Potential Undercoverage and Bias in Name-based Samples of Foreigners

Rainer Schnell[1)], Tobias Gramlich[1)], Mark Trappmann[2)]

[1)] University of Duisburg-Essen, Duisburg, Germany
[2)] Institute for Employment Research IAB, Nürnberg, Germany

www.methodenzentrum.de and www.iab.de

The 2011 ASA Spring Methodology Conference at Tilburg University

Overview

- The problem of screening for small/special populations
- The dictionary based solution: onomastics
- Problems of this approach
- Evaluation of a name based approach
- Data
- Results

The problem of screening for members of special populations

- In Germany there is no (exhaustive/available) sampling frame in order to sample for foreigners; one solution:
  1. Draw an "ordinary general population sample" and screen at the telephone/the door for members of the target population: expensive; inefficient; large initial sample needed
  2. More efficient and less expensive: use this sampling frame of the general population; screen it for members of the target population *before* ringing the door/calling them; draw a sample only out of the matches.
- What can typically be such a sampling frame? What characteristics one can use for screening for foreigners?

The dictionary approach

- Often, names are part of a sampling frame (telephone directories; lists from the registration office: lists of members' names; name plates; ...).
- Based on the name (first/last/combination) decide whether individual is a member of the target population.
- *Onomastics*: research on names; in this context especially mapping names to their ethnic/linguistic/cultural origin; or nationality.
- Typically, this is done using long lists of known mappings using *name dictionaries* or based on *expert judgment*.

The general problem of screening

- Is screening really effective (i.t.o. increased fraction of target population)?
- Problem: is screening efficient? How good is the performance of such a classification?
- Problem: possible bias due to wrong screening decisions/classifications?

Table 1: classification table

True value	Classification	
	domestic name	*foreign* name
domestic nationality	true negatives *rn*	false positives *fp*
foreign nationality	**false negatives** *fn*	**true positives** *rp*

Bias through screening

- Bias by fp: *overcoverage*, of little concern; decreases efficiency; increases survey costs
- Bias by fn: large problem, *undercoverage*; possible bias; typically remains undiscovered
- $\text{Bias}_{(\text{fn})}$, if false negatives are systematically different from true positives:
  $\text{Bias}_{(\text{fn})} = \bar{Y}_{(rp)} - \bar{Y}_{(fn)}$

Evaluation of onomastic driven samples

- In cooperation with the IAB, Nürnberg: Evaluation of onomastic based samples
- Data: Respondents to the PASS survey (wave 1 to 3)
- True status known (PASS Scientific Use File): nationality, origin, migrational background
- Covariates to study potential bias due to *fn* also come from the PASS SUF.

Respondents to the PASS survey

- PASS survey is conducted yearly since 2006 by the IAB.
- About *labour market and social security* after labor market reforms in Germany.
- Household panel with all persons >=16 years within the household being interviewed.
- Target population: households receiving social welfare benefits and general population households with low economic status.
- Of course, this is no *normal or realistic* sampling frame, but for evaluation of potential bias in onomastic screening this is irrelevant.

Evaluation using the PASS SUF

- About 21000 names of respondents to 3 waves of the PASS panel survey.
- Permission to classify names **within** the IAB (names **never** have left the institute)
- First names and surnames have been classified separately using a naive Bayes' classifier.
- Names have been classified into the largest groups of foreigners in Germany (Turkey, Italy, Greece, former Yugoslavia, Russia, Poland – partly grouped).
- Evaluation of classification using SUF data from wave 1 and 2.

Training data

- Bayes' classifier was trained using names of all employed persons.
- List of names is absolutely anonimized:
  - Separate lists for first and last names
  - Above a certain frequency
  - Names from 2004
  - No other characteristics but frequencies by nationality given first and last names separately.
- In total 112 831 respectively 493 974 different first and last names in **Germany** from these separate lists.
- These lists correspond each to about 30mio persons in Germany.

Classification of names

- We use a different method to classify names.
- No dictionary or expert judgment incorporated!
- *Automatical* classification based on relative frequency of bi- or trigrams .
- Separate judgment for first and last name.
- Result: probability for each country given the first or last name.

Classification of names

- $P_{(Country)}$, $P_{(Name)}$, $P_{(Name|Country)}$ known from name lists, $P_{(Country|Name)}$ results from Bayes' Theorem.
- Additionally: Names split up in parts (n-grams; substrings of length n).
  - e.g. bigrams ($n = 2$) or trigrams ($n = 3$)
  - e.g. 'Peter' consists of the bi- and trigrams {PE,ET,TE,ER} {PET,ETE,TER}
- Classification is based on the relative frequency of n-grams.
- Advantage: allows for errors in names (typos, different spelling, ...)
  (especially important when classifying automatically without manual review).

Results

- Is onomastic screening effective?
- Use bi- or trigrams?
- How to combine the separate judgment of first and last name?
- What about the quality criteria of the classification?
- Is there bias due to false negatives?

Efficiency

- Does onomastic sampling increase fraction of foreign population compared to SRS?
- Fraction of foreigners is at least doubled respectively five times the fraction compared to SRS
(when looking at specific nationalities even higher).

Table 2: Fraction of foreigners...

	foreign nationality		turkish nationality		italian nationality	
	fraction	gain	fraction	gain	fraction	gain
... in population	8.6	–	2.7	–	0.4	–
... if first and last name class. foreign	17.7	x2.1	5.9	x2.2	0.8	x2.0
... if first or last name class. foreign	46.6	x5.4	19.8	x7.3	2.5	x6.3
... if first and last name class. in same foreign nat.	52.1	x6.1	25.7	x9.5	2.7	x6.8

Bi- or trigrams?

Table 3: True positives using bi- or trigrams

	classification ...			
	last name		first names	
Nationality (true value)	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
Yugoslavia[a)]	0.48	0.60	0.31	0.52
Poland[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)] including eastern European neighboring states
[c)] including member states of the former Soviet Union

- Trigrams produce higher proportions of true positive classifications of persons with foreign nationality.
- Different proportions of true positives for different nationalities.

Combination of names? Classify nationality or origin?

Table 4: Quality criteria of classifying first and last names

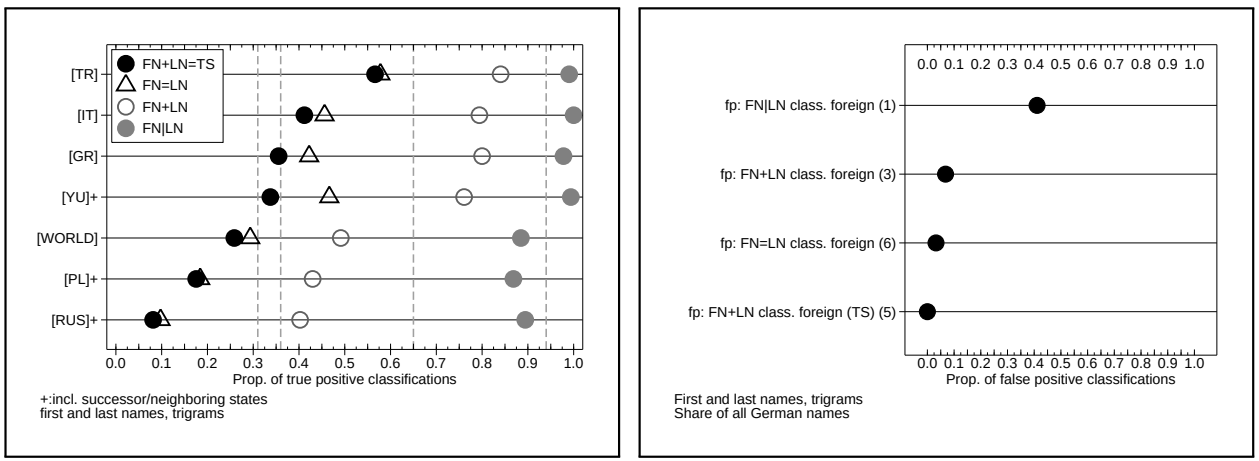
Classification	sensitivity	specitivity	ppv	npv	error rate
	foreign nationality				
only first name	0.57	0.87	0.21	0.97	0.15
only last name	0.68	0.89	0.29	0.98	0.12
First or last name	0.85	0.79	0.20	0.99	0.21
First and last name	0.40	0.97	0.44	0.96	0.06
	foreign origin				
only first name	0.47	0.87	0.30	0.93	0.17
only last name	0.48	0.90	0.35	0.94	0.15
First or last name	0.69	0.80	0.20	0.99	0.21
First and last name	0.26	0.97	0.51	0.92	0.10

ppv: positive predictive value; npv: negative predictive value

- combination of first and last name classification increases rp-rate (first *or* last name classified foreign).
- resp. increases rn-rate (first *and* last name classified foreign).
- overall error rate is lower when classifying nationality rather than origin.

Quality of classification – true positives and false positives

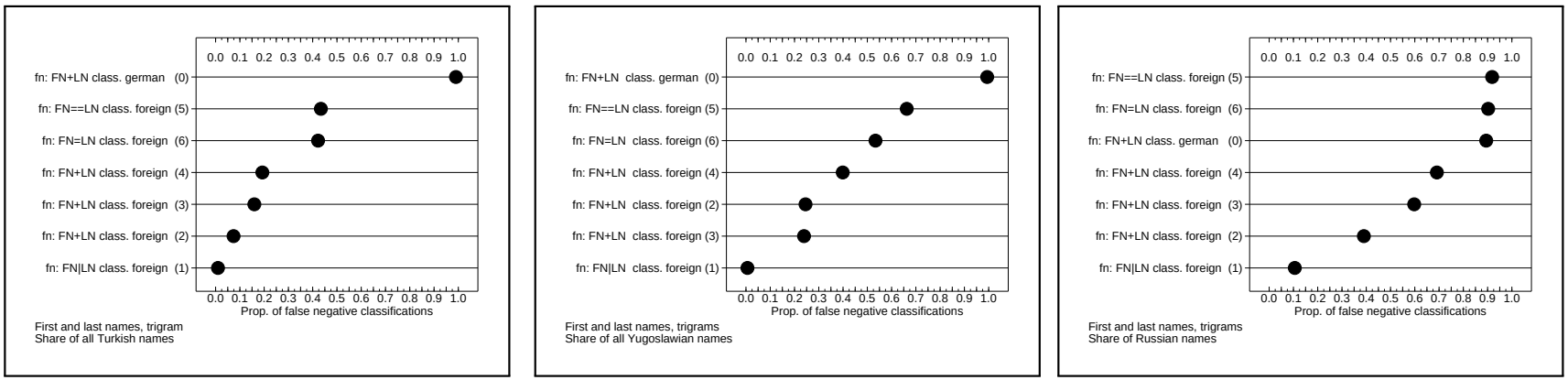
Figure 1: True positive and false positive classifications



- Classification works differently for different nationalities (best for Turkish persons, worst fur Russians)
- Success depends on classification rule (FN *OR* LN classified foreign vs. FN *and* LN match specific nationality).
- Depending on rule proportion of false positive classifications increases.

Quality of classification: false negatives

Figure 2: False negative classifications (TR, YU, RUS)



- Proportion of false negatives lowest if classification of FN *or* LN have to match an unspecific foreign nationality
- Proportion of false negatives highest when classification of FN *and* LN have to match the specific foreign nationality

Efficiency and quality

- Automatic onomastic screening is effective
- Efficiency depends on classification rule
- Quality also depends on classification rule
- Onomastic screening works differently for names from different countries

Bias in demographic variables

Table 5: Bias: sex, age, marital status, and region

	Sex	Age	Marital Status		Region
	% female	years	% married	% single	% East
True value (all foreigners)	52.2	37.6	51.5	21.3	14.0
fn1: FN & LN class. German	+12.2	+5.0	+4.0	+2.5	−5.1
fn2: FN | LN class. German	+7.8	+1.8	+1.9	−2.0	−0.9

- Higher fraction of females and married persons among false negatives
- False negatives are older and less frequent residing in Eastern Germany

Bias in demographic variables

Table 6: Bias: distribution of hh income

	mean	median	p10	p90	gini[a)]
			hh income		
True value (all foreigners)	1474	1300	620	2500	0.304
fn1: FN & LN class. German	+202	1600	652	3500	0.297
fn2: FN | LN class. German	+88	1300	622	2700	0.289
			ind. net income		
True value (all foreigners)	1165	1050	350	2000	0.352
fn1: FN & LN class. German	+533	1400	475	3000	0.343
fn2: FN | LN class. German	+152	1100	389	2000	0.338

[a)] all foreigners, respectively *without* false negatives

- Higher individual and household income among false negatives.
- Concentration of income is overestimated without false negatives.

Bias in labor force status

Table 7: Bias: labor force status

	employed %	unemployed %	maternity leave %
		all foreigners	
True values (all foreigners)	19.7	42.4	3.3
fn1: FN & LN class. German	+10.6	−6.4	+5.7
fn2: FN | LN class. German	+2.4	−2.6	+1.0
		female foreigners	
True values (female foreigners)	12.8	36.0	6.0
fn1: FN & LN class. German	+10.0	−6.2	+8.0
fn2: FN | LN class. German	+3.0	+0.2	+1.1

- Higher proportion of employed persons/on maternity leave among false negatives.
- Lower proportion of unemployed persons.

Bias in education

Table 8: Bias: education (highest degree)

	all foreigners			
	without degree	Haupts.	mittl. Reife	Abitur
True values (all foreigners)	17.0	29.3	20.9	23.8
fn1: FN & LN class. German	−6.1	−13.5	−7.8	+17.8
fn2: FN | LN class. German	−6.3	−3.4	−4.3	+7.0
	female foreigners			
True values (all foreigners)	19.1	24.9	23.0	24.2
fn1: FN & LN class. German	−5.2	−12.6	+1.6	+21.9
fn2: FN | LN class. German	−7.6	−2.7	+4.3	+7.5

- Lower proportion of lower educational degrees among false negatives
- especially women classified false negative often have higher educational degrees.

Bias in religion, subj. indicators

Table 9: Bias: religion, share of Muslims

	member of rel. comm. %	Muslims %	"very religious" %
True values (all foreigners)	74.1	51.8	15.9
fn1: FN & LN class. German	−8.5	−42.0	+1.3
fn2: FN | LN class. German	−6.6	−42.0	−2.7

- Large differences (lower proportions) for false negatives in the membership in religious communities and especially in the proportion of Muslims

Table 10: Bias: subj. indicators (satisfaction with...)

	means of responses on 11-points scale				
	health	flat	living standard	life in general	social participation
True values (all foreigners)	7.1	6.8	5.9	6.4	6.4
fn1: FN & LN class. German	−0.1	+0.4	+0.6	+0.3	+0.1
fn2: FN | LN class. German	+0.0	+0.2	+0.1	+0.2	−0.1

- No/small bias in subjective indicators introduced by false negatives.

Bias in language use

Table 11: Bias: language use

	language different from German used ...				
	in pers. int.	in hh int.	mainly in hh	mainly in hh	with friends
True values (all foreigners)	16.9	17.9	52.4	73.2	44.8
fn1: FN & LN class. German	−7.5	−8.5	−21.0	−26.8	−17.9
fn2: FN | LN class. German	+0.1	+0.2	−2.8	−9.9	−1.6

- Large differences in the language use of false negatives within the household/among friends.
- Lower proportions of languages other than German.

Summary and conclusion

- Onomastic screening and sampling is effective.
- There is a trade off between efficiency and potential bias
- There is large bias due to false negatives, but not necessarily on all variables
- Especially variables connected to *integration* show large biases
- Especially for female foreigners bias due to false negatives is large

Performance of screening

- Several criteria to judge quality of classification:
  1. Sensitivity = rp-rate = $\frac{\text{rp}}{\text{rp}+\text{fn}}$
  2. Specifity = rn-rate = $\frac{\text{rn}}{\text{rn}+\text{fp}}$
  3. positive predictive value = $\frac{\text{rp}}{\text{rp}+\text{fp}}$
  4. negative predictive value = $\frac{\text{rn}}{\text{rn}+\text{fn}}$
- Generally, if *rp* increases also *fp* increases; if *rn* increases also *fn* increases.
- Generally, 1-4 *are unknown*, since true status of classification is not known (at least *rn* and *fn* are typically unknown).

Performance of screening

- 1610 persons with foreign nationality; overall performance
- $RP_{(VN|NN)} = 93.7\%$, $RP_{(VN+NN)} = 63.4\%$, $RP_{(VN=NN)} = 37.5\%$
- $FN_{(VN+NN)} = 6.3\%$, $FN_{(VN|NN)} = 36.6\%$

Data: nationalities in PASS

Table 12: PASS nationalities, wave 1 and 2

Nationality	Persons	in %
Germany	19 341	90.7
Turkey	632	3.0
Russia[a)]	287	1.4
Yugoslavia[b)]	209	1.0
Poland[c)]	127	0.6
Italy	84	0.4
Greece	56	0.3
other World	525	2.5
Total	21 327	100.0

[a)] including former members of the former Soviet Union

[b)] including successor states

[c)] together with eastern European neighboring countries

Data: training data

Table 13: number of first and last names

nationality	number of names in lists: first names: names	first names: persons	last names: names	last names: persons
Germany	58 757	28 309 791	383 592	27 551 167
Yugoslavia[a)]	10 494	313 193	21 973	262 425
Turkey	8 137	605 029	20 835	587 175
Poland[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
other World	25 517	470 446	34 039	286 887
Total	112 831	30 178 813	493 974	29 004 057

[a)] including successor states

[b)] together with eastern European neighboring countries

[c)] including member states of the former Soviet Union

Classification of names

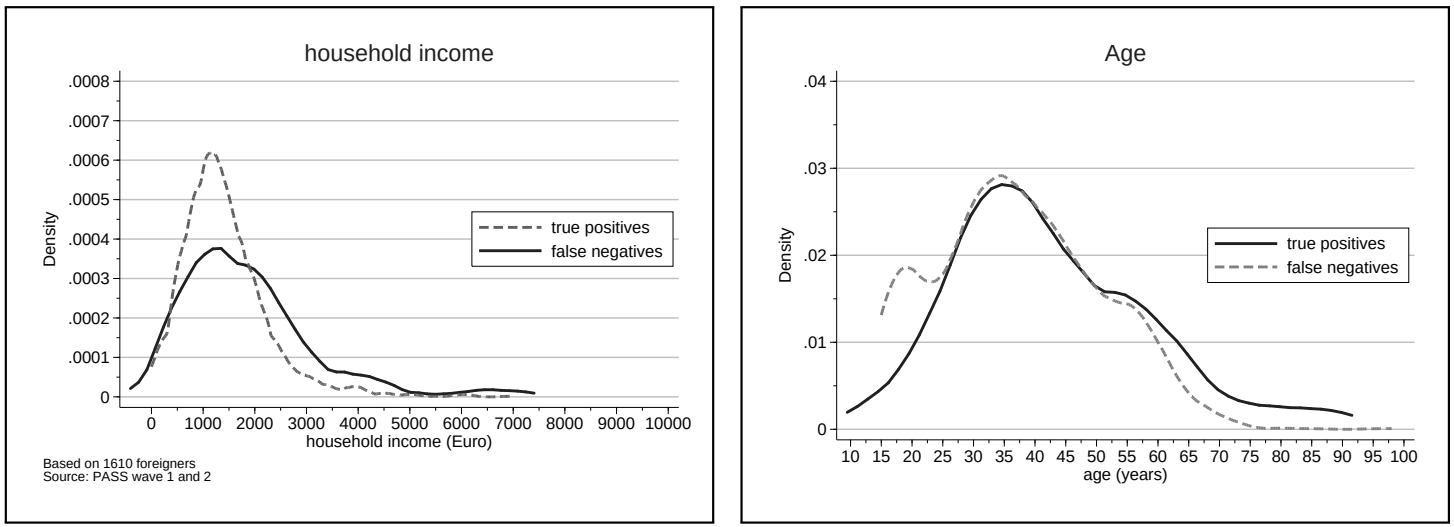
- $P_{(Country)}$, $P_{(Name)}$, $P_{(Name|Country)}$ known from name lists, $P_{(Country|Name)}$ results from Bayes' Theorem
- `classify a name according to the highest probability for a country given this name`

Table 14: example: 'Tobias'

Nationality	Persons 'Tobias'	Persons Total	P_{Name}	$P_{Country}$	$P_{Name\|Country}$	$P_{Country\|Name}$
A	**1000**	15000000	**0.625**	0.980	0.0001	**0.0001**
B	500	150000	0.313	0.015	**0.0033**	**0.0708**
C	50	30000	0.031	0.003	0.0017	0.0177
D	50	20000	0.031	0.002	0.0025	0.0398

Differences in distribution of income, age and household size

Figure 3: Differences in the distribution of net hh income and age



Results: Bias "index of integration"

- PCA with 14 dependent variables (age, sex, employment status, education, language use, hh size, subj. indicators, health indicators)
- 7 PCs with Eigenvalue>=1, screeplot: 1 PC
- Scores from the 1st PC: Are there differences in this "index of integration"?
- Significant differences between all foreigners and false negatives (median test)

Table 15: Bias: index of integration (PC Scores)

	Mean scores on...		
	1. PC	2. PC	3. PC
True values (all foreigners)	0.00	−0.01	−0.29
fn1: FN & LN class. German	+0.33	+0.19	+0.73
fn2: FN | LN class. German	+0.13	+0.05	+0.29