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

FDZ-Nutzerkonferenz, Nürnberg April 2011

Übersicht

- Problemstellung
- Onomastik
- Problem
- Evaluation
- Daten
- Ergebnisse

Problemstellung

- für Ausländerstichprobe in Deutschland existiert kein (allg. zugänglicher) vollständiger Sampling Frame
  1. "normale Bevölkerungsstichprobe" ziehen, an der Haustür/am Telefon screenen und nur Mitglieder der Grundgesamtheit befragen:
     teuer, nicht effizient, u.U. sehr große Ausgangsstichprobe nötig
  2. effizienter und billiger: bereits Sampling Frames der "normalen Bevölkerung" nach der Zielgruppe screenen und Stichprobe aus den "Treffern" ziehen
     Sampling Frames? verfügbare Merkmale zum Screenen?

Onomastik

- oftmals Namen in Sampling Frame verfügbar (Telefonbuch, EWA, Mitgliederlisten, Türschilder ...)
- Entscheidung aufgrund des (Vor-, Nach-) Namens (Kombination) ob Person/Haushalt zur Grundgesamtheit gehört oder nicht
- "Onomastik": Namenskunde; hier insb. Zuordnung eines Namens zu einer ethnischen/sprachlichen Herkunft, Nationalität
- i.d.R. lexikon-basiert

Problem

- ist Screenen effektiv?
- Problem: Effizienz der Klassifikation?
- Problem: Güte der Entscheidung?
- Problem: Bias bei der Zuordnung/Entscheidung der Zugehörigkeit zur Grundgesamtheit?

Tab.: Ergebnis einer Klassifikation

Wahrer Status	Klassifikation	
	dt. Name	ausl. Name
dt. Staatsangehörigkeit	Richtig-negativ *rn*	Falsch-positiv *rp*
ausl. Staatsangehörigkeit	**Falsch-negativ** *fn*	**Richtig-positiv** *rp*

- jeweils alleine nicht aussagekräftig, erst gemeinsam sinnvolle Bewertung

Bias

- Bias durch fp: "Overcoverage", geringeres Problem, verringert Effizienz, erhöht Screeningkosten
- Bias durch fn: schwerwiegendes Problem, Undercoverage, systematischer Bias möglich, bleibt i.d.R. unerkannt
- $\text{Bias}_{(\text{fn})}$ dann, wenn falsch-negative systematisch unterschiedlich sind zu richtig-positiven, d.h. $\text{Bias}_{(\text{fn})} = \bar{Y}_{(rp)} - \bar{Y}_{(fn)}$

Güte und Effizienz

- mehrere Gütekriterien einer Klassifikation:
  1. Sensitivität = Rp-Rate = $\frac{\text{rp}}{\text{rp}+\text{fn}}$
  2. Spezifität = Rn-Rate = $\frac{\text{rn}}{\text{rn}+\text{fp}}$
  3. positiver Vorhersagewert = $\frac{\text{rp}}{\text{rp}+\text{fp}}$
  4. negativer Vorhersagewert = $\frac{\text{rn}}{\text{rn}+\text{fn}}$
- i.d.R. Erhöhung *rp* auf Kosten *fp*, *rn* auf Kosten *fn*,...
- 1-4 i.d.R *aber nicht angebbar*, da wahrer Status unbekannt (zumindest *rn* und *fn*)

Evaluation onomastischer Stichproben

- Kooperation mit dem IAB, Nürnberg: Evaluation onomastischer Stichproben
- Datenbasis: Namen aller PASS-Respondenten, Welle 1 bis 3
- Wahrer Wert bekannt (Nationalität, Herkunft, Migrationshintergrund) aus PASS-SUF
- Kovariaten aus PASS-SUF zur Untersuchung auf Bias durch Falsch-negative

Datenbasis: PASS-Namen

- Respondenten aus PASS, Welle 1 bis 3: rd. 21 000 Namen
- Genehmigung zur Klassifikation **im** IAB (Namen haben IAB **nie** verlassen)
- (separate) Klassifikation Vornamen, Klassifikation Nachnamen, anhand naivem Bayes-Klassifikator
- Klassifikation in größten Ausländergruppen in Deutschland (und im PASS):
  TR, IT, GR, YU, RUS, PL (Länder z.T. in Gruppen zusammengefasst)

Datenbasis: PASS-SUF

- Evaluation der Klassifikation der Namen anhand des PASS-SUF (Wellen 1 und 2)
- Haushaltspanel mit Befragung aller Personen im Haushalt
- Grundgesamtheit: Personen aus ALGII-Leistungsbezieherhaushalte bzw. aus Haushalten mit niedrigerem sozialem bzw. wirtschaftlichem Status
- keine ‘normale’/realistische Grundgesamtheit zur Stichprobenziehung, aber zur Demonstration ist das irrelevant!

Datenbasis: PASS-SUF

Tab.: Nationalitäten im PASS, Welle 1 und 2

Staatsangehörigkeit	Personen	in %
Deutschland	19 341	90.7
Türkei	632	3.0
Russland[a)]	287	1.4
Jugoslawien[b)]	209	1.0
Polen[c)]	127	0.6
Italien	84	0.4
Griechenland	56	0.3
restl. Welt	525	2.5
Insgesamt	21 327	100.0

[a)]: einschl. ehem. Staaten der Sowjetunion

[b)]: einschl. Nachfolgestaaten

[c)]: und osteuropäische Nachbarländer

Datenbasis: Namensdatenbank

- insg. 112 831 Vornamen und 493 974 unterschiedliche Nachnamen **in Deutschland** in separaten Namenslisten
- entspricht Namen von jeweils rd. 30mio Personen in Deutschland
- Namenslisten absolut anonymisiert:
  - Listen getrennt nach Vor- bzw. Nachname
  - nur Vor- bzw. Nachnamen über Mindesthäufigkeit
  - Alter der Namenslisten
  - keine Merkmale ausser Vor- bzw. Nachname, Nationalität, Häufigkeit

Datenbasis: Namensdatenbank

Tab.: Anzahl Vor- und Nachname in Namenslisten

Staatsangehörigkeit	Anzahl Namen in Namensliste			
	Vornamen		Nachnamen	
	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
Polen[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
restl. Welt	25 517	470 446	34 039	286 887
Insgesamt	112 831	30 178 813	493 974	29 004 057

[a)]: ehem. Jugoslawien und Nachfolgestaaten

[b)]: Polen und osteuropäische Nachbarländer

[c)]: Russland und ehem. Staaten der Sowjetunion, Russische Föderation

Klassifikation der Namen

- kein Lexikonabgleich!
- automatische Klassifikation aufgrund *relativer* Bi- bzw. Trigramm-Häufigkeiten
- jeweils für Vor- und Nachname getrennt
- Ergebnis: W'keit für jede Nationalität

Klassifikation der Namen

- $P(Land)$, $P(Name)$, $P(Name|Land)$ bekannt aus Namensliste, $P(Land|Name)$ durch Umformung Bayes'-Theorem
- `Klassifiziere den Namen in diejenige Nationalität, für die die relative Häufigkeit des Auftretens des Namens gegeben eine Nationalität maximal ist`

Tab.: fiktives Beispiel: 'Tobias'

Nationalität	Anzahl Personen 'Tobias'	insgesamt	$P(Name)$	$P(Land)$	$P(Name\|Land)$	$P(Land\|Name)$
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

Klassifikation der Namen

- zusätzlich: Zerlegung der Namen in n-Gramme (Teilstrings der Länge n)
- bspw. Bigramme ($n = 2$) oder Trigramme ($n = 3$)
- bspw. 'Peter' hat die Bi- bzw-. Triramm-Menge {PE,ET,TE,ER} {PET,ETE,TER}
- Klassifikation also auf Basis der relativen n-Gramm-Häufigkeiten
- Vorteil: kein exakter Abgleich mit Namensliste, sondern fehlertolerant (wichtig insbesondere für vollautomatisches Screenen ohne man. Review)

Ergebnisse

- screenen effektiv?
- Bi- oder Trigramme?
- Vor- oder/und Nachname?
- Vergleich der Gütekriterien
- Bias?

Ergebnisse: Effektivität

- bei Screenen höherer Anteil von Elementen der Zielpopulation als in SRS?
- Verdoppelung bzw. verfünffachter Anteil der Zielpopulation durch onomastisches Screenen (bei spez. Gruppen noch darüber hinaus!)

Tab.: Anteil Ausländer...

	ausl. Staatsang.		türk. Staatsangeh.		ital. Staatsang.	
	Anteil	Gewinn	Anteil	Gewinn	Anteil	Gewinn
... in Grundgesamtheit	8.6	–	2.7	–	0.4	–
... wenn Vor- & Nachname ˜=D	17.7	x2.1	5.9	x2.2	0.8	x2.0
... wenn Vor- | Nachname ˜=D	46.6	x5.4	19.8	x7.3	2.5	x6.3
... wenn Vor- = Nachname ˜=D	52.1	x6.1	25.7	x9.5	2.7	x6.8

Ergebnisse: Bi- oder Trigramme?

Tab.: Übereinstimmung (RP) Klassifikation Nachname mit Nationalität, n-Gramme

	Klassifikation ...			
	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
Polen[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)]: einschl. Nachfolgestaaten
[b)]: und osteuropäische Nachbarländer
[c)]: einschl. ehem. Staaten der Sowjetunion

Ergebnisse: Vor- oder/und Nachnamen? Nationalität oder Herkunft?

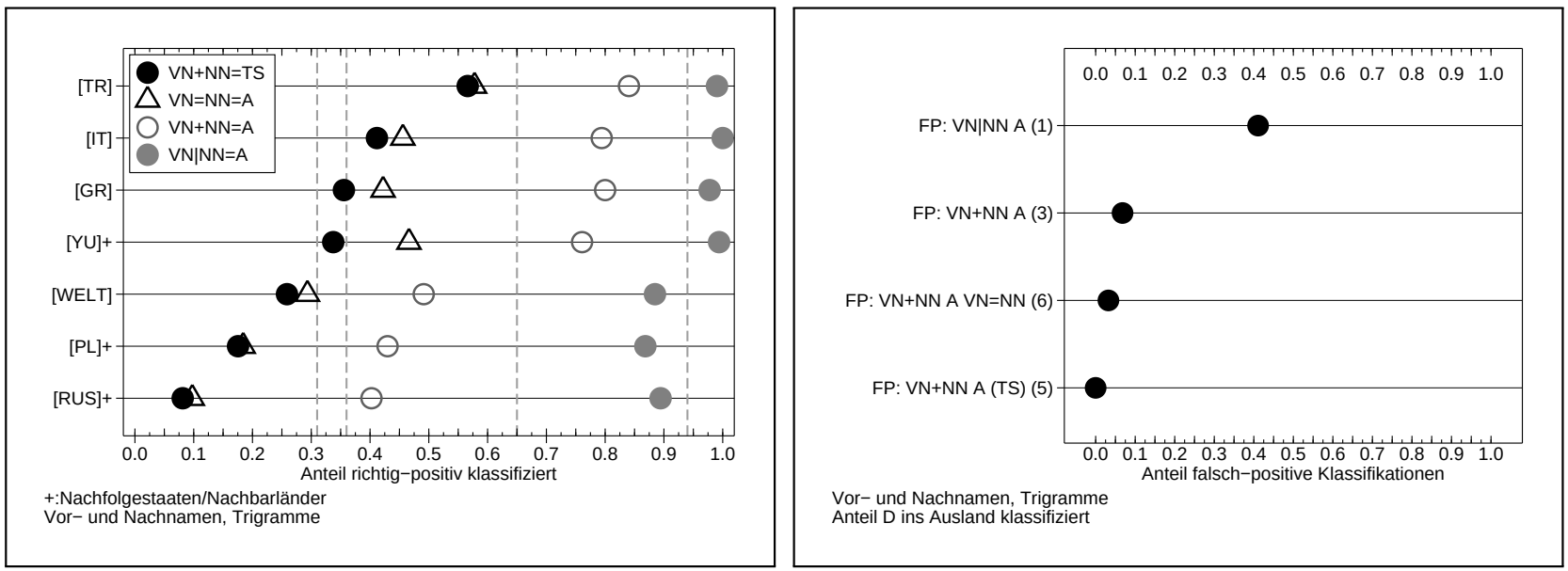
Tab.: Güte der Klassifikation von Vor- und Nachnamen

Klassifikation	Sensitivität	Spezifität	PPV	NPV	Irrtums'wkeit
	ausl. Staatsangehörigkeit				
Vorname allein	0.57	0.87	0.21	0.97	0.15
Nachname allein	0.68	0.89	0.29	0.98	0.12
Vor- oder Nachname	0.85	0.79	0.20	0.99	0.21
Vor- und Nachname	0.40	0.97	0.44	0.96	0.06
	außerhalb Deutschlands geboren				
Vorname allein	0.47	0.87	0.30	0.93	0.17
Nachname allein	0.48	0.90	0.35	0.94	0.15
Vor- oder Nachname	0.69	0.80	0.20	0.99	0.21
Vor- und Nachname	0.26	0.97	0.51	0.92	0.10

PPV: positiver Vorhersagewert; NPV: negativer Vorhersagewert

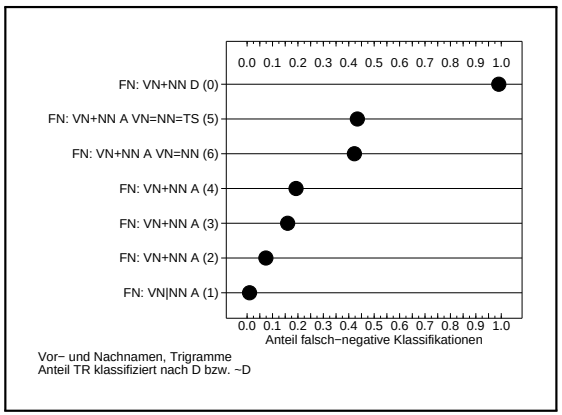
Ergebnisse: Klassifikation: Richtig-positive, falsch-positive Klassifikationen

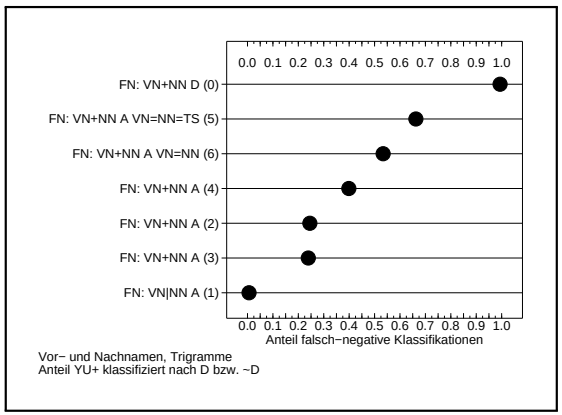
Abb.: Richtig-positive, falsch-positive Klassifikationen

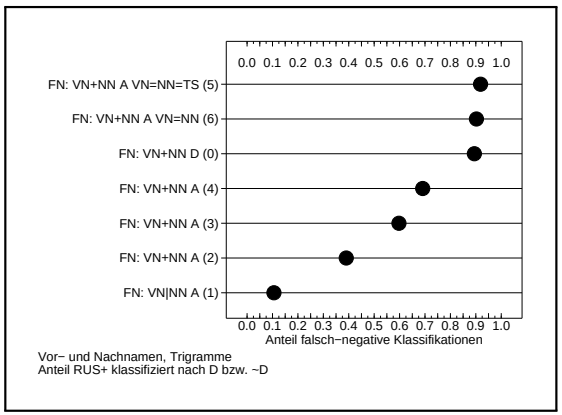


Ergebnisse: Güte der Klassifikation: falsch-negative Klassifikationen

Abb.: falsch-negative Klassifikationen, TR, YU, RUS







Ergebnisse: Effizienz und Güte

- (vollautom.) onomastisches Screenen ist effizient
- Effizienz unterscheidet sich je nach Klassifikationsregel
- Güte der Klassifikation unterscheidet sich nach Klassifikationsregel
  (und je nach Kosten von Falschklassifikationen)
- Screenen funktioniert für verschiedenen Nationalität unterschiedlich gut

Ergebnisse: Bias Demographie

Tab.: Bias (Geschlecht, Alter, Familienstand und Region)

	Geschlecht	Alter	Fam.stand		Region
	% weibl.	Jahre	% verh.	% ledig	% Ost
alle Ausländer (wahrer Wert)	52.2	37.6	51.5	21.3	14.0
FN1: VN & NN D	+12.2	+5.0	+4.0	+2.5	-5.1
FN2: VN | NN D	+7.8	+1.8	+1.9	-2.0	-0.9

Tab.: Bias: Haushaltsgröße, Wohnungsgröße, PKW, Haushaltsnettoeinkommen

	Haushaltsgröße	Wohnungsgröße	PKW im Haushalt	HH-Nettoeinkommen
	Personen	m^2	in %	€
alle Ausl. (wahrer Wert)	3.3	75.8	46.7	1474
FN1: VN & NN D	-0.5	+9.2	+9.8	+202
FN2: VN | NN D	-0.3	-3.6	-0.5	+88

Ergebnisse: Bias Demographie

Tab.: Bias: Verteilung Einkommen

	Durchschn.	Median	p10	p90	Gini[a)]
		HH-Nettoeinkommen			
alle Ausl. (wahrer Wert)	1474	1300	620	2500	0.304
FN1: VN & NN D	+202	1600	652	3500	0.297
FN2: VN | NN D	+88	1300	622	2700	0.289
		ind. Nettoeinkommen			
alle Ausl. (wahrer Wert)	1165	1050	350	2000	0.352
FN1: VN & NN D	+533	1400	475	3000	0.343
FN2: VN | NN D	+152	1100	389	2000	0.338

[a)]: alle Ausländer bzw. *ohne* FN1 bzw. FN2

Ergebnisse: Bias Erwerbsstatus

Tab.: Bias: Erwerbsstatus (alle Ausländer)

	erwerbstätig Anteil	arbeitslos Anteil	Mutterschutz/Elternzeit Anteil
		alle Ausländer	
alle Ausländer (wahrer Wert)	19.7	42.4	3.3
FN1: VN & NN D	+10.6	-6.4	+5.7
FN2: VN | NN D	+2.4	-2.6	+1.0
		weibl. Ausländer	
alle Ausländer (wahrer Wert)	12.8	36.0	6.0
FN1: VN & NN D	+10.0	-6.2	+8.0
FN2: VN | NN D	+3.0	+0.2	+1.1

Ergebnisse: Bias Schulbildung

Tab.: Bias Bildung (höchster Schulabschluss)

	ohne Abschluss	Haupts.	mittl. Reife	Abitur
		alle Ausländer		
alle Ausländer (wahrer Wert)	17.0	29.3	20.9	23.8
VN + NN D (*fn1*)	-6.1	-13.5	-7.8	+17.8
VN | NN D (*fn2*)	-6.3	-3.4	-4.3	+7.0
		weibl. Ausländer		
alle Ausländer (wahrer Wert)	19.1	24.9	23.0	24.2
VN + NN D (*fn1*)	-5.2	-12.6	+1.6	+21.9
VN | NN D (*fn2*)	-7.6	-2.7	+4.3	+7.5

Ergebnisse: Bias Religion, und Zufriedenheiten

Tab.: Bias: Religion, Religiosität, Anteil Muslime)

	Mitglied Religionsgem. Anteil	Muslime Anteil	"sehr religiös" Anteil
alle Ausländer (wahrer Wert)	74.1	51.8	15.9
FN1: VN & NN D	-8.5	-42.0	+1.3
FN2: VN | NN D	-6.6	-42.0	-2.7

Tab.: Bias subjektive Indikatoren (Zufriedenheiten)

	Durchschnitt auf Skala 0-10				
	Gesundheit	Wohnung	Lebensstandard	Leben allg.	soz. Teilhabe
alle Ausländer (wahrer Wert)	7.1	6.8	5.9	6.4	6.4
FN1: VN & NN D	-0.1	+0.4	+0.6	+0.3	+0.1
FN1: VN | NN D	+0.0	+0.2	+0.1	+0.2	-0.1

Ergebnisse: Bias Sprache und Migrationshintergrund

Tab.: Bias: Sprache

			andere Sprache als Deutsch...		
	im P-Int.	im HH-Int.[a)]	überw. im HH[b)]	überw. im HH[c)]	im Freundeskreis[c)]
alle Ausländer (w.W)	16.9	17.9	52.4	73.2	44.8
FN1: VN + NN D	-7.5	-8.5	-21.0	-26.8	-17.9
FN2: VN | NN D	+0.1	+0.2	-2.8	-9.9	-1.6

[a)]: Sprache des Haushaltsinterviews (HHV)
[b)]: lt. Auskunft des HHV im Haushaltsfragebogen
[c)]: lt. Auskunft im Personenfragebogen

Tab.: Bias: Migrationsgeschichte)

	selbst zugez.	Eltern[a)]	Eltern[b)]
alle Ausländer (wahrer Wert)	83.9	84.3	81.9
FN1: VN & NN D	-0.2	+3.2	-7.6
FN2: VN | NN D	+3.3	-1.4	-2.7

[a)]: mindestens 1 Elternteil nicht in D. geboren
[b)]: beide Elternteile nicht in D. geboren

Ergebnisse: Bias "Integrationsindex"

- Hauptkomponentenanalyse mit 14 unabh. Variablen (Alter, Geschlecht, Erwerbsstatus, Bildung, Interviewsprache, Haushaltsgröße, subj. Bewertungen, Gesundheit)
- 7 PC mit Eigenwert >=1, Screeplot: 1 PC
- Scores auf der ersten Hauptkomponente als "Integrationsindex"; Unterschiede?
- Unterschiede zwischen allen Ausländern und FN jeweils signifikant (Mediantest)

Tab.: Bias Integrationsindex (Scores Hauptkomponentenanalyse)

	Hauptkomponentenanalyse, Scores ...		
	1. HK	2. HK	3. HK
alle Ausländer (wahrer Wert)	0.00	-0.01	-0.29
FN1: VN & NN D	+0.33	+0.19	+0.73
FN2: VN | NN D	+0.13	+0.05	+0.29

Schlussfolgerung

- Screening durch onomastische Klassifikation ist effektiv
- Trade off zwischen Effizienz und möglichem Bias
- z.T. großer Bias durch falsch-Negative, aber nicht notwendigerweise auf allen Variablen
- insbesondere Variablen, die mit "Integration" zusammenhängen weisen z.T. sehr großen Bias auf ("harte" Variablen: Erwerbsstatus, Bildung, Einkommen; weniger bei "weichen", subjektiven Indikatoren)
- insbesondere bei Frauen ist der Bias durch falsch negative Klassifikation groß