Die Berechnungsmethode zur Ermittlung der Nichtanerkennungsquote nach der Registerzählung 2011

Statistik Austria
Fachbereich Registerzählung

Kurzfassung

Im Rahmen der Probezählung 2006 wurde ein Verfahren entwickelt, welches für den nachfolgenden jährlich im Rahmen der Statistik des Bevölkerungsstandes gemäß § 10 (7) FAG 2017[1] durchzuführenden Registerabgleich, den Anteil der „Karteileichen" an der Menge von Verdachtsfällen schätzt. Diese sogenannte Nichtanerkennungsquote ergab sich aus einer komplexen Rechenvorschrift, welche auf den Erfahrungen bei der Durchführung der Wohnsitzanalyse für das Jahr 2006 basierte.

Mit der nunmehr für den Stichtag 31.10.2011 durchgeführten Wohnsitzanalyse der Registerzählung wurde mit der Erfahrung und dem Datenbestand von mittlerweile 5 Jahren ein Modell entwickelt, welches die Wahrscheinlichkeit für das Vorliegen einer „Karteileiche" auf Basis eines statistischen Standardverfahrens erlaubt.

Dieses Modell weist all jenen Verdachtsfällen hohe Nichtanerkennungswahrscheinlichkeiten zu, welche in ihren demographischen Merkmalen jenen Datensätzen sehr stark ähneln, die sich im Rahmen der vergangenen Wohnsitzanalysen als Löschfälle erwiesen haben.

[1] Gültige Rechtsgrundlage bis zum Stichtag 31.10.2015: Finanzausgleichsgesetz 2008 § 9 (9).

1 Einleitung

Im Zuge der Probezählung 2006 wurde erstmals das Konzept der Lebenszeichenanalyse zur Ermittlung der tatsächlichen Bevölkerungszahl entwickelt [1] und bei der jährlichen Bevölkerungsfeststellung für den Finanzausgleich sowie der Registerzählung 2011 erfolgreich angewandt.

Dabei wird ausgehend vom Zentralen Melderegister versucht, jede Person mit einer gültigen Hauptwohnsitzmeldung durch andere Register zu bestätigen. Diese als sogenannte Lebenszeichen titulierten Einträge in Registern, wie beispielsweise jenem des Hauptverbands der Sozialversicherungsträger, des Arbeitsmarktservice oder der Dienstgeberdaten des Bundes und der Länder, erheben in Verbindung mit dem Zentralen Melderegister einen derart starken Bestätigungsanspruch, dass diese Datensätze unmittelbar als Wohnbevölkerung der betreffenden Gemeinde gezählt werden, sofern sie nicht unter die gesetzliche 90- oder 180 – Tage – Regel fallen („Technische Nichtanerkennungen")[2].

Die Komplementärmenge zur oben beschriebenen Hauptgruppe an Datensätzen („Zählfälle") stellt zwar einen kleinen, jedoch sehr untersuchungswürdigen Personenkreis dar: Es handelt sich demnach um Hauptwohnsitzmeldungen zu welchen im Rahmen des registerbasierten Erhebungsverfahrens keinerlei bestätigende Einträge in anderen Registern gefunden werden konnten. Diese im Sprachgebrauch der Registerzählung als „Klärfälle" bezeichneten Datensätze stellen zu einem gewichtigen – aber eben nicht zum vollständigen – Teil Löschfälle dar, welche den betreffenden Gemeinden als Wohnbevölkerung abzuerkennen sind.

Im Zuge der Probezählung 2006 und auch der Registerzählung 2011 wurde von Statistik Austria an diese Personengruppe per RSB – Verfahren[3] eine Benachrichtigung

[2] Dabei handelt es sich um Bestimmungen des Registerzählungsgesetzes [2]: 90-Tage-Regel (§ 7 Abs. 3): Personen werden zum Stichtag gezählt, wenn sie sich um den Stichtag herum mehr als 90 Tage in Österreich aufhalten. 180-Tage-Regel (§ 7 Abs. 2): Personen, die um den Stichtag herum weniger als 180 Tage in einer Gemeinde gewohnt haben und danach wieder in die Gemeinde zurückkehren, aus der sie vorher gekommen sind, werden nicht in der Stichtagsgemeinde gezählt, sondern in der Gemeinde des früheren und auch späteren Hauptwohnsitzes.

[3] Ein RSB – Brief („Rückscheinbrief") ist ein behördliches Schriftstück, das einer Bestätigung durch einen Empfänger bedarf.

versandt, in welcher um eine Bestätigung des fraglichen Hauptwohnsitzes ersucht wurde. Hierdurch konnten die Löschfälle unmittelbar festgestellt werden.[4]

Für die außerhalb einer Registerzählung statt findenden Bevölkerungsfeststellungen ergibt sich damit die Ausgangslage, dass zwar die Menge an fraglichen Klärfällen alljährlich ermittelt werden kann, aus diesen jedoch die Menge an Löschfällen möglichst exakt – allerdings ohne eine tatsächliche Briefbefragung – abzuleiten ist. Wie bereits bei Festlegung der Formel zu Nichtanerkennungen von Hauptwohnsitzen für die Finanzausgleichsjahre 2009 und 2010 angedacht, wird der Status einer Person in einem Jahr durch eine Markovkette modelliert und die Wahrscheinlichkeit der Löschung durch statistische Klassifikation (logistische Regression) geschätzt. Daraus wird dann die gemäß Finanzausgleichsgesetz zu ermittelnde Bevölkerungszahl berechnet.

2 Definitionen und Begriffe

Technische Nichtanerkennungen (TNA): Personen im ZMR, für die Verstorbenenmeldungen vor dem Stichtag vorliegen bzw. die unter die Regelungen nach § 7 Abs. 2 oder 3 Registerzählungsgesetz fallen, werden nicht anerkannt. Dies sind 90-Tage-Fälle, Kit-Fälle, 180-Tage-Fälle.

Zählfall (ZF): Personen mit einer gültigen Hauptwohnsitzmeldung, die durch andere Register bestätigt sind.

Klärfall (KF): Personen mit einer gültigen Hauptwohnsitzmeldung, die durch andere Register NICHT bestätigt sind.

Löschfall (LF): Personen aus der Menge der Klärfälle, für die keine Bestätigung des Hauptwohnsitzes vorliegt. Diese werden auch als „Karteileichen" bezeichnet.

Anerkennungen (A): Personen aus der Menge der Klärfälle, für die letztendlich eine Bestätigung des Hauptwohnsitzes vorliegt bzw. die nach Anwendung des statistischen Modells anerkannt werden.

Nichtanerkennungen (NA): Personen aus der Menge der Klärfälle, für die keine Bestätigung des Hauptwohnsitzes vorliegt bzw. die nach Anwendung des statistischen Modells nicht anerkannt werden.

[4] Aufgrund des Registerzählungsgesetzes wurde den betroffenen Gemeinden auch die Möglichkeit gegeben, für die so ausgelisteten zweifelhaften Hauptwohnsitze noch andere Lebenszeichen zu erbringen (schriftliche Bestätigungen der BürgerInnen). Der Vorgang kann folglich als recht erschöpfend bezeichnet werden.

3 Ausgangslage

Technisch gesehen steht damit alljährlich eine Menge an Datensätzen mit exakten Adressdaten (und damit verbunden, in aller Regel einer recht genauen Erfassung der Wohnsituation) sowie einer Reihe von demographischen Personenmerkmalen zur Verfügung. Letzteres allerdings mit der Einschränkung der Datenlage, welche das Zentrale Melderegister in Bezug auf die Demographie zu bieten hat, da es sich dabei exakt um jene Meldesätze handelt, zu denen das registerbasierte Verfahren keine weiteren Informationen anreichern konnte.

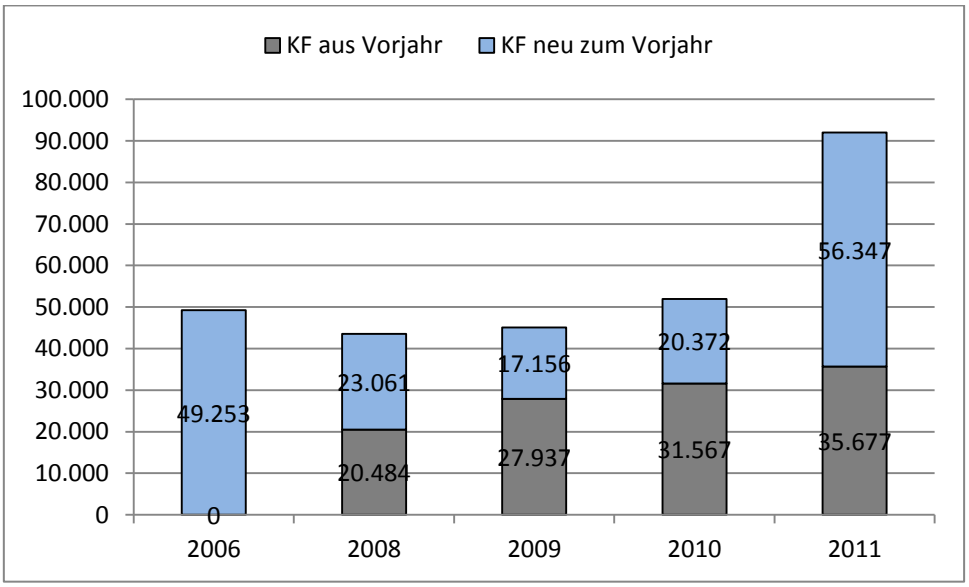

Abbildung 1: Entwicklung der Klärfälle sowie des Anteils an im Vergleich zum Vorjahr neuen Meldungen in den Jahren 2006 - 2011

Allein aus diesem jährlich erhaltenen Bestand lassen sich bereits bei bloßer zeitlicher Betrachtung, gewisse Schlüsse ziehen: prüft man, welche Datensätze bereits im Vorjahr ohne weitere Lebenszeichen (und damit ebenfalls Klärfälle) waren, so lässt sich unmittelbar feststellen, dass ein recht hoher Stock an unveränderten, alljährlich wiederkehrenden Verdachtsfällen in dieser Masse aufscheint.

Zusätzlich zu diesem Bestand existieren darüber hinaus die fraglichen Klärfälle aus den Erhebungen 2006 und 2011 mit der Markierung all jener Datensätze, welche schließlich tatsächlich zu löschen waren. Dieser historische Datenbestand bietet damit eine umfassend aufdeckende Sicht auf die andernfalls verborgenen Charakteristi-

ka[5] dieses Datenbestandes. Wie die modellhafte Grafik zeigt, verbleibt somit als zu ermittelnde Größe die Übergangswahrscheinlichkeit, dass ein Klärfall zu löschen ist.

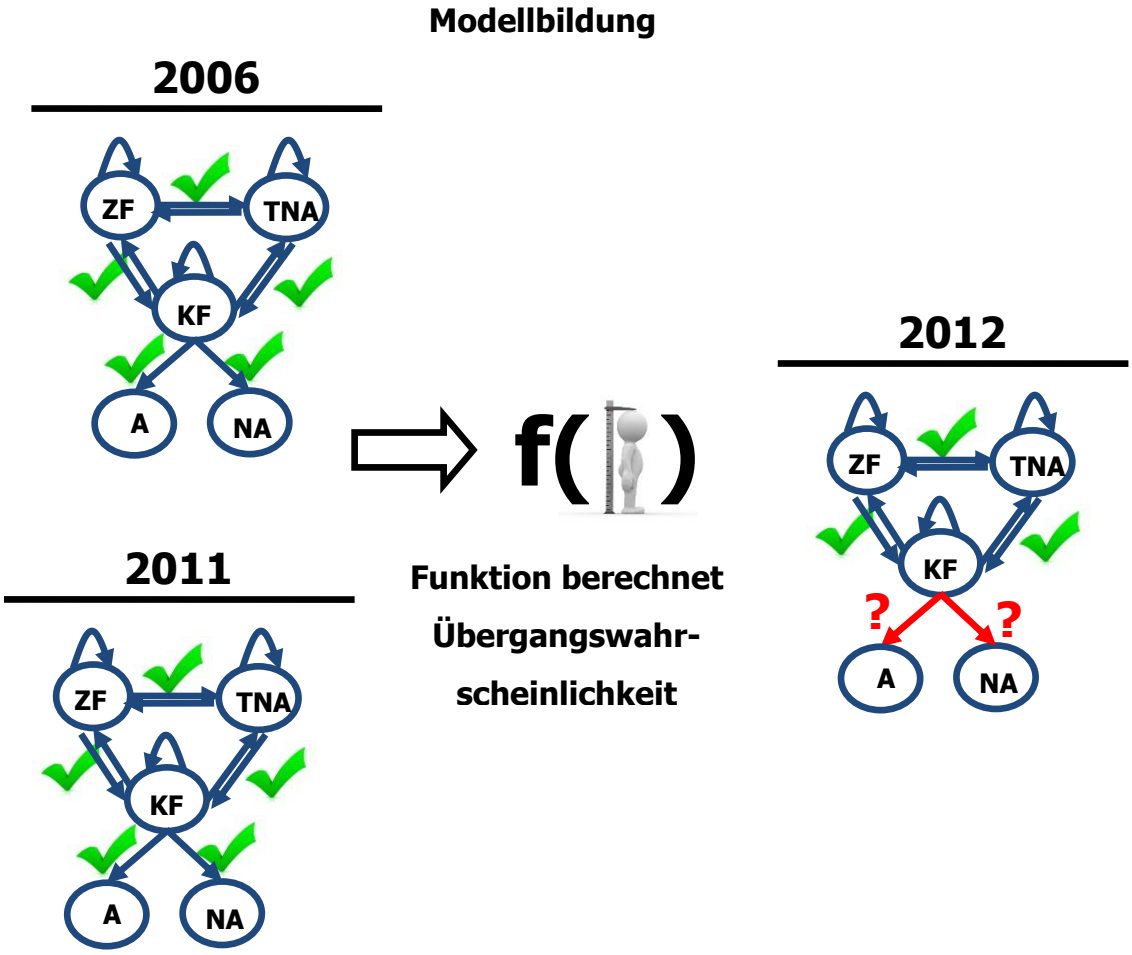

Abbildung 2: Veranschaulichung von Zustandsübergangsdiagrammen im Rahmen der neuen Nichtanerkennungsquote

Mit Hilfe dieses Datenbestandes von etwa 140.000 Einzeldatensätzen lassen sich nun Rückschlüsse auf die verborgene Struktur der Löschfälle ziehen, indem alle enthaltenen Merkmale auf ihre Eigenschaft hin untersucht werden, ob sie Anerkennungen von Nichtanerkennungen zu trennen vermögen.

[5] Dies deshalb, da zwischen den Erhebungen keine vollständig aufklärenden Briefbefragungen erfolgt sind.

4 Trennvariablen in der historischen Datenbasis

Bei der Suche nach Trennvariablen unter den Merkmalen der historischen Datenbasis, ist es wichtig festzustellen, dass hierbei die Verteilung der Löschwahrscheinlichkeit der Gesamtgruppe eine vitale Rolle spielt: nur Eigenschaften, welche die Datenmenge in Gruppen unterteilen, die sich in Bezug auf das Löschmerkmal deutlich von der Gesamtheit unterscheiden, können verwendet werden.[6]

Grundsätzlich teilt sich die Menge an Klärfällen so auf, dass auf jeden anerkannten Datensatz beinahe zwei Nichtanerkennungen[7] kommen.

Anzahl der Klärfälle:	**141.277**
davon wurden gelöscht:	92.665
davon wurden gezählt:	48.612

Tabelle 1: Die Aufgliederung der Klärfälle 2006 und 2011 nach ihrem tatsächlichen „Nichtanerkennungsstatus", wie er in den beiden Briefbefragungen ermittelt wurde

Auf Basis der Analyse aller verfügbaren potentiellen Trennvariablen ist es möglich, sämtliche Merkmale formal auf ihr Löschverhältnis im historischen Datenbestand hin zu untersuchen. Dabei können einfache Thesen wie „Unterscheiden sich männliche von weiblichen Datensätzen?" oder „Gibt es Unterschiede hinsichtlich des Geburtslandes?" genauso recht unmittelbar beantwortet werden, wie erweiterte Fragestellungen, nach der Frage des Einflusses des Alters zum Zeitpunkt des Zuzuges oder ähnliches.

Dazu ist es lediglich notwendig die Ausprägungen des betrachteten Merkmals gegen das Löschverhältnis dieser Untergruppe der Gesamtdatenbasis zu betrachten. Für eine statistische Modellbildung ist es notwendig Merkmalsausprägungen (oder Kombinationen davon) zu finden, die eine möglichst starke Abweichung in Bezug auf das Löschverhältnis von der Gesamtmasse aufweisen.

Der Übergang zur eigentlichen Nichtanerkennungsquote ist für die Menge an Klärfällen direkt durchzuführen: Es wird lediglich eine Abbildung gesucht, welche den mit

[6] Wenn im Folgenden von einem Löschmerkmal die Rede ist, dann ist damit stets jene Markierung der Datensätze gemeint, welche erkenntlich macht, dass sich der zugrundeliegende Meldedatensatz im Rahmen der Briefbefragungen 2006 und 2011 als „Löschfall" entpuppt hat.

[7] Genauer formuliert 92.665 / 48.612 = 1,906

Hilfe von Trennvariablen repräsentierten in Frage kommenden Datensätzen einen einzelnen Wert zwischen 0 und 1 zuordnet, der die Wahrscheinlichkeit für das Vorliegen eines „Löschfalls" möglichst gut schätzt. Diese Problemstellung wurde sehr unmittelbar mit einer klassischen, logistischen Regression gelöst.

5 Logistisches Regressionsmodell

Die Entscheidung für ein Regressionsmodell fußte auf einer Reihe von Gründen. Die vorliegende historische Datenbasis eignet sich in ihrer Ausprägungsstruktur gut für einen derartigen Ansatz.
Nachvollziehbarkeit und Transparenz stellen im Rahmen der Registerzählung bei allen Überlegungen stets wichtige Leitgedanken dar: Das Verfahren eines linearen, logistischen Modells stellt ein Standardwerkzeug aus dem statistischen Methodeninventar dar. Das umgesetzte lineare, logistische Modell hat die Form

$$\operatorname{logit}(\pi) \equiv \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}$$

wobei α die Schätzkonstante und der Vektor β der Form

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)'$$

die gesuchten Koeffizienten für die potentiellen Trennvariablen darstellen.

Erhält man nun in den Folgejahren einen Datensatz mit allen benötigten Trennvariablen, wie dieser im Zuge der registerbasierten Bevölkerungsfeststellung für den Finanzausgleich alljährlich vorliegt, so lassen sich die entsprechenden Wahrscheinlichkeiten auf Einzeldatensatzebene zuordnen.

6 Gliederung der Grundgesamtheit nach Gemeindeklassen

Obgleich der Kern der Ermittlung der Nichtanerkennungsquote der beschriebene Ansatz eines Regressionsmodells ist, finden noch einige verfahrensinhaltliche Modifikationen statt, welche den Datenrealitäten in der Registerzählung geschuldet sind.

Es soll den verschiedenen Gemeindegruppen Rechnung getragen werden und eine getrennte Behandlung erfolgen. Statt lediglich ein einziges Regressionsmodell für

Gesamtösterreich abzuleiten, wurden die Gemeinden in unterschiedliche Gruppen eingeteilt, sodass das Verfahren dem unterschiedlichen Naturell der einzelnen Gemeinden explizit gerecht wird. Die Gruppierung erfolgte für kleine und mittlere Städte und Gemeinden auf Basis der Anzahl der Klärfälle pro Gemeinde, Großstädte werden aufgrund ihrer strukturellen Unterschiede als gesonderte Gruppe(n) kategorisiert. Somit sind die Modelle abgesichert gegen Änderungen von Gemeindegrenzen, die Zugehörigkeit zu einer Gruppe hängt für kleinere und mittlere Gemeinden/Städte lediglich von der Anzahl ihrer Klärfälle ab.

Nach einigen Versuchen mit anderen Vorgehensarten (nach Urbanisierungsgrad, geographischen Gegebenheiten etc.) erwies sich dieser einfache Ansatz als der zielführendste, die tatsächliche Realität abzubilden. Die historische Datenbasis wurde demnach in der Form aufgeteilt, dass sie fünf Gemeindegruppen widerspiegelt und schließlich dermaßen getrennt zur separaten Modellbildung übergeben.

Historische Datenbasis von 2006 und 2011			
Gemeinden mit	**Gruppe**	**Anzahl Gemeinden**	**Anzahl Löschfälle**
0 bis kleiner 15 Klärfälle	1	1.276	4.284
15 bis kleiner 100 Klärfälle	2	883	16.402
100 bis kleiner 10.000 Klärfälle	3	96	11.104
Ballungsräume (Städte über 100.000 EW)	4	4	19.478
10.000 Klärfälle und mehr (Wien)	5	1	41.397
Gesamtergebnis		2.260	92.665

Tabelle 3: Die Einteilung der Gemeinden nach der Anzahl ihrer Klärfälle in der historischen Datenbasis zur Behandlung mit unterschiedlichen Regressionsmodellen

Fälle die bereits in den Jahren 2006 und/oder 2011 geklärt wurden (anerkannt oder nicht anerkannt), werden in den Folgejahren analog behandelt. Zur Bewertung mit

der geschätzten Löschwahrscheinlichkeit verbleiben also nur vollständig neue Klärfälle.

Diese subtile Modifikation bringt neben ihrer sachlichen Notwendigkeit noch andere Vorteile wie eine größere Modellstabilität in Bezug auf die historische Variabilität der gemeindespezifischen Löschquoten.

7 Ermittlung der Schwellwerte zur Nichtanerkennung von Klärfällen

Nachdem nun jedem Klärfall eine Löschwahrscheinlichkeit zugeordnet wurde, stellt sich die Frage, welche Datensätze tatsächlich als Löschfälle zu klassifizieren sind.

Nach der Anwendung der von der historischen Datenbasis abgeleiteten Modelle auf die selbige, werden die Datensätze pro Gemeindegruppe nach den erhaltenen Löschwahrscheinlichkeiten absteigend sortiert, sowie die Anzahl der tatsächlich beobachteten Nichtanerkennungen A ausgezählt. Die Schwelle ab der künftig ein Löschfall angenommen wird, stellt die dem an A-ter Stelle stehenden Datensatz zugewiesene Wahrscheinlichkeit dar.

8 Validierung des Regressionsmodells

Ziel des Regressionsmodells ist es, Löschfälle bzw. „Karteileichen" in einer Masse von Klärfällen zu identifizieren. Neben diesem „Hauptziel " gibt es allerdings für die Modellbewertung nach klassischen statistischen Gesichtspunkten, den statistischen Kennzahlen und Tests, auch ein „Nebenziel", das die Modellauswahl beeinflusst. Es sollen zwar möglichst die richtigen Fälle markiert werden, es ist aber auch wichtig, die richtige Gesamtzahl der Löschfälle je Gemeinde zu finden.

8.1 Verfügbare / verwendete Variable

Wie auch bereits in Punkt 4 ausgeführt, ist bei der Variablenauswahl insbesondere darauf zu achten, dass zum Zeitpunkt der Durchführung der Analyse diese Informationen auch für den Bestand der Klärfälle zur Verfügung stehen. Da ein Klärfall per Definition ausschließlich im Melderegister und sonst in keinem anderen Register vorkommt, ist hier also eine Einschränkung auf die Merkmale des Melderegisters gegeben.

Ausnahme bilden Merkmale, die nicht die Person betreffen, sondern z.B. Gemeinden. Diese können auch Struktur- oder demographische Indikatoren sein. Allerdings sind diese Merkmale für alle Fälle einer Gemeinde gleich.

Umfangreiche Tests mit Faktoren- und Clusteranalyse haben letztendlich zu folgendem Variablensatz als potentiell relevant für die Ermittlung der Nichtanerkennungen geführt:

Variable	Quelle	Erläuterung
ALTER	ZMR	Alter zum Stichtag
ALTER_5	ZMR	Alter in 5-Jahresgruppen
ALTER_GR	ZMR	3 Altersgruppen (Kind, Erwerbsfähig, Pensionist)
ALTER_ZUZUG	ZMR	Alter zum Zeitpunkt des Zuzugs nach Österreich
ANZ_MELDAEND	ZMR	Anzahl Meldeänderungen der letzten 5 Jahre = Übersiedlungen
ANZ_WS	ZMR	Anzahl Wohnsitze (inkl. Nebenwohnsitze)
DEM_GESCHL	ZMR	Geschlecht
GEBDAT	ZMR	Geburtsdatum
GEBSTAAT	ZMR	Geburtsstaat
GEBSTAAT_GR	ZMR	Aggregierte Kategorie Geburtsstaat: Österreich, Deutschland, EU14, EU12, EU-Rest, Kontinente
STAAT_GR	ZMR	Aggregierte Kategorie Staatsbürgerschaft Österreich, Deutschland, EU14, EU12, EU-Rest, Kontinente
FAMST	ZMR	Familienstand
gebstaat_gr2	ZMR	Aggregierte Kategorie Geburtsstaat: Österreich, Deutschland, EU, Rest der Welt
gebstaat_gr3	ZMR	Aggregierte Kategorie Geburtsstaat: Österreich und Deutschland, Rest der Welt
ANT_UEBER60	ZMR (Gemeinde)	Anteil der über 60jährigen in der Gemeinde
ANT_UNTER15	ZMR (Gemeinde)	Anteil der unter 15jährigen in der Gemeinde
AZ_KF_GEM	ZMR (Gemeinde)	Anzahl Klärfälle in der Gemeinde
PARTNER	Beziehungstabelle	Nachweis eines Partners in der Beziehungstabelle (0=nein, unbekannt, 1=ja)
ANZAHL_BEZIEHUNGEN	Beziehungstabelle	Nachweis/Anzahl von weiteren Beziehungen in der Beziehungstabelle
PERS_M2_NTZ	GWR	Anzahl der Personen pro m² bezogen auf Wohnung excl. NWS
PERS_M2_NTZ_NWS	GWR	Anzahl der Personen pro m² bezogen auf Wohnung inkl. NWS
PERS_M2_OBJ	GWR	Kategorie Objekt: Personen pro m² im Gebäu-

		de
ANZ_NTZ_OBJ	GWR	Anzahl NTZ = Wohnungen in einem Objekt = Gebäude

8.2 Untersuchte Modelle

Um das bestmögliche Modell für die Ermittlung der Löschfälle zu finden und auch um eine Abschätzung für die Stabilität des Modells zu bekommen, wurden mehrere statistische Berechnungsmethoden untersucht. Aufgrund der resultierenden Ergebnisse wurde die Entscheidung hinsichtlich der Verwendung der logistischen Regression getroffen, wobei die unterschiedlichen Verfahren zu sehr ähnlichen Ergebnissen geführt haben.

8.2.1 Logistische Regression

In Punkt 5 wurde bereits formal die logistische Regression behandelt.
Für die fünf Gemeinde- bzw. Klärfallgruppen wurden zunächst in einem schrittweisen Verfahren für eine 70%ige Trainingsstichprobe jene Variablen entwickelt, die zur Erklärung beitragen. Das jeweilige Modell wurde anschließend anhand der 30%igen Validierungsstichprobe hinsichtlich der Stabilität überprüft, d.h. der Anteil der Fehlklassifikationen wurde verglichen. Für alle fünf Gruppen konnten stabile Modelle ermittelt werden.

Einer genaueren Betrachtung zur Beurteilung der Modellgüte wurden auch die ROC-Kurven[8] unterzogen. Die Abweichungen bzw. Übereinstimmungen sind alle im akzeptablen Bereich.

Der Einzelfall stellt lediglich einen Aspekt der zu lösenden Aufgabe dar. Neben möglichst hoher Treffergenauigkeit auf Personenebene ist eine möglichst gute Gesamt-Löschzahl je Gemeinde ebenso wichtig.

---

[8] Die ROC-Kurve (***Receiver Operating Characteristic***) stellt den Zusammenhang zwischen 1 – Spezifizität auf der x-Achse und Sensitivität auf der y-Achse in Abhängigkeit vom Trennwert dar.
Die Fläche unter der Kurve kann als Maß für die Güte der Klassifikation angesehen werden.
Faustregel:
0,7 < ROC < 0,8 akzeptable Diskrimination
0,8 < ROC < 0,9 exzellente Diskrimination
ROC > 0,9 außergewöhnliche Diskrimination

Die Abweichungen der tatsächlichen zu den modellierten Löschungen je Gemeinde wurden ebenfalls einer genauen Betrachtung unterzogen.

Gemeindegruppe	in % der Bevölkerung	Anzahl Gemeinden	in % der Gemeinden
0 bis kleiner 15 Klärfälle	>1,0-1,7%	9	0,71
	>0,5-1,0%	29	2,27
	>0-0,5%	821	64,34
	0	417	32,68
15 bis kleiner 100 Klärfälle	>1,0-2,2%	11	1,25
	>0,5-1,0%	41	4,64
	>0-0,5%	741	83,92
	0	90	10,19
100 bis kleiner 10.000 Klärfälle	>1,0%	0	0,00
	0,5-1,0%	3	3,13
	>0-0,5%	90	93,75
	0	3	3,13
Ballungsräume (Städte über 100.000 EW)	bis 0,17%	1	25,00
	unter 0,1%	3	75,00
10.000 Klärfälle und mehr (Wien)	*0,1 %*[9]	1	100,00

Insgesamt sind bei der Hälfte aller Gemeinden Abweichungen von weniger als 0,5 % festzustellen.

Eine weitere Prüfung der Modellgüte erfolgte hinsichtlich des Anteils der Löschungen an der Bevölkerung in der jeweiligen Gemeinde- bzw. Klärfallgruppe , da manche Gruppen aus sehr wenigen Gemeinden bestehen, auch für Gesamt-Österreich. Die Verteilungsanalysen zeigten keine Auffälligkeiten.

8.2.2 Boost

Zur Modellierung der Nichtanerkennungsquote wurde alternativ versucht, ein geeignetes Modell mittels Boosting zu finden.

Boosting (engl. „Verstärken") ist ein Algorithmus der automatischen Klassifizierung, der mehrere schlechte Klassifikatoren zu einem einzigen guten Klassifikator verschmilzt. (Quelle: Wikipedia) Er wurde von Freund und Shapire [4] entwickelt und ist

[9] Schätzung aufgrund der vorliegenden Daten 2012

im R-Paket Ada (Autoren: Mark Culp, Kjell Johnson, and George Michailidis) umgesetzt. Die darin enthaltenen Funktionen ada und adatest wurden auf die Löschdaten angewandt.

Es wurden Modelle mit einer unterschiedlichen Anzahl an Iterationsschritten gerechnet, welche aber keine wesentlichen Verbesserungen brachten. Hier muss immer ein Trade-Off zwischen falsch klassifizierten Löschfällen bzw. Nicht-Löschfällen gemacht werden.

Bei der Ergebnis-Betrachtung der richtig, beziehungsweise falsch vorhergesagten Löschungen erkennt man, dass diese Methode in etwa gleich gut ist wie die eine logistische Regression, wobei sie für die Gemeinden mit vielen Klärfällen etwas schlechtere Resultate lieferte, für jene mit weniger Klärfällen etwas bessere.

8.2.3 Tree = Entscheidungsbaum

Eine weitere Methode, die zur Modellierung der Nichtanerkennungsquote untersucht wurde, ist das Tree-Verfahren.

Entscheidungsbäume sind eine Methode zur automatischen Klassifikation von Datenobjekten und damit zur Lösung von Entscheidungsproblemen. Sie werden außerdem zur übersichtlichen Darstellung von formalen Regeln genutzt. Ein Entscheidungsbaum besteht immer aus einem Wurzelknoten und beliebig vielen inneren Knoten sowie mindestens zwei Blättern. Dabei repräsentiert jeder Knoten eine logische Regel und jedes Blatt eine Antwort auf das Entscheidungsproblem. (Quelle: Wikipedia)

Die Komplexität und Semantik der Regeln sind von vornherein nicht beschränkt. Bei binären Entscheidungsbäumen kann jeder Regelausdruck nur einen von zwei Werten annehmen. [5]

Die Untersuchung der Modelle bei Einsatz von Entscheidungsbäumen liefert ähnliche Ergebnisse wie die Verfahren der logistischen Regression oder auch Boost.

8.2.4 Qualitätssicherung

Die gewählte Methode sowie die genaue methodische und technische Vorgehensweise wurden von Univ.-Prof. Grossmann geprüft.

Jährlich, also bei jeder Anwendung des Modells, erfolgt eine umfangreiche Datenprüfung hinsichtlich der Homogenität der Klärfälle. Ändert sich die Struktur und treten besondere Häufungen auf, muss das Modell angepasst werden. Darüber hinaus ist

die legistische Voraussetzung für eine periodische Überprüfung von Klärfällen analog zu § 5 Registerzählungsgesetz in Vorbereitung.

9 Zusammenfassung

Mit der Registerzählung 2011 steht nun erstmalig ein umfangreicher historischer Datenbestand zur Masse der Klärfälle zur Verfügung, der die Bildung eines statistischen Modells zur Schätzung der individuellen Löschwahrscheinlichkeit ermöglicht. Dabei wurde ein Standardverfahren – nämlich ein lineares, logistisches Regressionsverfahren – angewandt um ein plausibles Modell zu erhalten, welches fortan – zumindest bis zur nächsten Registerzählung bzw. Wohnsitzanalyse - alljährlich im Rahmen der registerbasierten Bevölkerungsfeststellung verwendet werden kann, um die Nichtanerkennungsquoten auf Gemeindeebene zu ermitteln.

Das Regressionsmodell wurde anhand einer Reihe von statistischen Kennzahlen (Nullhypothesentests, Chi-Quadrat–Tests etc.) überprüft und auf Basis von verschiedenen Versuchen der Querüberprüfungen mit alternativen Verfahren (Entscheidungsbäume, andere Regressionsmodelle) validiert.

10 Literaturverzeichnis

[1] Statistik Österreich: Bericht über die Probezählung 2006

[2] Bundesgesetz über die Durchführung von Volks-, Arbeitsstätten-, Gebäude- und Wohnungszählungen (Registerzählungsgesetz), BGBl. I Nr. 33/2006 i.d.g.F.

[3] SAS Institute Inc: SAS/STAT(R) 9.22 User's Guide, Juni 2010

[4] Yoav Freund and Robert E. Schapire (1997); A Decision-Theoretic Generalization of Online Learning and an Application to Boosting, Journal of Computer and System Sciences, 55(1):119-139)

[5] J.R. Quinlan: *Induction of decision trees*, Machine learning, 1(1):81-106, 1986