjahr (ansonsten wären sie Typ 1 Fälle). Wenn also zwei Generationen in einem Haushalt leben, haben wir die zur Herstellung des Analysezusammenhangs nötige Angabe des „Namen des Kindes" sowie eine Vielzahl von Informationen über zumindest ein Elternteil, welches ebenfalls als Befragter im SOEP auftaucht. Die Beschränkung, welche die Analysepopulation des Typs 2 charakterisiert, liegt darin, daß sich die vorhandenen Informationen nicht direkt auf den Zeitpunkt der Namensvergabe beziehen. Daraus ergeben sich einige Limitationen, bspw. bezüglich einer Analyse von schichtspezifischen Einflüssen auf die Namensvergabe, beziehungsweise all jener Merkmale, die sich nach der Geburt des Kindes geändert haben könnten und nicht durch die SOEP-Biographiefragebatterie abgedeckt werden. Dennoch können relativ problemlos

Informationen analysiert werden, die als zeitinvariant zu behandeln sind, sowie, mit Abstrichen, auch zeitvariierende Variablen, da der zeitliche Abstand zwischen dem Jahr der Erhebung der zu benutzenden Elterninformationen und dem Jahr der Namensvergabe in jeweiligen Variablen definiert werden kann. Man muß hier jeweils theoretisch begründen, warum man ein gewisses „time lag“ vertreten kann.

Als Besonderheit sind im SOEP einige Haushalte identifizierbar, in denen Eltern, Kinder und Enkelkinder zusammenleben. Diese kleine Gruppe kann Gegenstand besonderer Analysen, etwa der intergenerationalen Weitergabe von Vornamen sein.

Typ 3 Fälle:

Die dritte Gruppe, die „Typ 3“ Fälle der SOEP Analysepopulation besteht aus einer großen Zahl von Personen, genauer gesagt aus „Allen“, die jemals mit dem SOEP in Berührung gekommen sind, insofern sie nicht Teil der Analysepopulationen Typ1 und Typ2 sind. Die Kategorie Typ 3 weist gleichzeitig die größten Limitationen auf. Hier stehen die Vornamen, das Geburtsjahr, die Nationalität und lediglich retrospektiv erhobene Elterninformationen (Kindern wurden über ihre Eltern befragt) zur Verfügung. Die Eltern dieser Befragungspersonen nahmen niemals aktiv an einer Befragung des SOEP teil. Es wurden jedoch im Rahmen der „Biographiebatterie“ Basismerkmale über die Eltern erhoben (z. B. Beruf des Vaters als man 15 Jahre alt war). Für eine Vielzahl von Forschungsfragen bietet auch diese Gruppe der Befragten ausreichende Informationen. Will man jedoch Hintergrundvariablen der Eltern in der Analyse berücksichtigen, so sind diese nur sehr begrenzt vorhanden und beziehen sich nicht auf den Zeitpunkt der Vergabe der Vornamen.

Tabelle 4: Übersicht über die Analysemöglichkeiten von SOEP Personengruppen nach technischer Verfügbarkeit

Typ I (Kategorie A)	Typ II (Kategorie B)	Typ III (Kategorie C)
Befragte Eltern von Kindern, die zwischen 1983/84 und heute in „SOEP-Haushalte“ geboren wurden.	Befragte, die Teil des SOEP sind, jedoch als „Eltern im Haushalt“ von Kindern, die vor 1984 geboren wurden, enthalten sind. Man erfährt den Namen zumindest eines Kindes, wenn dieses im HH lebt und damit zu den Befragten des SOEP gehört.	Alle weiteren Personen im SOEP
N: 4687 (Stand 2004)	N: 15036 (Stand 2004)	N.: 34904 (Stand 2004)
Alle relevanten Elterninformationen sind vorhanden	Eingeschränkte Analysemöglichkeiten bezüglich Elterninformationen	Sehr eingeschränkte Analysemöglichkeiten bezüglich Elterninformationen
Probleme: • Sicherstellung der leiblichen Elternschaft (relevant bei mehreren Kindern innerhalb einer Familie mit einem unterschiedlichem Elternteil)	Probleme: • Zeitpunkt der Namensvergabe liegt außerhalb des SOEP -„Universums“. Die vorhandenen Informationen über die Eltern können dadurch nicht exakt auf den eigentlichen Namensvergabeprozeß angewendet werden. • Stiefelternschaften	Probleme: • Es sind bestenfalls Basisinformationen über Eltern vorhanden, welche die Namensvergabe vornahmen. (bspw. der Beruf der Eltern als die Person, deren Namen wir kennen 15 Jahre alt war.) • Stiefelternschaften
Abgedeckter Analysezeitraum: 1984 - derzeit	Abgedeckter Analysezeitraum: 1967-1984: (Namensträger: minderjährige Kinder von Befragungspersonen) sowie vor 1967: Namensträger sind die Befragungspersonen, deren Eltern im Haushalt leben und damit auch Befragungspersonen sind. Unter Umständen erstreckt sich der Analysezeitraum bis heute, nämlich dann, wenn in verschiedenen Wellen die Anzahl der Befragten erhöht wurde und Familien mit Kindern dazukamen, die vor der Erstbefragung geboren wurden.	Abgedeckter Analysezeitraum: Ende des 19. Jahrhunderts bis derzeit

4. 2 Die Generierung von geburtszeitpunktbezogenen Elterninformationen in „neuen“ Variablen

Zur Durchführung verschiedener Analysen im Rahmen des Projektes wurden verschiedene Arbeitsdatensätze und neue Variablen gebildet. Dies ist nötig, um die Informationen über die Eltern mit denen über die Kindern zum „richtigen“ Zeitpunkt (wie oben beschrieben) sinnvoll zu verknüpfen und die so gewonnenen Informationen in statistische Modelle und Berechnungen einfließen zu lassen. Für verschiedene Fragestellungen werden, auch

angesichts der unterschiedlichen „Qualität“ (Typ 1-3) der Informationen, verschiedene Arbeitsdatensätze nötig. Generell orientiert sich die Generierung der Arbeitsdatensätze und Variablen jeweils an den im folgenden exemplarisch beschriebenen Schema: Die relevanten Informationen über die Eltern der Namensträger werden aus den jeweiligen Teildatensätzen der SOEP Datenbank herausgefiltert und aus verarbeitungstechnischen Gründen in vier Arbeitsdatensätzen zusammengefaßt, die im folgenden, versehen mit einem Code über ihre Zugehörigkeit (Analysepopulationen Typ1-3), wieder zusammengeführt werden können.

Tabelle 5: Arbeitsdatensätze mit vier unterschiedlichen Informationsqualitäten: technische Definition

1.	*namestyp1.dta:* Enthält alle Personen, welche in einen existierenden Panel-Haushalt hineingeboren wurden, oder bei dem der Panel-Haushalt im Jahr nach der Geburt entstanden ist. Von diesen Personen liegen normalerweise vollständige Informationen beider Elternteile zum Zeitpunkt der Geburt vor.
2.	*namestyp2.dta:* Enthält alle Personen, welche zumindest in einer Panel-Welle die Haushalts-Stellung ”Kind“ innehatten, und die nicht bereits Teil des Typ-1- Datensatzes sind. Nicht verwendet wurden allerdings ”Pflegekinder“, da deren Vornamen normalerweise nicht von den Pflegeeltern vergeben werden. Von den Typ-2 Beobachtungen liegen in der Regel keine Eltern-Informationen zum Zeitpunkt der Geburt vor.
3.	*namestyp3.dta:* Enthält alle Personen, welche nicht bereits Typ-1 oder Typ-2 Beobachtungen sind. Von diesen Beobachtungen liegen lediglich die Angaben der Befragten selbst über ihre Eltern vor.
4.	*names.dta:* Enthält alle Personen.

Die Teildatensätze namestyp* enthalten immer nur diejenigen Variablen, welche für den jeweiligen Typ auch tatsächlich gebildet werden konnten.

- Variablen über die Person selbst
- Zeitinvariate Variablen über die Eltern (Elternangabe)
- Zeitvariierende Variablen zum Zeitpunkt der Geburt (Elternangabe)
- Proxy-Variablen für die zeitvariierenden Variablen aus dem ”ersten Beruf“ (Elternangabe)
- Proxy-Variablen für die zeitvariierenden Variablen aus der ”ersten erreichbaren Angabe“ (Elternangabe)
- Proxy-Variablen für die zeitvariierenden Variablen aus den Angaben der Personen über Ihre Eltern (Kindangabe)
- Längsschnittinformationen aus allen erreichbaren Panelwellen über die Eltern (Elternangabe)

Die Benennung der Variablen folgte dabei einem einheitlichen Schema. Bei diesem Schema wurden zunächst die zeitinvariaten und die zeitvariierenden Variablen unterschieden.[8]

[8] Hierbei handelt es sich um eine Auswahl, weitere Indikatoren wären möglich.

Tabelle 6: Die Variablen und ihre Bezeichnungen:

Zeitinvariat

Hhnr	Unveränderl. Haushaltsnummer
Plsample	Stichprobenart
Sex	Geschlecht
Gebjahr	Geburtsjahr
Gebland	Geburtsland
Sozland	Land letzter Schulabschluss
Sozbul	Bundesland letzter Schulabschluss
Konf	Index Konfession
Meanrel	Religiosität
Varrel	Variabilität Religiosität

Zeitvariierend

Hnr	Haushaltsnummer	pin	Politisches Interesse
Ise	ISEI-Status	nbi	Kein Berufsabschluss
Pip	Parteiidenifikation Partei	aut	Autonomie
Net	Befragungsstatus	zul	Zufriedenheit Leben
Mag	Maginitude Prestigi	asb	Schulabschl. im Ausland
Pii	Intensität Parteiidentifikation	ied	ISCED-Klassif.
Nat	Nationalität	bhi	HH-Brutto-Einkommen
Bra	(NACE) Branche	abb	Berufsabschl. im Ausland
Zug	Zufriedenheit Gesundheit	cas	CASMIN
Sbi	Schulbildung	nhi	HH-Netto-Einkommen
Tre	Treiman Prestigi	fam	Familienstand
Zua	Zufriedenheit Arbeit	bst	Berufl. Stellung
Bbi	Berufl. Bildung	bul	Bundesland
Egp	EGP-Klassen	isc	ISCO-1988
Zuw	Zufriedenheit Wohnung	pid	Parteiidentifikation j/n
Hbi	Hochschulbildung	est	Erwerbsstatus
Occ	Berufsklassen (DeStatis)		

Allen Variablen mit Elterninformationen schließt sich dann ein Kürzel für den Elternteil an, wobei **m** für die Mutter und **v** für den Vater steht. sbim bezeichnet mithin die Schulbildung der Mutter, sbiv die Schulbildung des Vaters.

Letzter Bestandteil des Variablennamens ist schließlich ein bis zu vier Zeichen langes Kennzeichen über den inhaltlichen Status der Variable. Diese Kennzeichen haben folgende Bedeutung:

Tabelle 7: Inhaltlicher Status der Variablen

(Kein Kennzeichen)	Angabe zum Zeitpunkt der Geburt
1st	Proxyvariable aus den Angaben zum ersten Beruf
one	Erste erreichbare Angabe nach der Geburt
bio	Angabe des Kindes über Ihre Eltern
1984, 1985 . . . 2002	Längsschnittinformation aus dem angegebenen Jahr

In bestimmten Fällen konnten aus den Proxy-Variablen die entsprechenden Angaben zum Zeitpunkt der Geburt gebildet werden: Dann nämlich, wenn ein Elternteil in der ersten erreichbaren Panelwelle angab noch immer im ersten Beruf zu sein, und das Geburtsjahr des Kindes zwischen dem Jahr des Berufseintritts und der ersten erreichbaren Beobachtung lag.

Nur in diesen Fällen wurde die Angabe zum ersten Beruf bereits in die entsprechende Variable zum Zeitpunkt der Geburt eingesetzt. In allen anderen Fällen wurde auf eine Entscheidung über die für diverse Analysen tatsächlich zu verwendende Proxy-Variable verzichtet. Dies muß man inhaltlich/theoretisch von Fall zu Fall entscheiden.

Weiterhin sei erwähnt, daß in einigen wenigen Fällen (ca. 70) eine Angabe zu einer Variable zum ”Zeitpunkt der Geburt“ auch bei denjenigen Variablen vorhanden sein kann, bei denen keine Proxy-Variable gebildet wurde. Dies betrifft insbesondere die subjektiven Indikatoren wie Lebenszufriedenheit oder Parteineigung. Ursache hierfür ist, daß die Informationen über die Identität der Eltern aus unterschiedlichen Quellen stammen können. Insbesondere kann es sein, daß der Pointer auf den Vater aus einer Welle stammt, welcher nicht der Welle der ersten Erfassung eines Kindes entspricht. In diesem Fall kann einer Typ-2-Beobachtung die Information einer Person zugespielt werden, der erst später als Vater des Kindes in Erscheinung getreten ist, die Information sich aber trotzdem auf den Zeitpunkt der Geburt beziehen. Dies kann mit Hilfe der *lag*-Variablen vermieden werden.

Die Erzeugung der sogenannten Eltern-Pointer (Verknüpfung der Kinder mit ihren Eltern) erscheint auf den ersten Blick trivial, ist es aber nicht. Auch hier gibt es Unterschiede je nach Qualität der Informationen. Bei den Typ-1-Beobachtungen wird als Mutter diejenige Person identifiziert, die in der Datei „biobirth“ als Mutter aufgeführt ist. Findet sich keine Information in „biobirth“, so wird statt dessen eine Information in den „*****kind“-Datensätzen gesucht. Findet sich hier keine Information zum Zeitpunkt der Geburt, so wird die nächste erreichbare Information verwendet. Vater ist diejenige Person, welche zum Zeitpunkt der Geburt in den „*kind“-Datensätzen als ”Partner der Mutter“ bezeichnet wird. Findet sich keine Information zum Zeitpunkt der Geburt, so wird die nächste erreichbare Information verwendet. Das Problem dieses Vorgehens ist, daß die Person, welche in einer späteren Welle als ”Vater“ oder als ”Mutter“ bezeichnet wird, nicht dieselbe Person sein muß, wie diejenige, die es zum Zeitpunkt der Geburt war. Die Variablen „lagmnr“ bzw. „lagmpnr“ enthalten daher Informationen über die Zeitdifferenz zwischen dem Geburtsjahr und der Pointer-Information. Dabei ist zu beachten, daß die Zeitdifferenz zwischen dem Geburtsjahr und dem Zeiger auf die Mutter auf Null gesetzt wurde, wenn die Information aus dem Datensatz „biobirth“ stammt. Man kann mit Hilfe der Lag-variablen die Auswertung auf diejenigen Beobachtungen begrenzen, bei deren Eltern es sich sicher um dieselben handelt, wie diejenigen zum Zeitpunkt der Geburt.

Bei Typ-2-Beobachtungen wurde als Mutter diejenige Person als Mutter definiert, die in der Datei „biobirth“ als Mutter aufgeführt wird. Fehlt diese Information, so wurde auf folgendes Verfahren zurückgegriffen, welches auch für die Väter angewandt wurde:
1. Beschränkung auf diejenigen Haushalte, in denen zumindest eine Person ”Kind“ des Haushaltsvorsitzenden oder dessen Lebenspartners ist (Werte 3 oder 12 auf der Variablen „?stell“ der Datensätze „?pbrutto“).
2. Mütter sind die weiblichen Haushaltsvorsitzenden oder die weiblichen Lebenspartner des Haushaltsvorsitzenden. Väter sind die männlichen Haushaltsvorsitzenden oder die männlichen Lebenspartner der Haushaltsvorsitzenden. In 3 Haushalten (mit Kindern) haben der Haushaltsvorsitzende und der Lebenspartner des Haushaltsvorsitzenden gleiches Geschlecht.

Für diese Haushalte wurde kein Vater bzw. Mutter-Pointer generiert. Es ist zu beachten, daß die Information über den Eltern-Pointer bei Typ-2-Beobachtungen definitionsgemäß nicht aus dem Jahr der Geburt stammen können. Es wurde darum wie bei den Typ-1-Beobachtungen eine Lag-Variable gebildet, welche die Zeitdifferenz zwischen dem Geburtsjahr und dem Zeiger abbildet. Dabei ist zu beachten, daß die Zeitdifferenz zwischen dem Geburtsjahr und dem Zeiger auf die Mutter auf Null gesetzt wurde, wenn die Information aus dem Datensatz „biobirth“ stammt.

Für die Typ-3-Beobachtungen wurden ausschließlich die Angaben der Kinder über Ihre Eltern verwendet. Diese sind im Datensatz „bioparen“ abgelegt. Die Konstruktion von Pointern erübrigt sich für diese Beobachtungen.

Angesichts der unterschiedlichen Qualitäten der Informationen über die Elterncharakteristika ist es nicht verwunderlich, daß man im Gesamtdatensatz bei einigen Variablen auf eine hohe Anzahl an „Missings“ stößt. Um das Arbeiten mit „Missings“ zu erleichtern, wurde eine einheitliche Klassifizierung angestrebt:

Tabelle 8: Vercodung fehlender Werte

-1	Keine Angabe (Angabe verweigert)
-2	Trifft nicht zu (z.B. wg. Fragebogenfilter)
-3	Angabe offenbar Fehlerhaft (lt. SOEP)
-4	Non-match bei der Längsschnittinformation
-5	Angaben zum Elternteil nicht vorhanden

4.3. Probleme, Limitationen und Besonderheiten im Umgang mit umstrukturierten SOEP Daten

Die vier gebildeten Datensätze (und damit der Gesamtdatensatz) sind in ihrer Beschaffenheit also ein umstrukturiertes Teil-SOEP. Dabei haben wir uns für ein „weites Format" entschieden, d. h. die „Namensträger" sind die „Fälle" und die Angaben über die „Namensvergebenden" sind die Variablen. Es handelt sich um einen Querschnittsdatensatz. Die Angaben, die wir über die „Namensvergeber" haben, die sich nicht auf den direkten Zeitpunkt der Namensvergabe beziehen, wurden dennoch als Zusatzinformationen beibehalten. Man kann sich vorstellen, daß dies den Umfang der Datenmatrix leicht in Regionen von über 1400 Variablen je Befragten anschwellen läßt. Ein Großteil dieser Variablen wird wahrscheinlich nie benutzt werden, so daß ein zusätzlicher, handlicher Datensatz – „master.dta" - gebildet wurde, welcher nur die relevanten und gesicherten Informationen beinhaltet.

Nichts desto trotz bieten die umfangreicheren Datensätze eine Kontrollmöglichkeit zur Überprüfung untypischer Angaben. Außerdem ist jederzeit die Verknüpfung mit den Standard-Datensätzen des SOEP möglich. Somit wird eines der datentechnisch bedingten Probleme zumindest teilweise überprüfbar: die Validität der Informationen, welche durch eine mögliche Diskrepanz zwischen Erhebungszeitpunkt und den tatsächlichen Umständen zum Zeitpunkt der Vergabe der Namen zustande gekommen sein kann (nur bei Typ-2 und Typ-3 Beobachtungen). Außerdem bieten sie die Informationen für Spezialanalysen, beispielsweise über zeitvariierende Indikatoren.

Ein weiteres datentechnisches Problem liegt in der Sicherstellung der leiblichen Elternschaft. Dies kann bei Typ-2 und insbesondere bei Typ-3 Beobachtungen Probleme bereiten. Stiefelternteile, die bspw. kurz nach der Geburt des Kindes an die Stelle des leiblichen Elternteils traten, werden gemeinhin als Eltern klassifiziert, obwohl sie im Sinne bestimmter Forschungsfragen die notwendige Bedingung des Einflusses auf die Vergabe des für ein Kind gewählten Vornamens nicht erfüllen. Dieses Problem läßt sich nicht beheben. Man kann jedoch davon ausgehen, daß diese Fälle eher die Ausnahme als die Regel sein werden. Auch Typ-1 Beobachtungen weisen unter Umständen ein ähnliches Problem auf; man hat jedoch die Möglichkeit, Fehler durch die Verwendung von Biographieangaben weitestgehend auszuschließen. Des weiteren ist es sinnvoll anzunehmen, daß der zum Zeitpunkt der Geburt

im Haushalt lebende Mann, auch wenn er nicht der leibliche Vater des Kindes ist, einen Einfluß auf die Vergabe des Namens hatte. Dies ist bei Typ-1 Beobachtungen immer möglich. Mütter können in der Regel sicher zugewiesen werden.

5 Abschließende Bemerkungen

Es sei hier nochmals darauf hingewiesen, daß die hier erläuterten Vercodungen der Vornamen und die Prozeduren zur Umstrukturierung der SOEP Daten im Rahmen des DFG-Projektes „Vornamen im Wandel“ durchgeführt wurden. Die Erstellung des Datensatzes ist durch die spezifischen Fragestellungen des Projekts angeleitet. Die erläuterten Prozeduren können als „Leitfaden“ für zukünftige Auswertungen der Vornamen des SOEP dienen, erheben jedoch keinen Anspruch auf Vollständigkeit und Applikationsfähigkeit auf andere als die bearbeiteten Fragestellungen.

Die sich aus den Umstrukturierungen ergebenden Arbeitsdatensätze sind temporär, d.h. sie werden auf der Grundlage von extra geschriebenen Computerprogrammen für die jeweiligen Analysezwecke generiert, jedoch nicht in der Standard-Datenbank des SOEP gespeichert. Die Replizierbarkeit der Analysen ergibt sich aus der Dokumentation der jeweiligen Programmierungen als STATA Do-files. Diese sind nicht Teil der allgemeinen SOEP Datenweitergabe.

Nach Abschluß des Projektes wird ein im Rahmen dieses Projektes entstandener Vornamensfile in die SOEP Datenweitergabestrategie integriert werden. Dieser Datensatz wird nicht den eigentlichen Vornamen des Befragten (dies ist datenschutzrechtlich nicht möglich), sondern pro Vornamen (Person) die bis zu vier vergebenen Codes (Regionenbezug und kulturhistorische Herkunft 1-3) enthalten. Da es sich bei diesen Codes gleichwohl um datenschutzrechtlich sensible Angaben handelt, werden Analysen mit diesen Variablen nur unter besonderen Datenschutzmaßnahmen möglich sein.[9]

[9] Der Zugang zu den Vornamen selbst, nämlich zur Generierung von anderen als den hier erarbeiteten Codes, wird neben den allgemeinen Zugangsvoraussetzungen gemäß Datenweitergabevertrag des SOEP zusätzlich mit weiteren Beschränkungen belegt sein. Ähnlich wie bei den SOEP „GEOcodes“ kann ein Zugriff nur auf gesonderten Antrag gewährt werden. Analysen mit den sensitiven Vornamens-Files werden aus Datenschutzgründen außerdem nur auf einem besonders gesicherten Rechner im DIW Berlin möglich sein.

Es ist geplant den „Vornamenscodesfile“ auch in Zukunft zu pflegen, d. h. bei den jährlichen Erweiterungen des SOEP-Samples alle neu hinzugekommenen Vornamen mit den entsprechenden Codes zu versehen.

Literatur

Ahmed, S. (1999): *A Dictionary of Muslim Names*. New York

Burkhauser, Richard V./Michaela Kreyenfeld/Gert G. Wagner (1997): The German Socio-Economic Panel - A Representative Sample of Reunited Germany and its Parts, *DIW-Vierteljahrsheftet*, Vol. 66

Gerhards, J. (2003): *Die Moderne und ihre Vornamen. Eine Einladung in die Kultursoziologie*, Wiesbaden. Westdeutscher Verlag.

Gerhards, J. (2005): *The Name Game. Cultural Modernization and First Names*. New Brunswick und London: Transaction Publishers.

HaiskenDeNew, J. P./Frick, J. R. (2005): Desktop Companion to the German Socio Economic Panel Study (DTC), http://www.diw.de/deutsch/sop/service/dtc/dtc.pdf.

Ilčev, S. (1996): *Rečnik na ličnite i familni imena u Bъlgarite*. Sofija

Merkle, L. and Zimmermann, K. F. (1992): "Savings, remittances, and return migration", *Economics Letters*, Vol. 38, 77-81.

Olivart, J. M. A. (1993): Dic*cionario de Nombres de Personas*. Universitat de Barcelona

Schimmel A. (1992): *Herr „Demirci" heißt einfach „Schmidt". Türkische Namen und ihre Bedeutung*. Önel Verlag Köln 1992

Schmidt, Ch. M. (1994): The Country of Origin, Family Structure and Return Migration of Germany's Guest-Workers. In: *Vierteljaheshefte zur Wirtschaftsforschung*, No. 1-2, 119-125.

Schupp, J. und G. G. Wagner (2002): Maintenance of and Innovation in Long-term Panel Studies: The Case of the German Socio-Economic Panel (GSOEP). In: *Allgemeines Statistisches Archiv*, Vol. 86(2), pp 163-175.

Seibicke, W. (1996-2001): *Historisches Deutsches Vornamenbuch. Bd. 1 (A-E), Bd. 2 (F-), Bd. 3 (-Sa), Bd. 4 (Sc-Z)*. Walter de Gruyter. Berlin. New York

Voges, W. / Schmidt, C. (1996): Lebenslagen die Zeit kosten - Zum Zusammenhang von sozialer Lage, chronischer Erkrankung und Mortalität im zeitlichen Verlauf. In: Zapf, W. / Schupp, J. / Habich, R. (Hrsg.), *Lebenslagen im Wandel: Sozialberichterstattung im Längsschnitt* (Sozio-ökonomische Daten und Analysen für die Bundesrepublik, Band 7, S. 378-401). Frankfurt am Main / New York: Campus.

Wagner, G. G. / J. Schepers (1989): Soziale Differenzen in der Lebenserwartung - Neue empirische Ergebnisse für die Bundesrepublik Deutschland, in: *Zeitschrift für Sozialreform*, 35. Jg. Heft 11/12, , S. 670-682

Weiterhin: die Personennamen-Datei in der Personnamen-Beratungsstelle der Universität Leipzig, die ständig aktualisiert wird.