German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2013-04 | JULY 25, 2013

A new Name-Based Sampling Method for Migrants using n-grams

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

A new Name-Based Sampling Method for Migrants using $n-$grams*

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

[1]University of Duisburg-Essen
[2]Institute for Employment Research
[3]infas Institute for Applied Social Sciences
[4]Federal Office for Migration and Refugees

July 25, 2013

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 1.600 foreigners in the PASS panel. Finally, a CATI survey based on the new method in Hesse (Germany) is described.

Keywords: Onomastics, rare populations, bigrams, trigrams

*This is a slightly edited English version of a previously published peer-reviewed article (Schnell et al. 2013a). The authors thank Sabrina Torregroza for her translation of the manuscript. We have to thank two anonymous reviewers of the German manuscript for their helpful comments. The authors are indebted to Andreas Humpert and Klaus Schneiderheinze for helpful discussions of a previous version. The method suggested here, as well as the type of test have been proposed by Rainer Schnell and have been implemented by Tobias Bachteler and Jörg Reiher. Mark Trappmann was involved in the conception of this article and commented on different versions of the manuscript. Menno Smid conducted the CATI survey in Hesse (Germany). The classifications, calculations of the tests and prior drafts of the paper have been done by Tobias Gramlich. Inna Becher provided a literature review of sampling techniques for rare populations. The final version was written by Rainer Schnell.

1 Introduction

Populations of migrants and foreigners[1] in Germany are of vital importance for numerous fields of research, such as educational research or research on the labour market. However, the usual sampling procedures are not suitable for this population for two reasons: the proportion of certain groups of migrants in a general population is rather small and most methods for drawing random samples are therefore inefficient. Secondly, the most frequently used sampling frames (e.g. phone books, register of residents or patients) often do not contain information on the membership to the specific target population. Especially those who have a migration background but also the German citizenship are very hard to find with these sampling frames. Special sampling procedures are therefore needed for this population.

2 Sampling procedures for samples of migrants

Usually no complete sampling frame is available for migrants.[2] Therefore, special sampling procedures for "rare populations" have to be used (Kalton 2009). The upper limit for the classification as a "rare population" is typically a proportion of up to 10% of the total population (Kalton and Anderson 1986). So, despite having a proportion of – depending on definition – about 9.0% foreigners or 19.3% people with a migration background (Statistisches Bundesamt 2011b;a), foreigners in Germany form a rare population in this sense.

[1]The terms "foreigner" and "migrant" are obviously not congruent. For a discussion on the conceptual problems of these terms see e.g. Galonska et al. (2004) and Diefenbach and Weiß (2006). Hence, the empirical implementation of these concepts is not trivial, see e.g. Statistisches Bundesamt (2011b). Despite these problems, the terms "foreigner" and "migrant" are used interchangeably in the following text since at least one reference value is needed for the evaluation of every empirical procedure. Due to the lack of other available data, the method introduced below is evaluated using citizenship. The underlying concept of the proposed method however, is irrelevant for the technical procedure: if conceptually different reference data is available for the training of the procedure (names given migration background) other concepts may be used.

[2]Only the Federal Office for Migration and Refugees (BAMF) – the registry's owner – has used the Central Register of Foreigners (AZR) for some cases of sampling as part of its mandate (Babka von Gostomski 2008). Due to court decisions the micro data of the AZR are unavailable for any statistical use outside the BAMF.

2.1 Cluster samples

Lists of subpopulations are often used as sampling frames for rare populations. For example, this includes lists of members of associations or organizations or lists of participants. As an example, Rother (2010) was able to use official data of the Federal Office for Migration and Refugees (BAMF) and has drawn a cluster sample of foreigners out of all participants of integration courses.

Frequently, elements of the target population cluster geographically: This often leads to a high density of foreigners from certain countries in certain geographic areas ("Little Italy", "China Town"), so that standard sampling procedures can be applied efficiently (see e.g. Blane 1977; Ecob and Williams 1991).

For example, in some member states the European Survey on Minorities and Discrimination (EU-MIDIS) is limited to minority groups with a share of at least 5% of the total population and to certain areas with a "medium to high concentration" of the target population (European Union Agency for Fundamental Rights 2009: p. 18f). The main problem of these procedures is that members of the target population who live in these kind of areas can differ from those who live in areas with a low concentration, especially regarding the dependent variables (such as "integration" or "perception of discrimination").

2.2 Non-probability sampling procedures

Especially in Germany, samples of foreigners are often drawn using quota sampling, which leaves the selection of a specific respondent, at least partially, up to the interviewer. A well known example for this kind of sampling procedure is the study "Zuwanderer in Deutschland" (Immigrants in Germany) (Bertelsmann Stiftung 2009). Since the selection is done by the interviewers this might result in the selection of better integrated migrants. In general, quota sampling should not be used for scientific purposes, due to their uncontrollable effects (Schnell/Hill/Esser 2011:296-298).

Furthermore, network or snowball sampling are being used as sampling methods for foreigners or migrants, as done in a subsample of German migrants from the Soviet Union in the German Socio Economic Panel Study GSOEP (Burkhauser et al. 1997). Network-samples suffer from two problems: the selection probabilities usually cannot be calculated, and socially isolated people have smaller probabilities of selection. Network-samples should therefore be regarded as a last resort rather than a standard method to obtain reliable and valid results.

2.3 Screening

Very often, only sampling frames for general populations are available. Usually, all sampled elements of such a frame are examined if they belong to the rare subpopulation. This procedure is referred to as "screening". For example in Germany samples from local population registries can be drawn and foreigners be identified by their documented citizenship (see Granato 1999). An example is the sample of foreigners in the GSOEP study (Sample B) (Haisken-DeNew and Frick 2005). This approach is not suitable for all target populations, though. For example, screening based on nationality cannot be used for migrants who have the German nationality. Salentin (2007) therefore proposes screening based on the place of birth instead of nationality.[3] Although this procedure would find naturalized first generation migrants, it will miss their children and grandchildren, who were born in Germany. While it would be possible to ask about the parents' or even grandparents' country of birth in a survey, the validity of the given answers is probably rather questionable. There seem to be no empirical studies on this topic, though.

2.4 Name-based procedures

The actual screening-attribute often is not included in the sampling frame so that alternative attributes have to be used. In many cases only a list of names is available. Lists of names are being used throughout the world in many research projects concerned with populations of migrants, since names might indicate a regional or ethnic origin of the name bearer. In practice, in order to screen for migrants especially onomastic dictionaries of names with known affiliation (meaning exclusive or very frequent occurrence in the relevant group) to a certain country of origin are used or constructed in a time-consuming matter.

The onomastic method developed by Humpert and Schneiderheinze (2000) has become the standard procedure for sampling migrants in Germany. The idea is to compare and classify names of different nationalities from public directories with their onomastic name databases. Recent examples for the application of this procedure are the annual report of the Expert Council of German Foundations for Migration and Integration (SVR) (2010), the survey on integration by the Federal Institute for Population Research BiB (Mammey and Sattig 2002; Haug and Swiaczny 2003), as well as the study

[3]This would only be possible if "place of birth" is available for sampling. A close examination of all 16 registration laws in Germany revealed that under the current law, the variable can only be obtained for a citizen known by name before, but not for sampling).

"Muslims in Germany" (MiD, Brettfeld and Wetzels 2007). A review by Mateos (2007) described comparable approaches, but missed the paper by Humbert/Schneiderheinze (2000).

However, similar methods are being used in epidemiology, where the use of ad hoc compiled lists of frequent or typical names in certain countries are common. These lists are usually examined and edited by native speakers. The resulting lists are then used to screen a more general population sampling frame. One example is the work of Halm and Sauer (2005: pp. 43). They classified entries from a phone book based on a list containing about 10.000 "typical" surnames and about 7.000 "typical" first names. Similar procedures are being used in epidemiology throughout the world. As an example, Schwartz et al. (2004) created a list of common Arabic first and last names based on different publicly available resources[4] and manually examined them to draw a sample out of a Detroit cancer registry. In a similar way Lauderdale (2006) used a list of names of all those covered by social security to screen for Arabian women in birth registers in California.

3 A new sampling method: Classifying by $n-$grams

The method presented below is, in contrast to all methods previously used, not based on the classification of *full* names and suffixes, but on the classification of letter substrings (n-grams).[5]

The starting point of this procedure are lists which contain frequencies of names for different nationalities, separated by first and last name. Most suitable for the construction of such lists are census or social security databases.[6]

From the proportion of a certain name $P_{(name)}$, the proportion of all people of a certain nationality $P_{(nationality)}$, and the conditional proportion of a name from those with a certain nationality $P_{(name|nationality)}$ the conditional

[4]Various registers (such as birth and death records) which contained names as well as the corresponding birth countries; membership lists of (e.g. cultural-) clubs and associations; furthermore (Arabic-speaking) employees collected „typical" Arabic names out of phone books.

[5]The method described here was presented by the authors at the ESRA conference 2011 in Lausanne and at the ASA conference 2011 in Tilburg. The method was developed by Rainer Schnell, Tobias Bachteler and Jörg Reiher at the University of Konstanz in November 2006 and has been used for a number of sampling procedures and record linkage projects. A first description in German can be found in Schnell (2009). An earlier German draft was published as a discussion paper in June 2012 (Schnell et al. 2012).

[6]Shackleford (1998) and Lauderdale (2006) describe such lists of name frequencies for different nationalities using US census counts and all those persons covered by social security in the US.

probability for a certain nationality given a certain name $P_{(nationality|name)}$ can be calculated using Bayes's theorem:

$$P_{(nationality|name)} = \frac{P_{(name|nationality)} * P_{(nationality)}}{P_{(name|nationality)} * P_{(nationality)} + P_{(name|\neg nationality)} * P_{(\neg nationality)}}$$

The disadvantage of such a classification consists in the use of complete and correctly written names.[7] About 15-25% typing errors in names are typically reported in the literature on record-linkage (Winkler 2009). These errors presumably are especially concentrated on names of migrants. Methods based on exact matching are therefore expected to produce particularly many false negative non-matches for these cases. Accordingly, ratios would be systematically underestimated. A more error-tolerant method for name classification is therefore preferable.

One option is not to use the complete first or last names, but to split names into sets of several consecutive characters (n-grams) and use these sets of n-grams to classify the name. In computer science such n-gram based methods are being used for many problems of natural language processing, such as the construction of search engines or spelling checking programs.

Below, we therefore use the relative frequencies of all n-grams of a first or last name in order to calculate the conditional probability of a nationality given a set of n-grams.[8] In addition to the error tolerance, this method has yet another advantage: it does not need a complete dictionary of names. It may therefore also classify names which are not listed in a dictionary.

[7]The idea of classifying names based on Bayes's theorem appears to have been used in an implementation of the US-Bureau of the Census for the first time: Passel and Word (1980) and Perkins (1993) describe the construction of a list of "Spanish" names for the US-American census in 1980. Using Bayes's theorem conditional probabilities of belonging to the Hispanic subpopulation have been calculated. The census application is based on complete names; the procedure was not included in the sampling literature on rare populations. Additionally, some applications can be found in computational linguistics, where texts were classified by languages using n-grams (e.g. Cavnar and Trenkle 1994). Even less common is the automatic classification of names (Konstantopoulos 2007). Both works do use n-grams, but classify by n-gram-profile similarity and not using Bayes' theorem. These applications are not found in the sample literature, either.

[8]As the results of this study show, nationality can be used as a training attribute, despite the high ratio of German citizens among those with a migration background, as long as the conditional probabilities of different n-grams differ significantly between groups of interest.

Table 1 demonstrates the classification using the example of the name "Peter". We start by splitting the name "Peter" (length: 5 letters) into a set of n-grams (here bigrams, $n = 2$). The name consists of 4 bigrams: {PE,ET,TE,ER}. Since word or name beginnings and endings are often characteristic for a certain language,[9] it is recommended to add a blank in front and behind the actual name, so that additional n-grams are generated at the beginning and at the end of the name:

$$\{\_P,PE,ET,TE,ER,R\_\}$$

For each country, the number of occurrences of each specific bigram is then divided by the total number of names[10] per nationality. Following this, the product of this relative frequency of the n-grams in the name "Peter" is then calculated for each country.[11] This product is multiplied with the factor w.[12] A name will then be classified as belonging to the nationality for which this classification takes on the maximum.[13]

4 Training data for the classsification by $n-$grams

The central point of this new procedure is a database of names from which the n-gram-frequencies can be obtained. A database containing the names and their frequencies for all nationalities considered is therefore needed. Such databases are – especially outside of Germany – available from different sources and of varying quality. In Germany, lists based on the register of residents could be considered, but due to the necessity of negotiating with each

[9]This also applies for names: e.g. German male names usually do not end with an -a. Neither do they usually end with an -e, while Italian names with an -e at the end are often male name variants (e.g. the Italian names "Simone" (male) and "Simona" (female)).

[10]It is being divided by the number of names, and not by the number of n-grams. Differences between countries are therefore not further accentuated by in the length of the names.

[11]The contribution of an "unknown", undocumented n-gram for the classification of a name was not set to 0, because otherwise the entire expression in the multiplication of the overall probability would be 0, of course. For our application, the factor $\frac{1}{N}$ was used instead of 0. At this point, when using a more elaborate program, the application of n-gram-smoothing could be useful (Jurafsky and Martin 2009). For the first-time application on sampling however, the procedure chosen here proved to be more practicable.

[12]If $n_{names[j]}$ is the number of names from country j and $N = \sum n_{names[j]}$ is the number of names from all countries combined, this results in $w_{[j]} = \frac{n_{names[j]}}{N}$, which is the prior probability in Bayes' theorem.

[13]Using a scripting language, the computational implementation only consists of a few lines.

Table 1: Classification of the name "Peter"[a]

nationality	bigram	Count bigrams	total count names/country	$p_{country}$	$\Pi p_{country}$	correction factor w	classification-size (country|"Peter")
Germany	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			
eastern Europe[b)]	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			
former Yugoslavia	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			
Italy	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			
Turkey	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			
...	...	.....	.....	...	...	...	...
Total		14.221.676	30.535.580				

[a]The names from Greece and the former Soviet Union countries not reported here have been considered in the calculation.
[b] Poland and the eastern European neighbor states.

community separately, this approach is rather difficult in practice. An obvious alternative would be the use of telephone directories, from which names and frequencies, but not the corresponding nationality, could be obtained. If telephone directories from different countries would be used, all names listed there could be treated as "native", accepting the overall small probability of the inevitable misclassification of migrants each telephone directory contains. Alternatively, one could limit themselves to the most common names above a certain threshold (and which could be set differently for each country).

Easier to handle are, of course, appropriate lists from a single data source. For the development of the method presented here, lists separated by nationality and containing the frequencies of first and last names of all those employees subject to social insurance contribution were constructed in Germany for the first time.[14] These lists contain names and frequencies separated by nationality. For privacy reasons, first and last names were therefore not combined (e.g. "D – Peter Müller – n=5"), but listed only separately ("D – Peter – n=40.000" and "D – Müller – n=1.000").

Overall, the database of names contained 112.831 different first and 493.974 different last names for the nationalities classified here. Weighted according to their respective frequencies, this results in about 30 million cases. Table 2 shows the number of different names for the groups of countries considered for classification here.

[14]On request of the first author, this database has been compiled in 2005 for the first time by the German Research Data Center of the Federal Employment Agency at great efforts and after consulting the data protection officer of the Federal Employment Agency. The lists were only provided with subject to the conditions that the lists would be constructed separately for first and last names and that each name had to occur at least five times.

Table 2: Number of first and last names in the training database

	first name		last name	
nationality	names	people	names	people
Germany	58.757	28.309.791	383.592	27.551.167
former Yugoslavia[a]	10.494	313.193	21.973	262.425
Turkey	8.137	605.029	20.835	587.175
eastern Europe[b]	2.750	103.300	7.273	47.366
Italy	2.707	216.672	14.334	180.118
Greece	2.517	112.744	9.452	75.732
Russia[c]	1.952	47.638	2.476	13.187
rest of the world	25.517	470.446	34.039	286.887
Total	112.831	30.178.813	493.974	29.004.057

[a] former Yugoslavia and successor states

[b] Poland and eastern European neighbor countries

[c] Russia and successor states of the former Soviet Union

"Russia" contains all entries of names and frequencies made for successor states of the former Soviet Union and the former Russian Federation.[15] Under "former Yugoslavia", names were combined from all Yugoslavian successor states.[16] To increase the number of cases for eastern European countries, Poland and several eastern European neighbor states were combined into a group of countries ("eastern Europe").[17]

As described below, a panel data set was used for the validation of the method which, despite its size of nearly 19.000 respondents in each wave, only contained enough cases for the most common groups of foreigners in Germany. In this article, we therefore limit the study to the largest groups of foreigners in Germany.[18]

[15]In this data set this includes Estonia, Latvia, Lithuania, the Soviet Union, the Russian Federation, Ukraine, Belarus, Armenia, Azerbaijan, Kazakhstan, Kyrgyzstan, Tajikistan and Turkmenistan.

[16]This includes Bosnia and Herzegovina, Yugoslavia, Croatia, Macedonia, Albania and Slovenia. The Former Republic of Yugoslavia was at the time of the preparation of this database of names not yet disintegrated in Serbia and Montenegro.

[17]Covering Poland, Bulgaria, Hungary, Romania, former Czechoslovakia, Slovakia, and the Czech Republic.

[18]This limitation results solely from the size of the subgroups in the data set used for validation. In other applications a different grouping of countries could be chosen.

Figure 1: Procedure for constructing the training database

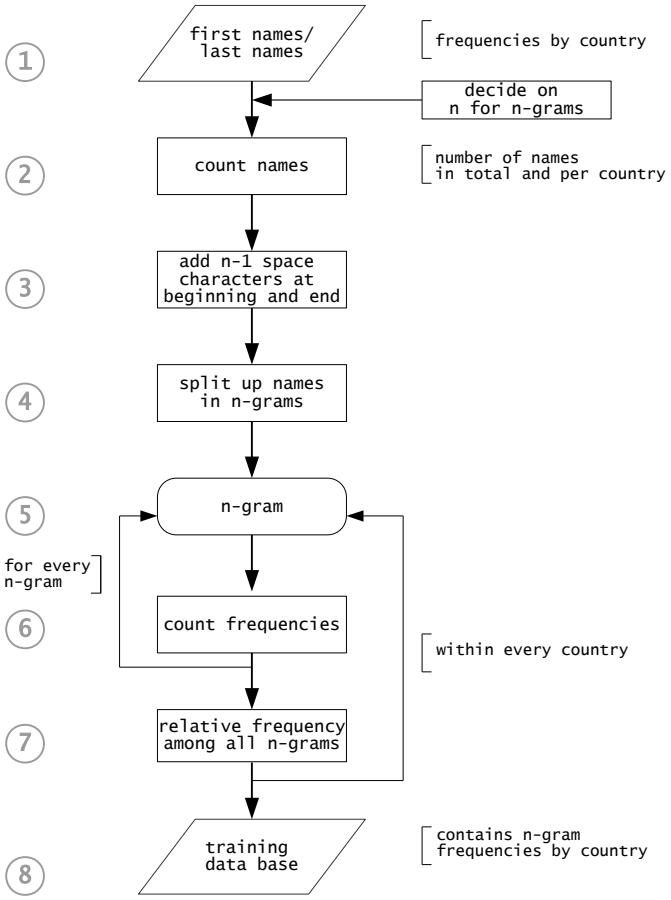


Figure 1 shows a diagram of the procedure used to construct the training database.[19]

5 Validation of the new method

The classification method described above was tested using two data sets: the data set of the German PASS panel survey and the German survey "Fear of Crime in Hesse". The results of both surveys are presented below.

[19]Since the procedure can, among other things, be used for commercial purposes (e.g. by direct marketing databases), the research group has decided not to publish the trainings data nor the classification results.

5.1 The panel "Labour Market and Social Security"

The German "Panel Labour Market and Social Security" (PASS) (Promberger 2007; Trappmann et al. 2010) is a household survey newly launched in 2007 by the Institute for Employment Research of the Federal Employment Agency (IAB).

The PASS-Panel consists of two subsamples: Sample 1 consists of households receiving social welfare benefits (according to the social code SGB II). This sample is updated with new SGB II-benefit recipients in each wave. Sample 2 is a general population household sample disproportionately stratified on low income. All household members 15 years of age and older are being interviewed in the selected households. Respondents who are over 65 years old receive a shortened version of the questionnaire. If possible, the interview was conducted by telephone (CATI), but if necessary, the interview was also conducted by an interviewer in person (CAPI).

The names of all participants to the first and second wave, separated by first and last names, were classified with the described classification method on the property of the IAB.[20] Overall, 18.795 first and second names were available for classification from the first two panel waves.

Table 3 shows the frequencies of the countries considered in PASS. The total of 1.610 foreign citizens are mainly from Turkey (31%), the successor states of the former Soviet Union (15%) and former Yugoslavia (10%). Only about 9% of the PASS respondents have a foreign citizenship; whereas the proportion of those born outside of Germany is much higher (17%). Here, people from the former Soviet Union build the largest share (23%) followed by those born in Turkey (about 18%) and in the eastern European neighbor states (15%).[21]

[20] All work on and with the names was done in the IAB in Nuremberg. The names themselves never left the IAB, only the "estimated nationality", estimated by the classification of the name, did. The names were only used for the classification; the authors never had access to the names.

[21] Only about 44% of those born outside of Germany have a foreign citizenship, while about 80% of the foreign citizens were not born in Germany. The level of agreement with nationality, given the country of birth, is highest for people born in former Yugoslavia (88%), followed by those born in Russia (78%) and Turkey (71%). The level of agreement with the country of birth, given the nationality, is highest for the eastern European citizens (93%), followed by the Greeks (90%), Italians and Turkish citizens (each 74%).

Table 3: Number of people and nationalities in PASS

	nationality		native country	
	people	in %	people	in %
Germany	17.140	91.2	15.757	83.8
Turkey	514	2.7	530	2.8
Russia[a]	246	1.3	697	3.7
former Yugoslavia[b]	163	0.9	170	0.9
eastern Europe[c]	114	0.6	444	2.4
Italy	68	0.4	49	0.3
Greece	45	0.2	33	0.2
rest of the world	460	2.5	1.091	5.8
unknown	45	0.2	24	0.1
total	18.795	100.0	18.795	100.0

[a] including countries of the former Soviet Union

[b] including successor states

[c] Poland and eastern European neighbor countries

5.2 Problems of classifying nationality and migration background

To evaluate the name classification, a comparison standard is required, which the automatic classification can be compared to. The PASS questionnaire offers several alternatives: self-reported nationality, self-reported country of birth, (where appropriate) birthplaces of parents or grandparents. The use of each of these variables would be fraught with problems (e.g. see summarizing Diefenbach and Weiß 2006). However, the focus of interest here is the extent of agreement of these criteria for classifying respondents during screening, which could then be followed by more detailed selection steps. The evaluation of the classification method has to be based on the same concept as the training data: in absence of other data sources, this is nationality.[22]

Apart from the inevitable substantive problems of each criterion for migration background or nationality, yet another problem is caused by the

[22]Since more elaborated concepts require the measurement of every single dimension, a survey is needed in which all these dimensions, in addition to the name, are collected from a very large sample so that rare populations can be classified as well. In Germany, datasets of this size can only be produced by Official Statistics for which neither para-data nor names can be obtained by law.

Table 4: Level of agreement between different criteria of migration status

Criteria	proportion agreement	Kappa
nationality – country of birth (all)	0.88	0.54
nationality – country of birth (foreign)	0.78	0.73
foreign nationality – born in a foreign country	0.89	0.53
nationality – origin parents[a] (all)	0.88	0.32
nationality – origin parents[a] (foreign)[b]	0.66	0.59
country of birth – origin parents[a] (all)	0.90	0.53
country of birth – origin parents[a] (foreign)[b]	0.47	0.34
country of origin father – mother (all)	0.99	0.99
country of origin father – mother (foreign)	0.98	0.98

[a] if applicable: origin of the mother

[b] when at least one parent is of foreign origin

change of nationality or name over time. In Germany about 1% foreigners are naturalised each year (with declining tendency). About 50 naturalisations can therefore be observed in PASS between wave 1 and 2. Changes in the last name are possible as well: Most of these are women who change their name after marriage with a German spouse. This process will result in names classified falsely negative.[23]

Table 4 shows the level of agreement between nationality or country of birth of the respondent and nationality or country of origin of his parents.[24] Due to the high proportion of German-borns, in almost 90% of the cases, the nationality reported in PASS matches the country of birth. Taking only foreigners into consideration, more than 75% of the cases are matches between nationality and country of birth, while the rate differs between countries. Considering only whether a person has a German or a foreign citizenship and was born in Germany or elsewhere, the level of agreement is about 90%.[25]

[23]About 10% of the annual marriages in Germany are binational marriages between a German and a non-German spouse (Haug 2010). 8.4% of around 8.700 couples in PASS consist of a German and a non-German partner.

[24]Kappa corrects the percentage of agreement for the proportion which results simply due to the marginal distributions. The value of Kappa is therefore lower than the percentage of agreement (Schnell/Hill/Esser 2011: 395).

[25]For about two thirds of the foreigners, their own nationality matches those of their father or mother. The level of agreement with regard to country of birth is about 47%.

Table 5: Possible results of a name classification

	name classification	
nationality	German	non-German
German	true negative TN	false positive FP
non-German	false negative FN	true positive TP

Due to the high concordance between the variables, the use of self-reported nationality as a criterion for the classification of names seems plausible.

5.3 Measures for the assessment of classification results

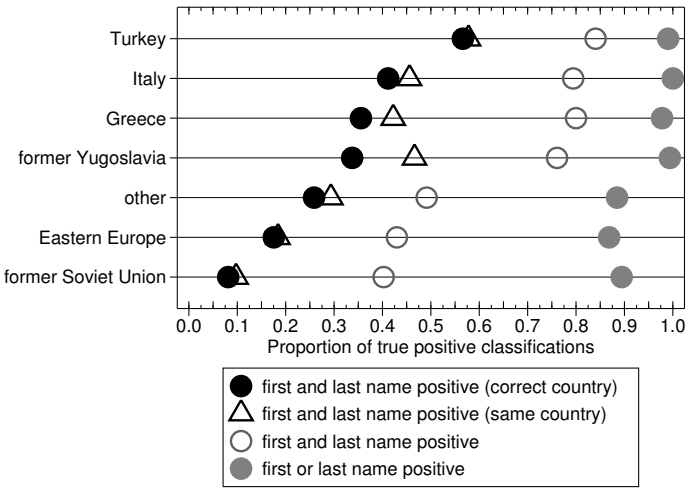
Following a classification of names, four results are possible (table 5): A classification is called true positive (TP) if a person with a foreign name actually has a foreign citizenship.[26] A true negative (TN) classification represents a person with a "German" name and a German citizenship. False positive classifiers (FP) indicate people with a German citizenship, who either have a foreign name or whose name has been falsely classified as foreign by the procedure. This may result in overcoverage.[27] A person is classified falsely negative (FN) if that person has a foreign citizenship, but either does not have a foreign name or this name has wrongly been classified as German. False negative classifications may result in bias due to undercoverage.

Only several of these measures considered together allow an assessment of the quality of the classification, since each criterion taken individually does not describe the classification adequately. Whether a classification method can be considered "good" or "poor" also depends on the consequences of false classifications. For the application discussed here, the effects of incorporating names, which were falsely classified as "foreign" and the effects of those falsely classified as "German" have to be considered. When using a screening method, high ratios of false positives are generally accepted since they can still be excluded in the following analyses. The ratio of true positives among those classified positive should be as high as possible.

[26]It must be noted however, that it is not about classifying "foreign" names, but about classifying the names of those with a foreign nationality. Names themselves can neither be "German" nor "foreign". Still, the automatic screening procedure is based on the idea that certain names will occur more or less often among those with a German citizenship than among those with a different citizenship.

[27]For errors due to under- or overcoverage see Lessler and Kalsbeek (1992).

Figure 2: Proportion of true positive classifications by country and classification rule. Database: PASS panel survey



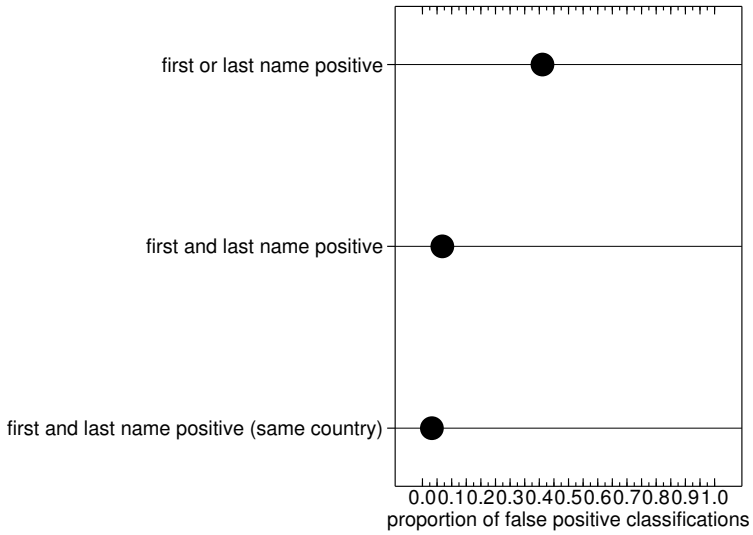
5.4 Combination rules for the classification of first and last names

The new method described above classifies first and last names separately. Therefore a decision has to be taken on how to consider these separate classifications for an overall name classification.[28]

The decision on the type of combination of first and last name classifiers has to be based on the costs of false classifications. A person could be counted as " foreign" if first *or* last name have been classified positive. This procedure will obviously lead to only few false negatives, but many false positive classifications. If however, a person is only considered "foreign" if the first *as well as* the last name are classified positive, more false negatives and less false positives are expected. An even more rigid rule would require that first and last name classifications *belong to the same foreign nationality* in order for the person to be classified as foreigner. Using this rule many false positive classifications could be avoided, but this would be at the cost of risking undercoverage bias due to more false negatives. The effects of different classification rules are examined in the following section.

[28]The problem of combining first and last name classifiers is common to all lexicon-based approaches which use separate dictionaries. This difficulty can be avoided if a training database containing the joint distribution of first and last names is available.

Figure 3: Proportion of false positive classifications (FP) by classification rule. Database: PASS panel survey



5.5 Classification results for the procedure: nationality

The figures 2 and 3 show the proportion of names classified true and false positive given nationality. It also shows different results depending on the classification rule used (combination of first and last names). Over all considered countries, the average proportion of names classified true positive is 90% if only the first or the last name needs to be classified negative (and therefore only produces 100%-90%=10% falsely negative classified persons' names). The proportion of true positive classifications is reduced to 65% if first and last name have to be classified as foreign. The third combination rule (positive classification of first and last name as belonging to the the same nationality) leads to an average proportion of over 35% true positives. Considering the proportion of cases for which first and last name are correctly classified, this still results in 30% correctly classified foreigners.

Differences between the countries become obvious in figure 2. Respondents from Turkey, Italy, Greek or former Yugoslavia show particularly good results in the automatic classification procedure described here. Especially for the more strict classification rules (first and last name classified positive) still about 4 out of 5 members of these countries are correctly classified.

Figure 3 shows the costs of different classification rules: weaker classification rules yield higher proportions of false positive classifications. The worst rule (a positive first or last name's classification is sufficient for the

Table 6: Agreement between classification rule and migration background. FN=first name, LN=last name

classification rule	proportion of agreement	Kappa
FN&LN neg., no migration background	0.67	0.32
FN&LN neg., immigrated themselves	0.34	-0.26
FN&LN neg., min. 1 parent immigrated	0.45	-0.04
FN&LN neg., min. 1 grandparent immigrated[a]	0.47	-0.00
FN|LN pos., no migration background	0.33	-0.30
FN|LN pos., immigrated themselves	0.66	0.29
FN|LN pos., min. 1 parent immigrated	0.55	0.04
FN|LN pos., min. 1 grandparent immigrated[a]	0.53	0.00
FN&LN pos., no migration background	0.19	-0.18
FN&LN pos., immigrated themselves	0.86	0.43
FN&LN pos., min. 1 parent immigrated	0.85	0.11
FN&LN pos., min. 1 grandparent immigrated[a]	0.86	-0.02
FN=LN pos., no migration background	0.21	-0.10
FN=LN pos., immigrated themselves	0.85	0.28
FN=LN pos., min. 1 parent immigrated	0.90	0.11
FN=LN pos., min. 1 grandparent immigrated[a]	0.91	-0.02

[a] Parents were born in Germany

classification of the person as "foreigner") yields almost 40% of people with German citizenship which are incorrectly classified as foreigners.

Any other rule results in less than 10% false positive classifications. The procedure described here yields similar results as the classification based on lists of names.[29] Depending on the target population, the combination rules can be chosen in a way that the proportion of true positives is considerably higher than 75% and the proportion of false positives is less than 10%. Therefore, most target populations of migration surveys in Germany can be sampled using the new classification procedure with much lower screening efforts than with random sampling.

[29]Simon and Kloppenburg (2007: 152) report a false positive rate of about 7% for an onomastic sample of Turkish households.

Table 7: Migration background by classification results of first or last name, column percentages. FN=first name, LN=last name

migration background	FN & LN negative	FN | LN positive	FN & LN positive	FN & LN +identical[1]
no migration background	88.9	57.0	20.5	15.7
foreign citizenship	1.0	17.6	46.1	51.4
born outside Germany	3.4	15.6	21.6	20.3
both parents born outside Germany	0.9	2.7	5.3	6.8
one parent born outside Germany	3.5	4.7	5.1	4.6
foreign language in the household	0.2	0.3	0.5	0.5
all grandparents born outside Germany	0.1	0.1	0.1	0.0
one grandparent born outside Germany	2.1	2.1	0.9	1.5

[1]first and last name classified as belonging to the same foreign nationality

5.6 Classification results for the procedure: migration background

Table 6 shows the proportion of agreement between the classification of first and last name and the migration background (regardless of nationality). The proportion of agreement between the variables "negative classification of first and last name" and "no migration background" is 0.67. The level of agreement between first and last name classified as "German" and their own migration background however, is only 0.34, which is considerably less than would be expected based on the marginal distribution alone (indicated by a negative kappa of -0.26). Did parents or grandparents immigrate, the agreement is still about 0.45. Kappa being close to zero indicates however, that this agreement can be explained solely by the marginal distribution.

As expected, the agreement between classification and migration background is further reduced if only the first or last name is classified positive. Only if the respondents themselves have a migration background, this agreement cannot be explained based on the marginal distribution alone. Are both first and last name classified as foreign or are both first and last name classified as the same nationality, the degree of agreement between the classification and the migration background is over 85%.

Looking at the proportion of false positive and false negative classifications (table 7), a similar result can be observed: the more strict the combination of first and last name classifiers, the higher the proportion of people with a migration background.

Table 8: Proportion of agreement between name classification and nationality by the type of n-grams used

	last name		first name	
country	bigrams	trigrams	bigrams	trigrams
Germany	0.90	0.84	0.87	0.68
Italy	0.69	0.79	0.42	0.51
Turkey	0.63	0.75	0.55	0.77
Greece	0.57	0.60	0.49	0.47
former Yugoslavia[a]	0.48	0.60	0.31	0.52
eastern Europe[b]	0.31	0.36	0.23	0.42
Russia[c]	0.17	0.14	0.33	0.58
total	0.87	0.81	0.83	0.67

[a]including successor states

[b]Poland and eastern European neighbor states

[c]including countries of the former Soviet Union

If first and last name are classified negative (as "German"), only 1% of those classified have a foreign citizenship and only about 7% have a migration background in the broader sense. If however, first or last name are classified positive (as "foreign"), already nearly half of those people actually have a migration background, with about 18% also having a foreign citizenship. A further 16% do not have a foreign citizenship but were born in a foreign country. The remaining 10% do not themselves have a migration background, but are migrants of the second (7,5%) or third generation.

If first and last name are classified positive, four out of five of those classified "positive" have a migration background. Using this rule only 46% are foreign citizens. Applying a more strict rule (first and last name have to be classified as the same foreign nationality) only about 16% do not have a migration background and over 50% have a foreign citizenship. This shows that by choosing appropriate classification rules, efficient screening results can be realized using this procedure.

5.7 On the choice of bigrams or trigrams for classification

In addition to the question on how to combine separate classifications for first and last names, a decision has to be made whether substrings of two or three

letters (bigrams or trigrams) should be used for the classification of names.[30] For a comparison between the classification of bigrams and trigrams, table 8 contains the proportion of correctly classified cases. It becomes clear, that the proportion of correctly classified names of foreigners is generally higher if trigrams being used instead of bigrams. Exceptions are first names of Russians and last names of Greeks.[31]

6 Validation of the procedure using a prospective study with a sample of Turkish names

In 2009, infas (Institute for Applied Social Science, Bonn) conducted a telephone survey among people with a Turkish migration background in Hesse. For this study, the procedure described here was used the first time for sampling. A large sample of last names from a telephone CD of Hesse was classified with this method.[32] Using the telephone numbers belonging to the names classified as "Turkish", 839 interviews were conducted. Since the migration background of the respondents has been asked asked in the survey, this study can be used as another partially validation of the method.

Table 9 shows the migration background up to the third generation for all those persons in the sample (previously classified as "Turkish"). 12% of those classified as Turkish do not have a Turkish migration background, 43% actually have a Turkish citizenship, another 13% do not (or no longer) have the Turkish citizenship, but were born in Turkey. Another 13% do not themselves have a migration background, but at least one of their parents was born in Turkey. Considering the origin of the grandparents and the language spoken in the household, more than two thirds (70,8%) of the 839 respondents classified as Turkish have a migration background in in a more strict sense (the person itself or its parents have a migration background).[33]

For this sample, the classification method has successfully reduced the number of screening contacts needed to generate a large probability sample.

[30] Longer n-grams are hardly suitable for the classification since longer n-grams implies a decrease in error-tolerance for matching names.

[31] With Germans, the splitting of the first and last name into bigrams leads to a higher proportion of correctly (negative) classified cases.

[32] The names classified positive were then screened by a member of the target population and corrected for a few "obvious" false positive cases. The additional screening was due to the requirements of the client of the survey.

[33] Out of the approximately 100 false positives without a Turkish migration background, about a quarter have a different foreign citizenship.

Table 9: Comparison of the classifications of the CATI-study with PASS

classification: Turkish	CATI Hesse %	PASS %
Turkish citizenship	43.4	49.6
born in Turkey	12.8	14.4
both parents born in Turkey	10.7	11.0
one parent born in Turkey	2.0	0.5
language spoken in household: Turkish	1.9	2.0
all grandparents were NOT born in Germany	16.9	10.4
no Turkish migration background	12.3	12.1
number of persons with names classified as "Turkish"	839	960

Table 9 also shows the results in PASS, when first or last names have been classified as "Turkish". The results are very similar.[34] Both samples taken together, only approximately 12% false positive classifications were observed. Overall, the CATI-study in Hesse confirmed the practicability and efficiency of screening using the procedure described.

7 Summary and evaluation

The procedure described here is suitable for efficient screening for foreigners and migrants in lists of names. Depending on the costs of false classifications, more or less strict classification rules can be chosen and, according to the task at hand, primarily false negative or false positive screening results can be avoided.[35]

The automatic name-based procedure described here, is more suited for some populations of migrants than for others. It has shown to be particularly useful for finding immigrants from the traditional immigration countries to Germany (Turkey, Italy, Greece and Yugoslavia). It is also applicable for

[34]This is even more remarkable taking into account that the study based on the PASS data did not use an additional manual review of the cases classified positive but only classified automatically. In PASS, all Turkish citizens have been correctly classified as "Turkish" by the procedure, meaning there were no false negative cases.

[35]The discussion on consequences of false negative classifications using name-based procedures in general or in comparison with other procedures exceeds the scope of this article. A more comprehensive analysis is subject of an other paper of the research group (Schnell et al. 2013b).

other groups (such as Russians and members of the eastern European neighbor countries), but is less efficient for these groups than for the traditional immigration countries. Even for the less suited subpopulations however, the efficiency of this procedure is still higher than it is for a pure random sample. Additionally, smaller selection effects than in quota or snowball procedures can be expected.

The method allows fast screening for foreigners or migrants in a general sampling frame (such as registers of residents or patients, phone books etc.). Another advantage is the avoidance of manual generation of entries in a name dictionary. By splitting the names into n-grams, the method described here is more error-tolerant with respect to the frequent variants of foreign names than a dictionary-based procedure. If suitable training data is available, the procedure is applicable in other countries, for other populations and other selection criteria.

References

Babka von Gostomski, C. (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, editor (2009). *Zuwanderer in Deutschland. Ergebnisse einer repräsentativen Befragung von Menschen mit Migrationshintergrund.* Bertelsmann Stiftung, Gütersloh.

Blane, H. D. (1977). Acculturation and drinking in an italian american community. *Journal of Studies on Alcohol*, 38(7):1324–1346.

Brettfeld, K. and Wetzels, P. (2007). *Muslime in Deutschland. Integration, Integrationsbarrieren, Religion sowie Einstellungen zu Demokratie, Rechtsstaat und politisch-religiös motivierter Gewalt.* Bundeministerium des Innern, Berlin.

Burkhauser, R. V., Kreyenfeld, M., and Wagner, G. G. (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.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Diefenbach, H. and Weiß, A. (2006). Zur Problematik der Messung von „Migrationshintergrund". *Münchner Statistik*, 3:1–14.

Ecob, R. and Williams, R. (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.* Elanders Hungary Kft., Budapest.

Galonska, C., Berger, M., and Koopmans, R. (2004). Über schwindende Gemeinsamkeiten: Ausländer- versus Migrantenforschung. Technical report, Wissenschaftszentrum Berlin (WZB).

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

Haisken-DeNew, J. and Frick, J. R. (2005). *Desktop Companion to the German Socio-Economic Panel (SOEP).* DIW Berlin, Berlin.

Halm, D. and Sauer, M. (2005). Freiwilliges Engagement von Türkinnen und Türken in Deutschland. Projektbericht, Stiftung Zentrum für Türkeistudien an der Universität Duisburg-Essen.

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

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

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

Infas (2009). *Methodenbericht des Projekts Kriminalitätsfurcht in Hessen.* infas Institut für Sozialforschung, Bonn.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: an Introduction to Natural LanguageProcessing, Computational Linguistics, and Speech Recognition.* Pearson, New Jersey, 2 edition.

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

Kalton, G. and Anderson, D. (1986). Rare populations. *Journal of the Royal Statistical Society. Series A*, 149(1):65–82.

Konstantopoulos, S. (2007). What's in a name? In *Proceedings of the Computational Phonology Workshop. 6th International Conference on Recent Advances in NLP*, Bulgarien.

Lauderdale, D. S. (2006). Birth outcomes for arabic-named women in california before and after september 11. *Demography*, 43(1):185–201.

Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling Errors in Surveys.* Wiley, New York.

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

Mateos, P. (2007). A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13(4):243–263.

Passel, J. S. and Word, D. L. (1980). 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, Denver, Colorado.

Perkins, C. 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, M., editor (2007). *Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung „Arbeitsmarkt und Soziale Sicherung“.* IAB-Forschungsbericht (12). Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

Rother, N. (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), editor (2010). *Einwanderungsgesellschaft 2010. Jahresgutachten 2010 mit Integrationsbarometer*. Sachverständigenrat deutscher Stiftungen für Integration und Migration, Berlin.

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

Schnell, R. (2009). Wie man Nadeln in Heuhaufen findet. Stichprobenverfahren für seltene und sehr seltene Bevölkerungsgruppen, Presentation, University of Duisburg-Essen, May 6, 2009; available as video at www.uni-due.de/methods. .

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., and Becher, I. (2012). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. Working Paper Wp-grlc-2012-02, German Record Linkage Center.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., and Becher, I. (2013a). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. *MDA – Methoden - Daten - Analysen*, 7(1):5–33.

Schnell, R., Hill, P., and Esser, E. (2011). *Methoden der empirischen Sozialforschung*. Oldenbourg, München, 9. edition.

Schnell, R., Trappmann, M., and Gramlich, T. (2013b). A Study of Assimilation Bias in Name-Based Sampling of Migrants. *submitted to Journal of Official Statistics*.

Schwartz, K. L., Kulwicki, A., Weiss, L. K., Fakhouri, H., Sakr, W., Kau, G., and Severson, R. K. (2004). Cancer among arab americans in the metropolitan detroit area. *Ethnicity & Disease*, 14(1):141–146.

Shackleford, M. (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.

Simon, E. and Kloppenburg, G. (2007). Das Fernsehpublikum türkischer Herkunft - Fernsehnutzung, Einstellungen und Programmerwartungen. Ergebnisse einer Repräsentativbefragung in Nordrhein-Westfalen. *Media-Perspektiven*, 3:142–152.

Statistisches Bundesamt (2011a). *Bevölkerung und Erwerbstätigkeit. Ausländische Bevölkerung. Ergebnisse des Ausländerzentralregisters*. Statistisches Bundesamt, Wiesbaden.

Statistisches Bundesamt (2011b). *Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund. Ergebnisse des Mikrozensus 2010*. Statistisches Bundesamt, Wiesbaden.

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

Winkler, W. E. (2009). Record linkage. In Pfeffermann, D. and Rao, C., editors, *Handbook of Statistics Vol. 29A, Sample Surveys: Design, Methods and Applications*, pages 351–380. Elsevier, North-Holland, Amsterdam.

IMPRINT

Publisher

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

Template layout

Christine Weidmann