German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2012-02 | JUNE 27, 2012

Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten

Rainer Schnell, Tobias Gramlich, Tobias Bachteler, Jörg Reiher, Mark Trappmann, Menno Smid, Inna Becher

Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten*

Rainer Schnell[1)], Tobias Gramlich[1)], Tobias Bachteler[1)], Jörg Reiher[1)], Mark Trappmann[2)], Menno Smid[3)], Inna Becher[4)]

[1)] Universität Duisburg-Essen

[2)] Institut für Arbeitsmarkt- und Berufsforschung

[3)] infas Institut für angewandte Sozialwissenschaft

[4)] Bundesamt für Migration und Flüchtlinge

27. Juni 2012

Abstract

The set of best methods for sampling migrant populations includes name-based sampling. So far this is done using either ad-hoc lists or onomastic dictionaries for the classification of names. This paper proposes a new name-based procedure, which uses a Bayes-classifier for the n-grams of the name. The new procedure is fault-tolerant of alternate spellings, and also allows the classification of names that are not found in dictionaries. It was tested using the names of about 1600 foreigners in the PASS panel. Finally, a CATI survey based on the new method in Hesse is described.

*Die Autoren danken Andreas Humpert und Klaus Schneiderhenze für hilfreiche Diskussionen früherer Manuskriptentwürfe. Das Verfahren und die Art des Tests wurde von Rainer Schnell vorgeschlagen und von Tobias Bachteler und Jörg Reiher umgesetzt. Mark Trappmann war an der Konzeption des Aufsatzes beteiligt, ermöglichte den Zugang zu den Daten und kommentierte unterschiedliche Fassungen des Manuskripts, Menno Smid leitete die CATI-Studie in Hessen. Die Berechnung der Tests und der erste Textentwurf erfolgte durch Tobias Gramlich, wobei Inna Becher eine Literaturdokumentation zur Verfügung stellte. Die finale Textversion stammt von Rainer Schnell.

Zusammenfassung

Zu den besten Verfahren für die Konstruktion von Migrantenstichproben gehören namensbasierte Stichproben. Hierfür werden bislang entweder ad-hoc-Listen oder namenskundliche Lexika für die Klassifikation von Namen verwendet. In diesem Beitrag wird ein neues Verfahren vorgeschlagen, das auf der automatischen Klassifikation eines Namens anhand der n-Gramme des Namens und der Anwendung des Bayes-Theorems basiert. Das neue Verfahren ist fehlertolerant gegenüber alternativen Schreibweisen und erlaubt auch die Klassifikation von Namen, die sich nicht in den Lexika finden. Das Verfahren wurde anhand der Namen der ca. 1600 Ausländer im PASS-Panel und einer CATI-Studie in Hessen untersucht.

1 Einleitung

Ausländer- und Migrantenpopulationen[1] in Deutschland sind für viele Forschungsbereiche von zentraler Bedeutung, z.B. für die Bildungs- und Arbeitsmarktforschung. Die in der Sozialforschung üblichen Stichprobenverfahren greifen aus zwei Gründen bei solchen Populationen nicht: Erstens stellen Migrantengruppen nur einen geringen Bevölkerungsanteil, daher sind die meisten Verfahren zur Ziehung von Zufallsstichproben hier ineffizient. Zweitens enthalten die meisten in der Praxis verwendeten Auswahlgrundlagen (Telefonbücher, Einwohnermelderegister oder Patientenregister) häufig keine aktuellen oder zugänglichen Informationen zur Zugehörigkeit zur Zielpopulation. Vor allem Personen mit Migrationshintergrund, die jedoch die deutsche Staatsbürgerschaft besitzen, lassen sich mit den zur Verfügung stehenden Auswahlgrundlagen kaum bestimmen. Daher müssen für diese Populationen spezielle Auswahlverfahren verwendet werden.

2 Stichprobenverfahren für Ausländerstichproben

Für Migranten in Deutschland liegen in der Regel keine vollständigen Auswahlgrundlagen zur Stichprobenziehung vor.[2] Daher müssen Stichprobenverfahren für seltene Populationen verwendet werden (Kalton 2009). Üblicherweise wird als obere Grenze einer „seltenen Population“ ein Anteil von maximal 10% an der Gesamtpopulation betrachtet

[1] Selbstverständlich sind die Begriffe „Ausländer“ und „Migrant“ nicht deckungsgleich; zu den damit verbundenen konzeptuellen Problemen siehe z. B. Galonska et al. (2004) und Diefenbach und Weiß (2006). Entsprechend ist die empirische Umsetzung der Konzepte nicht trivial, vgl. z. B. Statistisches Bundesamt (2011a). Die Bezeichnungen „Ausländer“ und „Migranten“ werden im Folgenden trotz dieser Probleme synonym verwendet, da für die Beurteilung jedes empirischen Verfahrens mindestens ein Referenzwert benötigt wird. Das im Folgenden vorgestellte Verfahren wird mangels anderer verfügbarer Daten anhand der Staatsangehörigkeit evaluiert. Für das technische Verfahren selbst spielt das zugrunde liegende Konzept keine Rolle: Verfügt man über konzeptuell andere Referenzdaten zum Training des Verfahrens, kann ein anderes Konzept eingesetzt werden.

[2] Lediglich das Bundesamt für Migration und Flüchtlinge (BAMF) hat im Rahmen seines Auftrages das von ihm selbst geführte Ausländerzentralregister (AZR) in einzelnen Fällen zur Stichprobenziehung verwendet (Babka von Gostomski 2008). Faktisch steht das AZR für keine praktischen Zwecke außerhalb des BAMF zur Verfügung.

(Kalton und Anderson 1986). Trotz eines Anteils an der Gesamtbevölkerung von insgesamt – je nach Definition – ca. 9.0% Ausländer oder 19.3% Personen mit Migrationshintergrund (Statistisches Bundesamt 2011a,b), stellen Ausländer in Deutschland daher in diesem Sinne eine „seltene“ Population dar.

2.1 Klumpenstichproben

Für seltene Populationen werden häufig Listen von Teilpopulationen als Auswahlgrundlagen verwendet. Dazu gehören z.B. Mitgliederlisten von Vereinen oder Organisationen. So konnte z.B. Rother (2010) für eine Ausländerstichprobe auf Verwaltungsdaten des Bundesamtes für Migration und Flüchtlinge (BAMF) zurückgreifen und eine Stichprobe aus allen Teilnehmern von Integrations- und Alphabetisierungskursen ziehen.

Oftmals klumpen Elemente der Zielpopulation geographisch stark: Dann ist die Dichte an Ausländern aus einzelnen Ländern in einzelnen geographischen Regionen so hoch („Little Italy“, „China Town“), dass sich Standardauswahlverfahren effizient auf diese Regionen anwenden lassen (siehe z.B. Blane 1977, Ecob und Williams 1991). So beschränkt sich z.B. die europäische Erhebung zu Minderheiten und Diskriminierung (EU-MIDIS) in einigen Mitgliedsländern auf Minderheitengruppen mit einem Mindestanteil von mindestens 5% an der Gesamtbevölkerung und vornehmlich auf Gebiete mit „mittlerer bis hoher Konzentration“ der Zielgruppen (European Union Agency for Fundamental Rights 2009, S. 18f). Das Problem dieser Verfahren besteht vor allem darin, dass sich die Personen in solchen Gebieten von Personen in Gebieten mit geringeren Konzentrationen in Hinsicht auf abhängige Variablen wie z.B. „Integration“ und „Diskriminierungswahrnehmung“ unterscheiden können.

2.2 Nicht auf Zufallsverfahren basierende Stichproben

Vor allem in Deutschland werden bei Stichprobenziehungen von Ausländern häufig Quotenstichproben verwendet, die die Auswahl mindestens auf einer Stufe letztendlich dem Interviewer überlassen. Ein bekanntes Beispiel für diese Art der Ziehung stellt die Publikation „Zuwanderer in Deutschland“ (Bertelsmann Stiftung 2009) dar. Gerade durch eine solche Auswahl durch den Interviewer dürften vor allem besser und gut integrierte Migranten erfasst werden. Allgemein sollte aufgrund der unkontrollierbaren Effekte bei Quotenauswahlen für wissenschaftliche Zwecke immer von Quotenauswahlen abgesehen werden (Schnell/Hill/Esser 2011:296-298).

Weiterhin werden auch Netzwerk- bzw. Schneeballverfahren zur Stichprobenziehung von Ausländern und Migranten eingesetzt, so z.B. für eine Teilstichprobe des SOEP von Spätaussiedlern aus der ehemaligen Sowjetunion (Burkhauser et al. 1997). Netzwerk-Stichproben leiden unter zwei Problemen: Die Auswahlwahrscheinlichkeiten sind in der Praxis kaum korrekt berechenbar und sozial isolierte Personen verfügen über geringere Auswahlwahrscheinlichkeiten. Netzwerkstichproben sind daher eher als allerletztes Mittel denn als Methode mit reproduzierbaren Ergebnissen anzusehen.

2.3 Screening

Oftmals liegen nur Auswahlgrundlagen für eine „Allgemeinbevölkerung“ vor (Schnell 1991). Einige Verfahren für die Auswahl seltener Populationen basieren darauf, dass ausgehend von einer solchen Liste alle Elemente einer Stichprobe daraufhin untersucht werden, ob sie zur Zielpopulation gehören. Dieses Vorgehen wird als „Screening“ bezeichnet. Für Stichproben von Ausländern können z.B. Einwohnermeldeamtsstichproben gezogen und dann Ausländer anhand ihrer registrierten Staatsangehörigkeit identifiziert werden (vgl. Granato 1999). Auch für die Ausländerstichprobe des SOEP (Stichprobe B) wurde aus Melderegistern gezogen (Haisken-DeNew und Frick 2005). Dieses Vorgehen eignet sich nicht für alle Zielgruppen. Für Personen, die die deutsche Staatsbürgerschaft besitzen, ist dieses Vorgehen des Screenens anhand der Nationalität nicht möglich. Salentin (2007) schlägt daher für Aussiedler vor, bei Einwohnermeldeamtsstichproben anstatt nach der Staatsangehörigkeit nach dem Geburtsort zu screenen.[3] Dieses Vorgehen eignet sich prinzipiell auch für alle anderen naturalisierten Migranten der ersten Generation. Dies gilt aber nicht für spätere Migrantengenerationen, also in Deutschland geborene Kinder oder Enkel der eingewanderten Migrantengeneration. Zwar wäre es möglich, in einer Befragung nach dem Geburtsland der Elternteile oder gar der Großeltern zu fragen, die Validität der Angaben dürfte aber eher fraglich sein. Empirische Studien hierzu scheinen nicht vorzuliegen.

2.4 Namensbasierte Verfahren

Häufig ist das eigentliche Screening-Merkmal nicht Bestandteil der Auswahlgrundlage, so dass Hilfsmerkmale verwendet werden müssen. In vielen Fällen stehen lediglich Namenslisten zur Verfügung. Namenslisten werden weltweit bei vielen Forschungsprojekten zu Migrantenpopulationen verwendet, da Namen auf die regionale oder ethnische Herkunft des Namensträgers hinweisen können. In der Praxis werden insbesondere oft namenskundliche Lexika von Namen mit bekannter Zugehörigkeit (d.h. ausschließlichem oder sehr häufigem Auftreten in der entsprechenden Gruppe) zu einem bestimmten Herkunftsland verwendet oder für diesen Zweck aufwendig erstellt.

In Deutschland hat sich das von Humpert und Schneiderheinze (2000) entwickelte onomastische Verfahren als Standardverfahren für sozialwissenschaftliche Stichprobenziehungen bei Migranten etabliert. Dabei werden Namen aus öffentlich zugänglichen Verzeichnissen anhand von ihnen erstellter Namensdatenbanken verschiedener Nationalitäten verglichen und entsprechend klassifiziert. Neuere Beispiele für die Anwendung dieses Verfahrens sind der Jahresbericht des Sachverständigenrates deutscher Stiftungen für Migration und Integration (SVR) (2010), der Integrationssurvey des Bundesinstituts für Bevölkerungsforschung BiB (Mammey und Sattig 2002, Haug und Swiaczny 2003) sowie die Studie „Muslime in Deutschland“ (MiD, Brettfeld und Wetzels 2007).

[3] Dies ist natürlich nur dann möglich, wenn die Weitergabe des Merkmals „Geburtsort“ zulässig wäre. Dies scheint nach der geltenden Rechtslage in Deutschland kaum möglich zu sein.

Prinzipiell ähnliche Verfahren werden in der Epidemiologie verwendet. Hierbei dominiert die Nutzung ad hoc zusammengestellter Listen häufiger oder typischer Namen einzelner Länder. Diese Listen werden in der Regel von muttersprachlichen Experten gesichtet und editiert. Die resultierenden Listen werden anschließend für das Screening einer allgemeineren Auswahlgrundlage verwendet. Ein entsprechendes Beispiel stellt die Arbeit von Halm und Sauer (2005, S. 43ff) dar. Anhand einer Liste mit rund 10.000 „typischer" Nachnamen und rund 7000 Vornamen klassifizieren sie Einträge im Telefonbuch. Ähnliche Verfahren sind in der Epidemiologie international verbreitet. So erstellen z.B. Schwartz et al. (2004) eine Liste häufiger arabischer Vor- und Nachnamen aus verschiedenen öffentlich zugänglichen Quellen[4] und sichten die resultierende Liste manuell, um schließlich eine Stichprobe aus einem Detroiter Krebsregister zu ziehen. Ähnlich verwendet Lauderdale (2006) eine Namensliste aller sozialversicherten Personen zur Erstellung einer Liste und zum Screening nach arabischstämmigen Frauen in kalifornischen Geburtsregistern.

3 Beschreibung des neuen Verfahrens

Das im Folgenden vorgestellte Verfahren basiert im Gegensatz zu allen bisher verwendeten Verfahren nicht auf der Klassifikation vollständiger Namen oder Namensendungen, sondern auf der Klassifikation von Buchstabenfolgen (n-Gramme).[5]

Den Ausgangspunkt des Verfahrens bilden dabei Listen, die für verschiedene Nationalitäten die jeweiligen empirischen Namenshäufigkeiten getrennt nach Vor- und Nachnamen aller Personen mit dieser Nationalität enthalten. Am geeignetsten für solche Listen sind Zensus- oder Sozialversicherungsdatenbanken.[6]

Aus dem Anteil eines bestimmten Namens $P_{(Name)}$, dem Anteil aller Personen mit einer bestimmten Nationalität $P_{(Nationalität)}$ und dem bedingten Anteil eines Namens unter Personen mit einer bestimmten Nationalität $P_{(Name|Nationalität)}$ kann mit Hilfe des Bayes-Theorems die bedingte Wahrscheinlichkeit für eine bestimmte Nationalität gegeben einen bestimmten Namen $P_{(Nationalität|Name)}$ berechnet werden:

$$P_{(Nationalität|Name)} = \frac{P_{(Name|Nationalität)} * P_{(Nationalität)}}{P_{(Name|Nationalität)} * P_{(Nationalität)} + P_{(Name|\neg Nationalität)} * P_{(\neg Nationalität)}}$$

[4] Verschiedene Register (z. B. Geburts- und Sterberegister), die Namen und entsprechend Geburtsländer enthielten; Mitgliederlisten entsprechender (z. B. Kultur-)Vereine und Vereinigungen; außerdem sammelten (arabisch sprechende) Mitarbeiter aus Telefonbüchern „typische" arabische Namen.

[5] Das hier beschriebene Verfahren wurde von den Autoren unter anderem auf der ESRA Konferenz in Lausanne 2011 und der ASA Konferenz 2011 in Tilburg vorgestellt. Das Verfahren wurde von Rainer Schnell, Tobias Bachteler und Jörg Reiher an der Universität Konstanz im November 2006 entwickelt und wurde von diesen in einer Reihe von Stichprobenziehungen und Record-Linkage-Projekten eingesetzt. Eine erste Beschreibung findet sich bei Schnell (2009).

[6] Shackleford (1998) und Lauderdale (2006) beschreiben solche Listen von Namenshäufigkeiten für verschiedene Nationalitäten nach Auszählungen des US-amerikanischen Zensus oder aller sozialversicherungspflichtig beschäftigten Personen in den USA.

Der Nachteil einer solchen Klassifikation besteht in der Verwendung vollständiger und korrekt geschriebener Namen.[7] In der Literatur zu Datenbereinigung in Datenbanken werden Schreibfehler bei ca. 20% der Fälle als typischer Wert betrachtet (Winkler 2009). Diese dürften sich bei Namen von Migranten konzentrieren, so dass mit besonders vielen Fällen zu rechnen ist, bei denen jedes Verfahren, das exakte Übereinstimmungen voraussetzt, Ausfälle produzieren wird. Entsprechend würden hierbei Anteilswerte systematisch unterschätzt. Daher ist ein fehlertolerantes Verfahren für die Namensklassifikation wünschenswert. Eine Möglichkeit besteht darin, für die Klassifizierung nicht die vollständigen Vor- oder Nachnamen zu verwenden, sondern die Namen in Buchstabenfolgen (n-Gramme) zu zerlegen und diese zur Klassifizierung der Namen zu verwenden. Solche n-Gramme werden in der Informatik für viele Probleme der Verarbeitung natürlicher Sprachen verwendet, so z.B. bei der Konstruktion von Suchmaschinen oder Programmen zur Rechtschreibprüfung.

Im Folgenden verwenden wir also die bedingten Häufigkeiten aller n-Gramme eines Vor- oder Nachnamens zur Berechnung der bedingten Wahrscheinlichkeit für eine Nationalität. Neben der Fehlertoleranz besitzt das Verfahren einen weiteren Vorteil: Es muss kein vollständiges Namenslexikon vorliegen. Es können also auch Namen klassifiziert werden, die nicht in einem Lexikon verzeichnet sind.

Tabelle 1 zeigt nun die Klassifikation am Beispiel des Namens „Peter“. Zunächst wird der Name „Peter“ (Länge 5 Buchstaben) in seine n-Gramm-Menge (hier Bigramme, n=2) aufgespalten. Der Name „Peter“ hat 4 n-Gramme der Länge 2: {PE,ET,TE,ER}. Da oftmals Wort- oder Namensanfänge und -endungen charakteristisch für verschiedene Sprachen sind,[8] empfiehlt sich das Anfügen von einem Leerzeichen vor und hinter die Namen, so dass zusätzliche n-Gramme am Namensanfang bzw. -ende entstehen:

$$\{\_\text{P,PE,ET,TE,ER,R}\_\}$$

Für jedes Land wird nun die Anzahl des Auftretens jedes einzelnen Bigramms dieser Bigrammmenge durch die Anzahl aller Namen[9] von Personen einer bestimmten Nationalität dividiert, und daraus für jedes Land dann das Produkt aus diesen relativen

[7] Diese Idee der Namensklassifikation durch das Bayes-Theorem scheint zum ersten Mal für eine Anwendung im US-Bureau of the Census benutzt worden zu sein: Passel und Word (1993) und Perkins (1993) beschreiben die Konstruktion einer „spanischen“ Namensliste für den US-amerikanischen Zensus 1980. Hierbei wurden mithilfe des Bayes-Theorems die bedingten Wahrscheinlichkeiten der Zugehörigkeit zu hispanischen Subpopulationen berechnet. Diese Zensus-Anwendung basiert auf vollständigen Namen; das Verfahren wurde in der Stichprobenliteratur für seltene Populationen nicht aufgenommen.

[8] Dies gilt auch für Namen: z. B. enden deutsche männliche Namen in der Regel nicht auf -a. Deutsche männliche Namen enden in der Regel auch nicht mit -e wohingegen in Italien Namen mit Endung -e oftmals die männliche Namensvariante sind (z. B. die italienischen Namen Simone (männlich) und Simona (weiblich)).

[9] Hier wird durch die Zahl der Namen dividiert und nicht etwa durch die Anzahl der n-Gramme. Dies hat sich bei unseren Anwendungen im Vergleich zu den Alternativen als vorteilhaft erwiesen.

Häufigkeiten der einzelnen n-Gramme des Namens „Peter“ gebildet.[10] Dieses Produkt wird anschließend mit dem Faktor w multipliziert.[11] Ein Name wird nun schließlich in dasjenige Land klassifiziert, für das diese Klassifikationsgröße maximal ist.

4 Trainingsdaten für das neue Verfahren

Zentraler Punkt des Verfahrens ist eine Namensliste, aus der diese n-Gramm-Häufigkeiten gewonnen werden können, d.h. ein Liste, die Namen und Häufigkeiten dieses Namens für alle (berücksichtigten) Nationalitäten enthält. Solche Listen sind – vor allem außerhalb Deutschlands – aus verschiedenen Quellen und in unterschiedlicher Qualität verfügbar. In Deutschland böten sich Listen auf der Basis der Einwohnermelderegister an; aufgrund der erforderlichen Kooperation vieler Gemeinden erscheint dieses Vorgehen in der Praxis eher schwierig. Eine naheliegende Alternative bestünde in der Verwendung von Telefonverzeichnissen, aus denen sich Namen und Häufigkeiten, hingegen aber nicht die zugehörige Nationalität der Personen gewinnen ließen. Werden hierbei Telefonverzeichnisse aus vielen verschiedenen Ländern verwendet, könnte man alle dort verzeichneten Namen als „einheimisch“ behandeln und die insgesamt geringe Wahrscheinlichkeit der unvermeidlichen Fehlklassifikationen durch Migranten akzeptieren oder sich auf die häufigsten Namen oberhalb je nach Land verschiedener Schwellenwerte beschränken.

Einfacher zu handhaben sind natürlich entsprechende Listen aus einer Datenquelle. Für die Entwicklung des hier vorgestellten Verfahrens wurden erstmalig nach Nationalität getrennte Listen von Vor- und Nachnamenshäufigkeiten aller in Deutschland sozialversicherungspflichtig beschäftigter Personen erstellt.[12] Diese Listen enthalten Namen und Häufigkeiten getrennt nach Staatsangehörigkeit. Aus Datenschutzgründen lagen Vor- und Nachname also nicht gemeinsam (z. B. „D – Peter Müller – n=5“), sondern nur getrennt („D – Peter – n=40 000“ und „D – Müller – n=1 000“) vor.

Insgesamt umfassten die Namenslisten für die hier klassifizierten Nationalitäten (bzw. Ländergruppen) 112 831 unterschiedliche Vor- und 493 974 unterschiedliche Nachnamen. Entsprechend gewichtet mit den jeweiligen Häufigkeiten entspricht dies jeweils rund 30 Millionen Personen. Die Tabelle 2 zeigt die Anzahl unterschiedlicher Namen der Namenslisten in den bei der Klassifikation berücksichtigten Ländergruppen.

[10] Der Beitrag eines „unbekannten“, nicht erfassten n-Gramms zur Klassifikation eines Namens wurde nicht auf 0 gesetzt, andernfalls würde der gesamte Ausdruck bei der Multiplikation zur Gesamtwahrscheinlichkeit natürlich auch 0. Bei unserer Anwendung wurde statt 0 der Faktor $\frac{1}{N}$ verwendet.

[11] Wenn $n_{Namen[j]}$ die Anzahl Namen des Landes j und $N = \sum n_{Namen[j]}$ die Anzahl Namen aller Länder bezeichnet, ergibt sich $w_{[j]} = \frac{n_{Namen[j]}}{N}$; das ist die Priorwahrscheinlichkeit aus dem Bayes-Theorem.

[12] Diese Liste wurde mit hohem Aufwand im Jahr 2005 auf Wunsch des Erstautors einmalig vom Forschungsdatenzentrum der Bundesagentur für Arbeit nach Rücksprache mit den Datenschützern der Bundesagentur unter der Auflage zur Verfügung gestellt, dass die Namenslisten nur getrennt für Vor- und Nachnamen erstellt wurden und jeder Name mindestens fünfmal in den Datenbanken erschien.

Tabelle 1: Klassifikation des Namens „Peter“[a)]

Nationalität	Bigramm	Anzahl Bigramme	Anzahl alle Namen/Land	p_{Land}	Πp_{Land}	Korrekturfaktor w	Klassifikations-größe (Land|“Peter“)
Deutschland	ER	5 004 637	28 309 791	0.1768	$8.09*10^{-8}$	0.9271	$7.50*10^{-8}$
	TE	2 689 490		0.0950			
	R+	2 477 673		0.0875			
	ET	1 750 568		0.0618			
	+P	929 959		0.0329			
	PE	767 277		0.0271			
LG Osteuropa	ER	17 968	254 788	0.0705	$8.42*10^{-9}$	0.0083	$7.03*10^{-11}$
	R+	14 392		0.0565			
	ET	12 271		0.0482			
	TE	10 936		0.0429			
	+P	10 828		0.0425			
	PE	6 134		0.0241			
eh. Jugoslawien	R+	49 621	723 561	0.0686	$2.79*10^{-10}$	0.0237	$6.61*10^{-12}$
	ER	37 698		0.0521			
	ET	32 209		0.0445			
	TE	9 597		0.0133			
	+P	9 154		0.0126			
	PE	7 567		0.0105			
Italien	ER	12 990	216 672	0.0600	$6.37*10^{-10}$	0.0071	$4.53*10^{-12}$
	PE	12 458		0.0575			
	+P	9 885		0.0456			
	ET	8 735		0.0403			
	TE	5 420		0.0250			
	R+	872		0.0040			
Türkei	ER	70 013	605 029	0.1157	$1.45*10^{-10}$	0.0198	$2.88*10^{-12}$
	ET	58 307		0.0964			
	R+	47 197		0.0780			
	TE	8 054		0.0133			
	+P	2 279		0.0038			
	PE	2 013		0.0033			
...	...	.....	.....	...	...	...	...
Summen		14 221 676	30 535 580				

[a)] Die hier nicht ausgewiesenen Namen aus Griechenland und der ehemaligen Sowjetunion wurden in den Berechnungen berücksichtigt.

Tabelle 2: Anzahl der Vor- und Nachnamen in den Trainings-Namenslisten

	Vornamen		Nachnamen	
Staatsangehörigkeit	Namen	Personen	Namen	Personen
Deutschland	58 757	28 309 791	383 592	27 551 167
Jugoslawien[a]	10 494	313 193	21 973	262 425
Türkei	8 137	605 029	20 835	587 175
Osteuropa[b]	2 750	103 300	7 273	47 366
Italien	2 707	216 672	14 334	180 118
Griechenland	2 517	112 744	9 452	75 732
Russland[c]	1 952	47 638	2 476	13 187
restliche Welt	25 517	470 446	34 039	286 887
Insgesamt	112 831	30 178 813	493 974	29 004 057

[a] ehemaliges Jugoslawien und Nachfolgestaaten

[b] Polen und osteuropäische Nachbarländer

[c] Russland und Staaten der ehemaligen Sowjetunion

Unter „Russland“ wurden alle Einträge von Namen und Häufigkeiten von Staaten der ehemaligen Sowjetunion und der ehemaligen Russischen Föderation zusammengefasst.[13] Unter „Jugoslawien“ wurden Namen aus allen jugoslawischen Nachfolgestaaten zusammengefasst.[14] Um die Fallzahlen für osteuropäische Länder zu erhöhen, wurden Polen und mehrere osteuropäische Nachbarländer zu einer Ländergruppe „Osteuropa“ zusammengefasst.[15] Für die im folgenden beschriebene Validierung des Verfahrens wurde auf einen Paneldatensatz zurückgegriffen, der trotz seiner Größe von nahezu 19 000 Personen pro Welle lediglich für die in Deutschland häufigeren Ausländergruppen ausreichend große Fallzahlen aufweist. Daher beschränken wir uns in dieser Arbeit auf die größten Gruppen von Ausländern in Deutschland.[16] Die Abbildung 1 zeigt zusammenfassend das Verfahren zur Erstellung der Trainingsdatenbank.

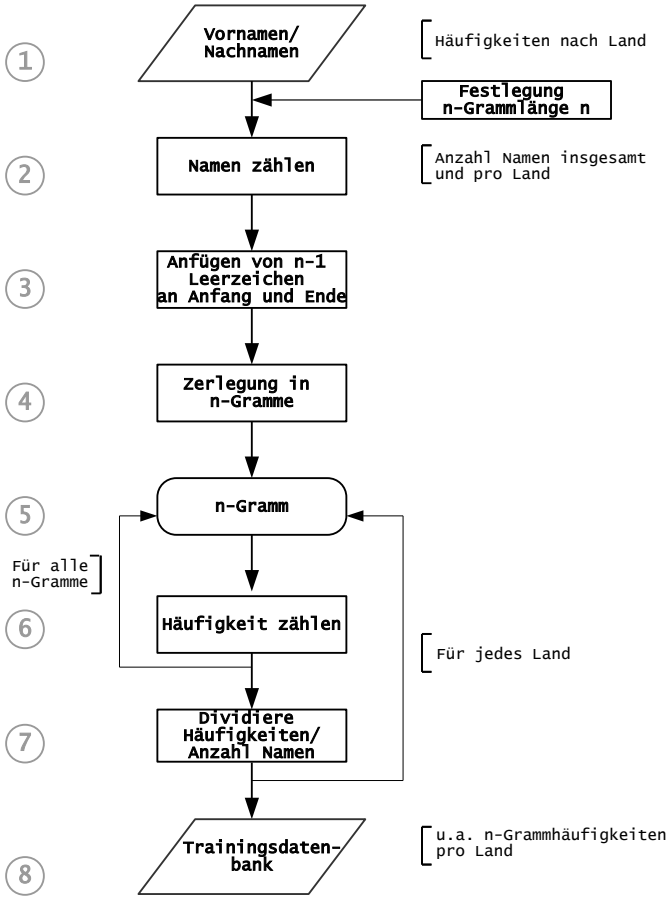
[13] Hierzu gehören in diesem Datensatz Estland, Lettland, Litauen, Sowjetunion, Russische Föderation, Ukraine, Weißrussland, Armenien, Aserbaidschan, Kasachstan, Kirgisistan, Tadschikistan und Turkmenistan.

[14] Hierzu gehören Bosnien und Herzegowina, Jugoslawien, Kroatien, Makedonien, Albanien und Slowenien. Die ehemalige Bundesrepublik Jugoslawien war zum Zeitpunkt der Erstellung der Namensliste noch nicht in Serbien und Montenegro zerfallen.

[15] Hierzu gehören Polen, Bulgarien, Ungarn, Rumänien, ehemalige Tschechoslowakei, Slowakei und die Tschechische Republik.

[16] Diese Beschränkung resultiert allein aus der Größe der Subgruppen im Validierungsdatensatz. In anderen Anwendungen kann eine andere – oder auch gar keine – Zusammenfassung der in den Trainingsdaten erfassten Länder zu Gruppen gewählt werden.

Abbildung 1: Schematische Darstellung des Verfahrens zur Erstellung der Trainingsdatenbank



5 Validierung des neuen Verfahrens

Das oben beschriebene Klassifikationsverfahren wurde anhand zweier Datensätze untersucht: Dem Paneldatensatz PASS des IAB und einer Stichprobe "Kriminalitätsfurcht in Hessen" in Zusammenarbeit mit dem Institut für angewandte Sozialwissenschaft (infas, Bonn). Die Ergebnisse beider Studien werden im Folgenden vorgestellt.

5.1 Das Panel „Arbeitsmarkt und Soziale Sicherung" PASS

Das „Panel Arbeitsmarkt und Soziale Sicherung" (PASS) (Promberger 2007, Trappmann et al. 2010) ist eine vom Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA) im Jahr 2007 neu gestartete Haushaltsbefragung.

Das PASS–Panel besteht aus zwei Teilstichproben: Stichprobe 1 besteht aus Haushalten aktueller Leistungsbezieher nach dem SGB II; sie wird zu jeder Welle um Neuzugänge in den Leistungsbezug aufgefrischt. Stichprobe 2 ist eine disproportional geschichtete allgemeine Bevölkerungsstichprobe. In den ausgewählten Haushalten waren alle Personen ab 15 Jahren zu interviewen, Personen über 65 Jahren erhielten einen verkürzten Fragebogen. Die Befragung erfolgte falls möglich telefonisch (CATI), bei Bedarf auch persönlich durch Interviewer vor Ort (CAPI).

In den Räumlichkeiten des IAB wurden die vom Datensatz getrennten Vor- und Nachnamen der Respondenten der ersten und zweiten Welle mit dem beschriebenen Klassifikationsverfahren klassifiziert.[17] Insgesamt lagen aus den ersten beiden Wellen 18 795 Vor -und Nachnamen von Teilnehmern zur Klassifikation vor.

Tabelle 3 zeigt die Häufigkeiten der hier berücksichtigten Länder im PASS. Die insgesamt 1610 Personen mit ausländischer Staatsangehörigkeit stammen vor allem aus der Türkei (31%), den Staaten der ehemaligen Sowjetunion (15%) und dem ehemaligen Jugoslawien (10%). Nur insgesamt 9% der PASS-Respondenten haben eine ausländische Staatsangehörigkeit; der Anteil der Personen, die außerhalb Deutschlands geboren sind, liegt mit 17% deutlich höher. Personen aus der ehemaligen Sowjetunion stellen hier mit 23% den größten Anteil an den im Ausland geborenen Personen dar, gefolgt von in der Türkei (rund 18%) oder in den osteuropäischen Nachbarländern geborenen Personen (15%).[18]

[17] Alle Arbeiten an und mit den Namen haben im IAB in Nürnberg stattgefunden; die Namen haben das IAB nicht verlassen, lediglich die aus der Klassifikation der Namen resultierende „geschätzte Nationalität". Die Namen wurden lediglich zur Klassifikation verwendet; die Autoren hatten zu keinem Zeitpunkt Zugriff auf die Namen.

[18] Nur rund 44% der außerhalb Deutschlands geborenen Personen haben auch eine ausländische Staatsangehörigkeit, demgegenüber sind rund 83% der Personen mit ausländischer Staatsangehörigkeit nicht in Deutschland geboren. Die Übereinstimmung von Nationalität gegeben das Geburtsland ist für Personen aus dem ehemaligen Jugoslawien am höchsten (88%), gefolgt von Personen, die in Russland (78%) oder der Türkei (71%) geboren wurden. Die Übereinstimmung des Geburtslands gegeben die Nationalität ist für osteuropäische Staatsangehörige am höchsten (93%), gefolgt von griechischen (90%), italienischen und türkischen Staatsangehörigen (jeweils 74%).

Tabelle 3: Anzahl Personen, Nationalitäten im PASS, Welle 1 und 2

	Staatsangehörigkeit		Geburtsland	
	Personen	in %	Personen	in %
Deutschland	17 140	91.2	15 757	83.8
Türkei	514	2.7	530	2.8
Russland[a]	246	1.3	697	3.7
Jugoslawien[b]	163	0.9	170	0.9
Osteuropa[c]	114	0.6	444	2.4
Italien	68	0.4	49	0.3
Griechenland	45	0.2	33	0.2
restliche Welt	460	2.5	1 091	5.8
unbekannt	45	0.2	24	0.1
Insgesamt	18 795	100.0	18 795	100.0

[a] einschließlich ehemaliger Staaten der Sowjetunion

[b] einschließlich Nachfolgestaaten

[c] Polen und osteuropäische Nachbarländer

5.2 Probleme der Klassifikation der Staatsangehörigkeit und des Migrationsstatus

Zur Evaluierung der Namensklassifikation wird ein Vergleichsstandard benötigt, mit dem die automatische Klassifikation verglichen werden kann. Dazu stehen aus dem PASS prinzipiell mehrere Alternativen zur Verfügung: die selbstberichtete Staatsangehörigkeit, das selbstberichtete Geburtsland, gegebenenfalls das Geburtsland der Eltern und Großeltern. Die Wahl jedes dieser Kriterien wäre mit Problemen behaftet (vgl. zusammenfassend Diefenbach/Weiß 2006). Im Zentrum des Interesses steht hier aber lediglich das Ausmaß der empirischen Übereinstimmung zwischen diesen Kriterien für eine Klassifikation im Rahmen eines Screenings, dem detailliertere Analysen folgen könnten. Die Bewertung des Klassifikationsverfahren muss sich nach dem verwendeten Konzept der Trainingsdaten richten: Dieses ist hier in Ermangelung anderer Datenquellen Staatsangehörigkeit.[19]

Abgesehen von den unvermeidlichen inhaltlichen Problemen jedes einzelnen Kriteriums für Migrationsstatus bzw. Staatsangehörigkeit entsteht ein Problem durch die Veränderung der Staatsangehörigkeit oder des Namens im Laufe der Zeit. In Deutschland werden mit sinkender Tendenz etwas mehr als 1% der Ausländer jährlich naturalisiert. Daher

[19] Da elaboriertere Konzepte Messungen der Einzeldimensionen erfordern, bedürfte man eines Surveys, bei dem neben den Namen diese Einzeldimensionen an einer sehr großen Stichprobe erhoben werden müssten, wenn man auch seltene Populationen klassifizieren möchte. Datensätze dieser Größe können in der Bundesrepublik nur durch die amtliche Statistik erhoben werden, wobei dabei aber dann prinzipiell die Weitergabe „erhebungstechnischer Hilfsmerkmale“ allgemein und von Namen oder daraus abgeleiteten Größen auf nahezu unüberwindliche Datenschutzprobleme stoßen würde.

Tabelle 4: Übereinstimmung zwischen Nationalität, Geburtsland und Herkunftsland der Eltern

Übereinstimmung zwischen	Anteil Übereinstimmung	Kappa
Nationalität – Geburtsland (Alle)	0.88	0.54
Nationalität – Geburtsland (Ausland)	0.78	0.73
ausländische Nationalität – im Ausland geboren	0.89	0.53
Nationalität – Herkunft Eltern[a] (Alle)	0.88	0.32
Nationalität – Herkunft Eltern[a] (Ausland)[b]	0.66	0.59
Geburtsland – Herkunft Eltern[a] (Alle)	0.90	0.53
Geburtsland – Herkunft Eltern[a] (Ausland)[b]	0.47	0.34
Herkunftsland Vater – Mutter (Alle)	0.99	0.99
Herkunftsland Vater – Mutter (Ausland)	0.98	0.98

[a] gegebenenfalls der Mutter

[b] bei ausländischer Herkunft mindestens eines Elternteils

werden in PASS knapp 50 Naturalisierungen zwischen Welle 1 und Welle 2 beobachtet. Ebenso sind Nachnamensänderungen möglich: Vor allem bei Frauen sind durch Heirat mit deutschen Ehepartnern falsch negative Klassifikationen möglich.[20]

Tabelle 4 zeigt die Übereinstimmungen einerseits zwischen Nationalität und Geburtsland sowie andererseits Nationalität bzw. Geburtsland mit dem Herkunftsland der Eltern.[21] Aufgrund des hohen Anteils gebürtiger Deutscher stimmen für fast 90% aller Personen die in PASS berichtete Nationalität und Geburtsland überein. Bei den Ausländern insgesamt stimmt in über 75% der Fälle Nationalität und Geburtsland überein, wobei die Übereinstimmung je nach Geburtsland variiert. Betrachtet man lediglich, ob eine Person eine deutsche oder irgendeine ausländische Staatsangehörigkeit besitzt und in Deutschland oder generell im Ausland geboren wurde, beträgt die Übereinstimmung der beiden Merkmale knapp 90%.[22] Diese relativ großen empirischen Übereinstimmungen der Merkmale lassen die Verwendung der (selbstberichteten) Nationalität als Kriterium für die Namensklassifikation nicht unplausibel erscheinen.

[20] Ca. 10% der jährlich geschlossenen Ehen in Deutschland sind binationale Ehen zwischen deutschem und nicht deutschem Partner (Haug 2010); im PASS bestehen 8.4% der rund 8700 Paare aus einem deutschen und einem ausländischem Partner.

[21] Kappa korrigiert die prozentuale Übereinstimmung um den Anteil der allein aufgrund der Randverteilung zu erwartenden Übereinstimmungen und liegt daher unter dem Prozentsatz der Übereinstimmungen (Schnell/Hill/Esser 2011: 395).

[22] Für knapp zwei Drittel der Ausländer stimmt die eigene Nationalität mit der der Mutter bzw. des Vaters überein. Die Übereinstimmung beim Geburtsland liegt bei 47%.

Tabelle 5: Mögliche Ergebnisse einer Namensklassifikation

	Namensklassifikation	
Staatsangehörigkeit	deutsch	nicht deutsch
deutsch	Richtig negativ RN	Falsch positiv FP
nicht deutsch	Falsch negativ FN	Richtig positiv RP

5.3 Klassifikationskriterien

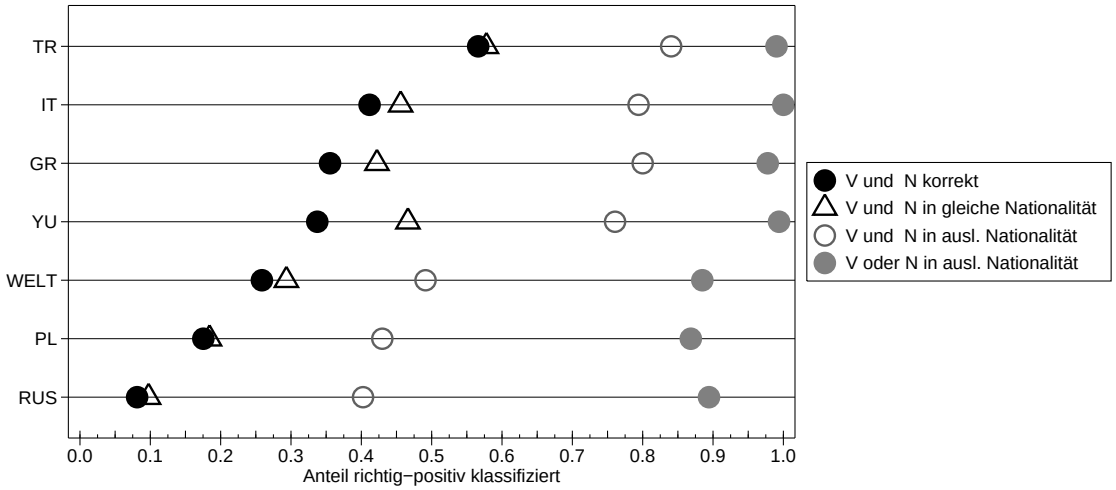
Nach einer Klassifikation der Namen sind vier Ergebnisse möglich (Tabelle 5): Als richtig positiv (RP) wird eine Klassifikation dann bezeichnet, wenn eine Person mit ausländischem Namen auch tatsächlich eine ausländische Staatsangehörigkeit besitzt.[23] Entsprechend bezeichnet eine richtig negative (RN) Klassifikation eine Person mit einem „deutschen" Namen und deutscher Staatsangehörigkeit. Fehlerhafte, falsch positive Klassifikationen (FP) bezeichnen Personen mit deutscher Staatsangehörigkeit, die entweder einen ausländischen Namen tragen, oder deren Namen vom Verfahren fälschlicherweise als ausländisch klassifiziert wurde. Durch falsch positiv klassifizierte Personen entsteht bei Stichproben möglicherweise Overcoverage.[24] Falsch negative (FN) Klassifikationen bezeichnen Personen mit einer ausländischen Staatsangehörigkeit, die entweder keine ausländischen Namen tragen oder deren Namen fälschlicherweise nicht als ausländisch klassifiziert wurde. Durch falsch negativ klassifizierte Personen besteht die Gefahr der Verzerrung der Auswahlgrundlage durch Undercoverage.

Erst mehrere dieser Kriterien gemeinsam erlauben eine Beurteilung der Güte der Klassifikation; jedes Kriterium allein beschreibt eine Klassifikation nur ungenügend. Ob ein Klassifikationsverfahren als gut oder weniger gut betrachtet werden kann, hängt von den Konsequenzen der Falschklassifikationen ab. In der hier beschriebenen Anwendung müssen die Effekte durch die Berücksichtigung fälschlicherweise als „ausländisch" klassifizierte Namen und die Effekte durch den Ausschluss fälschlicherweise als „deutsch" klassifizierte Namen gegeneinander abgewogen werden. Bei einer Screeninganwendung wird im Allgemeinen eher eine hohe Rate falsch Positiver akzeptiert, da diese im weiteren Verlauf der Analysen noch ausgeschlossen werden können. Wünschenswert wäre dabei, dass der Anteil der richtig Positiven an den positiv Klassifizierten insgesamt ($\frac{RP}{RP+FP}$) möglichst hoch sein sollte.

[23] Hier muss berücksichtigt werden, dass es genau genommen nicht um die Klassifikation „ausländischer" Namen geht, sondern um die Klassifikation der Namen von Personen mit ausländischer Nationalität; Namen an sich sind weder „deutsch" noch „ausländisch". Das automatische Screeningverfahren beruht aber darauf, dass bestimmte Namen unter Personen deutscher Nationalität häufiger oder weniger häufig auftreten als unter Personen einer anderen Nationalität.

[24] Zu den Fehlern durch Under- bzw. Overcoverage vgl. Lessler und Kalsbeek (1992).

Abbildung 2: Anteil richtig positiv Klassifikationen nach Ländern und Entscheidungsregel. V=Vorname, N=Nachname. Datenbasis: PASS



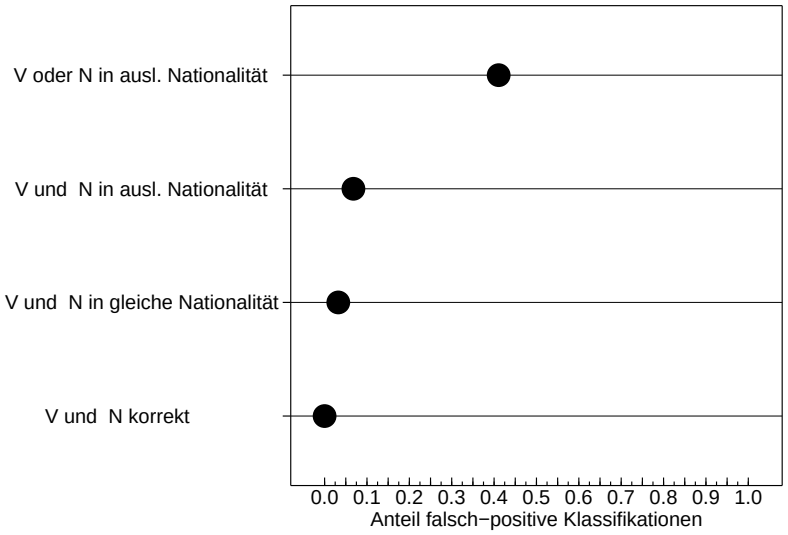
5.4 Kombinationsregeln für die Klassifikation von Vor- und Nachnamen

Das oben beschriebene neue Klassifikationsverfahren klassifiziert Vor- und Nachnamen getrennt. Daher muss eine Entscheidung getroffen werden, wie die separaten Klassifikationen von Vor- und Nachname für eine Gesamtklassifikation berücksichtigt werden sollen.[25]

Bei der Entscheidung über die Art der Kombination von Vor- und Nachnamensklassifikationen muss hier nach den Kosten falscher Klassifikationen entschieden werden. Ein Name könnte als ausländisch gelten, wenn Vor- oder Nachname als ausländisch klassifiziert wird. Dieses Vorgehen erzeugt offensichtlich wenige falsch negative, dafür mehr falsch positive Klassifikationen. Klassifiziert man hingegen einen Namen erst dann als ausländisch, wenn Vor- und Nachname als ausländisch klassifiziert werden, dann sind entsprechend mehr falsch Negative und weniger falsch Positive erwartbar. Eine noch rigidere Regel würde fordern, dass Vor- und Nachname in die gleiche ausländische Nationalität klassifiziert werden müssen, damit ein Name als ausländisch gilt. Durch diese Regel werden viele falsch positive Klassifikationen vermieden, allerdings auf Kosten eines möglichen Undercoverage-Bias durch mehr falsch Negative. Die empirische Wirkung unterschiedlicher Klassifikationsregeln wird daher im nächsten Abschnitt untersucht.

[25] Das Problem der Kombination von Vor- und Nachnamensklassifikation betrifft ebenso alle lexikonbasierten Ansätze, die – wie im Allgemeinen üblich – getrennte Lexika verwenden. Diese Schwierigkeit ließe sich vermeiden, stünde eine Trainingsdatenbank zur Verfügung, welche die gemeinsamen Häufigkeiten für Vor- und Nachnamen enthielte.

Abbildung 3: Anteil falsch positiver (FP) Klassifikationen nach Entscheidungsregel. V=Vorname, N=Nachname. Datenbasis: PASS



5.5 Klassifikationsergebnisse des Verfahrens: Staatsangehörigkeit

Die Abbildungen 2 und 3 zeigen getrennt die Anteile der richtig und falsch positiv klassifizierten Namen anhand der Staatsangehörigkeit mit den PASS-Daten. Je nach Klassifikationsregel (Kombination von Vor- und Nachnamen) zeigen sich unterschiedliche Ergebnisse. Der Anteil der richtig positiv klassifizierten Namen liegt im Durchschnitt über alle berücksichtigten Länder bei über 90%, wenn es ausreicht, dass Vor- oder Nachname nicht als deutsch klassifiziert wird (und erzeugt also nur 100%-90%=10% falsch negativ klassifizierte Namen). Der Anteil richtig positiver Klassifikationen sinkt auf 65%, wenn Vor- und Nachnamen als ausländisch klassifiziert werden. Die dritte Kombinationsregel (Klassifikation von Vor- und Nachnamen in dieselbe Nationalität) führt durchschnittlich zu einem Anteil richtig Positiver von über 35%. Betrachtet man den Anteil der Fälle, bei denen Vor- und Nachnamen korrekt klassifiziert werden, dann werden immer noch 30% aller Ausländer korrekt klassifiziert.

Deutlich zeigen sich in Abbildung 2 Unterschiede zwischen den hier betrachteten Ländergruppen. Insbesondere für Personen aus der Türkei, aus Italien, Griechenland oder dem ehemaligen Jugoslawien zeigt das hier beschriebene automatische Klassifikationsverfahren sehr gute Ergebnisse. Insbesondere bei der strengeren Klassifikationsregel (Vor- und Nachname positiv klassifiziert) werden unter Angehörigen dieser Länder im Durchschnitt noch 4 von 5 Personen korrekt klassifiziert.

Abbildung 3 zeigt die Kosten der unterschiedlichen Klassifikationsregeln: Je schwächer die Klassifikationsregel wird, desto mehr falsch positive Klassifikationen ergeben sich. Die schlechteste Regel (bereits ein „ausländischer“ Vor- oder Nachname genügt zur Klassi-

Tabelle 6: Übereinstimmung zwischen Klassifikation und Migrationshintergrund

Übereinstimmung zwischen Klassifikation und Migrationshintergrund	Anteil Übereinstimmung	Kappa
V&N neg., kein Migrationshintergrund	0.67	0.32
V&N neg., selbst zugezogen	0.34	-0.26
V&N neg., min. 1 Elternteil zugezogen	0.45	-0.04
V&N neg., min. 1 Großelternteil zugezogen[a]	0.47	-0.00
V|N pos., kein Migrationshintergrund	0.33	-0.30
V|N pos., selbst zugezogen	0.66	0.29
V|N pos., min. 1 Elternteil zugezogen	0.55	0.04
V|N pos., min. 1 Großelternteil zugezogen[a]	0.53	0.00
V&N pos., kein Migrationshintergrund	0.19	-0.18
V&N pos., selbst zugezogen	0.86	0.43
V&N pos., min. 1 Elternteil zugezogen	0.85	0.11
V&N pos., min. 1 Großelternteil zugezogen[a]	0.86	-0.02
V=N pos., kein Migrationshintergrund	0.21	-0.10
V=N pos., selbst zugezogen	0.85	0.28
V=N pos., min. 1 Elternteil zugezogen	0.90	0.11
V=N pos., min. 1 Großelternteil zugezogen[a]	0.91	-0.02

[a] Eltern sind in Deutschland geboren

fikation als „ausländisch“) führt zu einem Anteil von fast 40% der Personen mit deutscher Staatsangehörigkeit, die fälschlicherweise als ausländisch klassifiziert werden. Jede andere Regel führt aber zu weniger als 10% falsch positiver Klassifikationen.

Je nach beabsichtigter Zielpopulation lassen sich die Kombinationsregeln so wählen, dass der Anteil richtig positiver deutlich oberhalb von 75% und der Anteil falsch positiver Fälle unterhalb von 10% liegt. Sowohl für die klassischen Migrationspopulationen in Deutschland als auch für die Migranten aus den Nachfolgestaaten der Sowjetunion lassen sich daher mit dem neuen Klassifikationsverfahren Screening-Stichproben konstruieren, die weit effizienter sind als das Screenen reiner Zufallsstichproben.

5.6 Klassifikationsergebnisse des Verfahrens: Migrationsstatus

Die Tabelle 6 zeigt den Anteil der Übereinstimmung zwischen der Klassifikation von Vor- und Nachnamen und dem Migrationshintergrund (unabhängig von der Nationalität). Der Anteil der Übereinstimmung zwischen den Merkmalen „Klassifikation von Vor- und Nachname als deutsch“ und „kein Migrationshintergrund“ beträgt 0.67. Die Übereinstimmung zwischen einem als „deutsch“ klassifizierten Vor- und Nachnamen und eigenem

Tabelle 7: Migrationshintergrund nach Klassifikationsergebnis von Vor- bzw. Nachnamen aller PASS-Teilnehmer, Spaltenprozent

Migrationshintergrund	V & N negativ	V | N positiv	V & N positiv	V & N +identisch[1]
kein Migrationshintergrund	88.9	57.0	20.5	15.7
ausländische Staatsangehörigkeit	1.0	17.6	46.1	51.4
im Ausland geboren	3.4	15.6	21.6	20.3
2 Eltern im Ausland geboren	0.9	2.7	5.3	6.8
1 Elternteil im Ausland geboren	3.5	4.7	5.1	4.6
ausländische Sprache im Haushalt	0.2	0.3	0.5	0.5
alle Großeltern im Ausland geb.	0.1	0.1	0.1	0.0
ein Großelternteil im Ausland geb.	2.1	2.1	0.9	1.5

[1] positiv und Vor- und Nachnamen in dasselbe Land klassifiziert

Migrationshintergrund beträgt hingegen nur 0.34, dies ist deutlich weniger, als durch die Randverteilung allein erwartbar wäre (deutlich durch ein negatives Kappa von -0.26). Sind Eltern oder Großeltern zugezogen, beträgt die Übereinstimmung zwar immer noch ca. 0.45, das Kappa nahe Null zeigt aber, dass diese Übereinstimmung allein durch die Randverteilung erklärbar ist.

Erwartungsgemäß verringert sich die Übereinstimmung zwischen der Klassifikation und dem Migrationsstatus weiter, wenn nur Vor- oder Nachname als „ausländisch“ klassifiziert wird. Nur bei einem eigenen Migrantenstatus sind die Übereinstimmungen nicht allein durch zufällige Übereinstimmungen erklärbar. Sind Vor- als auch Nachname als ausländisch klassifiziert bzw. Vor- und Nachname in die gleiche Auslandskategorie klassifiziert worden, liegt das Ausmaß der Übereinstimmung der Klassifikation mit dem Migrationsstatus bei über 85%.

Betrachtet man die diesen Ergebnissen entsprechenden Anteile falsch positiver und falsch negativer Klassifikationen (Tabelle 7), zeigt sich ein vergleichbares Ergebnis: Je strenger die Kombination von Vor- und Nachnamensklassifikation, desto höher der Anteil der Personen mit Migrationshintergrund.

Werden Vor- und Nachnamen als „deutsch“ klassifiziert, haben lediglich 1% der so klassifizierten Personen eine ausländische Staatsangehörigkeit und nur rund 7% haben einen Migrationshintergrund im weiteren Sinne. Werden hingegen Vor- oder Nachnamen als ausländisch klassifiziert, haben bereits knapp die Hälfte der Personen auch tatsächlich einen Migrationshintergrund, darunter bereits rund 18% auch eine ausländische Nationalität. Weitere 16% haben zwar keine ausländische Nationalität, sind aber selbst im Ausland geboren. Die restlichen 10% haben keinen eigenen Migrationshintergrund, sondern sind Migranten der zweiten (7.5%) oder dritten Generation.

Werden Vor- und Nachnamen als „ausländisch“ klassifiziert, haben insgesamt 4 von 5 der so positiv klassifzierten Personen einen Migrationshintergrund. Bei dieser Regel sind

Tabelle 8: Anteile der Übereinstimmung der Namensklassifikation mit der Nationalität nach Art der verwendeten n-Gramme

	Nachname		Vorname	
Land (wahrer Wert)	Bigramme	Trigramme	Bigramme	Trigramme
Deutschland	0.90	0.84	0.87	0.68
Italien	0.69	0.79	0.42	0.51
Türkei	0.63	0.75	0.55	0.77
Griechenland	0.57	0.60	0.49	0.47
Jugoslawien[a]	0.48	0.60	0.31	0.52
Osteuropa[b]	0.31	0.36	0.23	0.42
Russland[c]	0.17	0.14	0.33	0.58
Insgesamt	0.87	0.81	0.83	0.67

[a] einschließlich Nachfolgestaaten

[b] Polen und osteuropäische Nachbarländer

[c] einschließlich ehemaliger Staaten der Sowjetunion

dann 46% ausländische Staatsangehörige. Wird das Kriterium zur Klassifikation weiter verschärft, so dass Vor- und Nachnamen nicht nur als ausländisch gelten, sondern beide auch in die gleiche Nationalität klassifiziert werden müssen, haben lediglich nur rund 16% der Personen keinen Migrationshintergrund und über 50% sind ausländische Staatsangehörige. Auch hier zeigt sich, dass bei Wahl der für die eigenen Zwecke geeigneten Klassifikationsregel effiziente Screeningkriterien durch das Verfahren realisiert werden können.

5.7 Zur Wahl von Bigrammen oder Trigrammen zur Klassifikation

Zusätzlich zur Frage, wie die separaten Klassifikationen von Vor- und Nachnamen kombiniert werden, muss eine Entscheidung getroffen werden, ob für die Klassifikation aus den Namen Buchstabenketten aus zwei oder aus drei Buchstaben (Bigramme bzw. Trigramme) verwendet werden sollen.[26] Tabelle 8 enthält zum Vergleich der Klassifikation durch Bi- oder Trigramme den Anteil korrekter Klassifikationen. Deutlich wird, dass der Anteil korrekter Klassifkationen unter den Ausländern bei der Verwendung von Trigrammen in der Regel höher ist als bei der Verwendung von Bigrammen. Ausnahmen hiervon sind Vornamen bei Russen und Nachnamen bei Personen mit griechischer Staatsangehörigkeit.[27]

[26] Längere n-Gramme eignen sich für die Klassifikation kaum, da längere n-Gramme für die häufigen kurzen Namen einen exakten und nicht mehr fehlertoleranten Abgleich implizieren.

[27] Bei Personen mit deutscher Staatsangehörigkeit erzeugt die Zerlegung der Vor- oder Nachnamen in Bigramme jeweils höhere Anteile korrekter Klassifikationen.

6 Validierung des Verfahrens anhand einer prospektiven Studie mit einer Stichprobe türkischer Namen

Infas führte im Jahr 2009 im Bundesland Hessen eine telefonische Befragung unter anderem von Personen mit türkischem Migrationshintergrund durch. Für diese Studie wurde das hier beschriebene Verfahren erstmals für eine Stichprobenziehung verwendet. Eine große Stichprobe aus den Namen einer Telefon-CD Hessens wurde mit dem Verfahren klassifiziert.[28] Mit den Telefonnummern, die zu den als „türkisch“ klassifizierten Namen gehörten, konnten 839 Interviews realisiert werden. Da in der Erhebung für die Befragten der Migrationshintergrund erhoben wurde, kann diese Studie für eine weitere teilweise Validerung des Verfahrens verwendet werden.

Tabelle 9 zeigt für alle als türkisch klassifizierten Personen der Stichprobe den Migrationshintergrund bis zur dritten Generation. 12% der als türkisch klassifizierten Personen haben keinen türkischen Migrationshintergrund, 43% besitzen tatsächlich die türkische Staatsangehörigkeit, weitere 13% sind zwar keine türkischen Staatsangehörigen (mehr), sind aber in der Türkei geboren. Weitere 13% haben keinen eigenen Migrationshintergrund, mindestens ein Elternteil ist jedoch in der Türkei geboren. Berücksichtigt man weiter die Herkunft der Großeltern und die Sprache, die im Haushalt gesprochen wird, haben von 839 als türkisch klassifizierte Personen mehr als zwei Drittel (70.8%) einen türkischen Migrationshintergrund im engeren Sinne (eigener Migrationshintergrund oder der Eltern).[29] Das Klassifikationsverfahren hat in dieser Stichprobe also erfolgreich die Zahl der Kontakte, die notwendig waren um durch Screening eine ausreichend große und zufallsbasierte Stichprobe zu gewinnen, reduzieren können.

Zum Vergleich zeigt Tabelle 9 das Ergebnis im PASS, wenn Vor- oder Nachnamen als türkisch klassifiziert werden. Die Ergebnisse sind hier sehr ähnlich.[30] In beiden Stichproben ergeben sich nur rund 12% falsch Positive. Insgesamt bestätigt die hessische CATI-Studie die praktische Einsatzfähigkeit und die Effizienz eines Screenings mit dem beschriebenen Verfahren.

[28] Die positiv klassifizierten Namen wurden in dieser Studie anschließend von einem Angehörigen der Zielpopulation gesichtet und um wenige „offensichtlich“ falsch-positive Fälle bereinigt. Diese zusätzliche Bereinigung war aufgrund des Auftraggebers unverzichtbar.

[29] Von den rund 100 falsch Positiven ohne türkischen Migrationshintergrund haben rund ein Viertel eine andere ausländische Staatsangehörigkeit.

[30] Das ist umso erstaunlicher, als für diese Studie anhand der PASS-Daten kein zusätzliches manuelles Review der positiv klassifizierten Fälle stattfand, sondern nur automatisch klassifiziert wurde. In PASS wurden alle türkischen Staatsangehörige vom Verfahren als „türkisch“ klassifiziert, d.h. es gab keine falsch negativen Fälle.

Tabelle 9: Vergleich der Klassifikationen der CATI-Studie in Hessen mit PASS

Klassifikation: türkisch	CATI Hessen %	PASS %
türk. Staatsangehörigkeit	43.4	49.6
in der Türkei geboren	12.8	14.4
beide Eltern in der Türkei geboren	10.7	11.0
ein Elternteil in der Türkei geboren	2.0	0.5
Sprache im Haushalt: Türkisch	1.9	2.0
alle Großeltern NICHT in D geboren	16.9	10.4
kein türk. Migrationshintergrund	12.3	12.1
Anzahl als „türkisch“ klassifiziert	839	960

7 Zusammenfassende Beurteilung

Das hier vorgestellte Verfahren eignet sich zum effizienten Screenen nach Ausländern und Migranten in Namenslisten. Je nach den jeweiligen Kosten von Falschklassifikationen lassen sich mehr oder weniger strenge Klassifikationsregeln wählen und entsprechend der jeweiligen Aufgabenstellung eher falsch negative oder eher falsch positive Screeningergebnisse vermeiden.[31]

Das hier beschriebene vollautomatische namensbasierte Verfahren eignet sich einige Migrantengruppen und Länder besser als für andere. Insbesondere bietet es sich für die Suche nach Personen an, die aus den klassischen Migrationsländern der Bundesrepublik stammen (Türkei, Italien, Griechenland und Jugoslawien). Für andere Gruppen (z.B. Russen und Angehörige osteuropäischer Nachbarländer) ist das vorgestellte Verfahren auch geeignet, besitzt aber für diese Gruppen eine geringere Effizienz als für die klassischen Migrationsländer. Aber auch bei den weniger geeigneten Subpopulation liegt die Effizienz des Verfahrens immer noch über dem der reinen Zufallsauswahl; ebenso können auch geringere Auswahleffekte erwartet werden als bei Quoten- oder Schneeballverfahren.

Das Verfahren ermöglicht mit geringem Aufwand in kurzer Zeit ein Screening nach Ausländer- oder Migrantenpopulationen in umfangreichen Auswahlgrundlagen, wie z.B. Einwohnermelderegistern, Telefonbüchern, Patientendateien usw. Ein weiterer Vorteil liegt darin, dass ein vorheriges manuelles Erstellen von Einträgen in einem Namenslexikon vermieden werden kann. Gegenüber den reinen Lexikonverfahren bietet das hier vorgestellte Verfahren durch die Zerlegung der Namen in n-Gramme den Vorteil, dass es fehlertoleranter gegenüber den in Registern häufigen minimalen Schreibvarianten insbeson-

[31] Die Frage der Konsequenzen falsch negativer Klassifikationen bei namensbasierten Verfahren im Allgemeinen oder im Vergleich dieses Verfahrens mit anderen Verfahren überschreitet den Rahmen dieser Arbeit. Eine entsprechende umfangreiche Analyse ist Gegenstand laufender Bemühungen der Arbeitsgruppe (Schnell et al. 2011a).

dere ausländischer Namen ist. Eine Erweiterung der grundsätzlichen Einsetzbarkeit des Verfahrens ist durch die Bereitstellung anderer Trainingsdaten zu erreichen. Schließlich ist das Verfahren ohne Modifikationen auch auf andere Länder anwendbar, falls dort entsprechende Trainingsdaten zur Verfügung stehen.

Literatur

Babka von Gostomski, Christian, 2008: Türkische, griechische, italienische und polnische Personen sowie Personen aus den Nachfolgestaaten des ehemaligen Jugoslawien in Deutschland. Erste Ergebnisse der Repräsentativbefragung „Ausgewählte Migrantengruppen in Deutschland 2006/2007“ (RAM). Working Paper 11, Bundesamt für Migration und Flüchtlinge, Nürnberg.

Bertelsmann Stiftung (Hg.), 2009: Zuwanderer in Deutschland. Ergebnisse einer repräsentativen Befragung von Menschen mit Migrationshintergrund. Gütersloh: Bertelsmann Stiftung.

Blane, Howard D., 1977: Acculturation and Drinking in an Italian American Community. Journal of Studies on Alcohol 38 (7): 1324–1346.

Brettfeld, Katrin, und *Peter Wetzels*, 2007: Muslime in Deutschland. Integration, Integrationsbarrieren, Religion sowie Einstellungen zu Demokratie, Rechtsstaat und politisch-religiös motivierter Gewalt. Berlin: Bundeministerium des Innern.

Burkhauser, Richard V., *Michaela Kreyenfeld* und *Gert G. Wagner*, 1997: The Immigrant Sample of the German Socio Economic Panel. Aging Studies Working Paper 7, Maxwell Center for Demography and Economics of Aging, Syracus, NY.

Diefenbach, Heike, und *Anja Weiß*, 2006: Zur Problematik der Messung von „Migrationshintergrund“. Münchner Statistik 3: 1–14.

Ecob, Russell, und *Rory Williams*, 1991: Sampling Asian Minorities to Asses Health and Welfare. Journal of Epidemiology and Community Health 45: 93–101.

European Union Agency for Fundamental Rights, 2009: EU-MIDIS Technical report: methodology, sampling and fieldwork. European Union minorities and discrimination survey. Budapest: Elanders Hungary Kft.

Galonska, Christian, *Maria Berger* und *Ruud Koopmans*, 2004: Über schwindende Gemeinsamkeiten: Ausländer- versus Migrantenforschung. Technischer Bericht, Wissenschaftszentrum Berlin (WZB).

Granato, Nadia, 1999: Die Befragung von Arbeitsmigranten: Einwohnermeldeamt-Stichprobe und telefonische Erhebung? ZUMA-Nachrichten 45 (23): 44–60.

Haisken-DeNew, John, und *Joachim R. Frick*, 2005: Desktop Companion to the German Socio-Economic Panel (SOEP). Berlin: DIW Berlin.

Halm, Dirk, und *Martin Sauer*, 2005: Freiwilliges Engagement von Türkinnen und Türken in Deutschland. Projektbericht, Stiftung Zentrum für Türkeistudien an der Universität Duisburg-Essen.

Haug, Sonja, 2010: Interethnische Kontakte, Freundschaften, Partnerschaften und Ehen von Migranten in Deutschland. Working Paper 33, Bundesamt für Migration und Flüchtlinge, Nürnberg.

Haug, Sonja, und *Frank Swiaczny*, 2003: Migrations- und Integrationsforschung in der Praxis. Das Beispiel BiB-Integrationssurvey. Zeitschrift für Angewandte Geographie 27 (1): 16–20.

Humpert, Andreas, und *Klaus Schneiderheinze*, 2000: Stichprobenziehung für telefonische Zuwanderumfragen. Einsatzmöglichkeiten der Namensforschung. ZUMA-Nachrichten 24 (47): 36–64.

Infas, 2009: Methodenbericht des Projekts Kriminalitätsfurcht in Hessen. Bonn.

Kalton, Graham, 2009: Methods for oversampling rare subpopulations in social surveys. Survey Methodology 35 (2): 125–141.

Kalton, Graham, und *Dallas Anderson*, 1986: Rare Populations. Journal of the Royal Statistical Society. Series A 149 (1): 65–82.

Lauderdale, Diane S., 2006: Birth Outcomes for Arabic-Named Women in California Before and After September 11. Demography 43 (1): 185–201.

Lessler, Judith T., und *William D. Kalsbeek*, 1992: Nonsampling Errors in Surveys. New York: Wiley.

Mammey, Ulrich, und *Jörg Sattig*, 2002: Determinanten und Indikatoren der Integration und Segregation der ausländischen Bevölkerung (Integrationssurvey). Projekt- und Materialdokumentation. Materialien zur Bevölkerungswissenschaft des Bundesinstituts für Bevölkerungsforschung, Nr. 105a. Wiesbaden: Bundesinstitut für Bevölkerungsforschung.

Passel, Jeffrey S., und *David L. Word*, 1993: Constructing the List of Spanish Surnames for the 1980 Census: An Application of Bayes' Theorem (Paper presented at the Annual Meeting of the Population Association of America) 1980. Technical Working Paper 4, Population Division, U.S. Bureau of the Census, Washington D.C.

Perkins, Colby R., 1993: Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results. Technical Working Paper 4, Population Division, U.S. Bureau of the Census, Washington D.C.

Promberger, Markus (Hg.), 2007: Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung „Arbeitsmarkt und Soziale Sicherung". IAB-Forschungsbericht (12). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.

Rother, Nina, 2010: Das Integrationspanel. Ergebnisse einer Befragung von Teilnehmenden zu Beginn ihres Alphabetisierungskurses. Working Paper 29, Bundesamt für Migration und Flüchtlinge, Nürnberg.

Sachverständigenrat deutscher Stiftungen für Integration und Migration (SVR) (Hg.), 2010: Einwanderungsgesellschaft 2010. Jahresgutachten 2010 mit Integrationsbarometer. Berlin: Sachverständigenrat deutscher Stiftungen für Integration und Migration.

Salentin, Kurt, 2007: Die Aussiedler-Stichprobenziehung. Methoden – Daten – Analysen 1 (1): 25–44.

Schnell, Rainer, 2009: Wie man Nadeln in Heuhaufen findet. Stichprobenverfahren für seltene und sehr seltene Bevölkerungsgruppen, Vortragsmanuskript, Universität Duisburg-Essen.

Schnell, Rainer, *Tobias Gramlich* und *Mark Trappmann*, 2011a: Potential Undercoverage and Bias in Name-based Samples of Foreigners. Vortrag auf der ASA Spring Methodology Conference, Mai 2011, Tilburg.

Schnell, Rainer, *Paul B. Hill* und *Elke Esser*, 2011b: Methoden der empirischen Sozialforschung. München, 9. Auflage: Oldenbourg.

Schwartz, Kendra L., *Anahid Kulwicki*, *Linda K. Weiss*, *Haifa Fakhouri*, *Wael Sakr*, *Gregory Kau* und *Richard K. Severson*, 2004: Cancer among arab americans in the metropolitan Detroit area. Ethnicity & Disease 14 (1): 141–146.

Shackleford, Michael, 1998: Actuarial Note: Name Distributions in the Social Security Administration Area, August 1997. Social Security Administration Acturial Note 139, Social Security Administration. Office of the Chief Actuary, Baltimore, MA.

Statistisches Bundesamt, 2011a: Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund. Ergebnisse des Mikrozensus 2010. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt, 2011b: Bevölkerung und Erwerbstätigkeit. Ausländische Bevölkerung. Ergebnisse des Ausländerzentralregisters. Wiesbaden: Statistisches Bundesamt.

Trappmann, Mark, *Stefanie Gundert*, *Claudia Wenzig* und *Daniel Gebhardt*, 2010: PASS: a household panel survey for research on unemployment and poverty. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 130 (4): 609–622.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Template layout

Christine Weidmann

www.record-linkage.de