Die Berechnungsmethode zur Ermittlung der Nichtanerkennungsquote nach dem Zensus 2021

VIKTOR SOMMER

Als ein Ergebnis der für den Stichtag 31.10.2021 durchgeführten Wohnsitzanalyse der Registerzählung wurde das seit dem Finanzjahr 2014 eingesetzte Modell, welches den Anteil an „Karteileichen" aus der Menge der Verdachtsfälle auf Basis eines statistischen Standardverfahrens schätzt, geprüft und adaptiert. Dieses aktualisierte Modell wird ab dem Finanzjahr 2024 für die Berechnung der sogenannten Nichtanerkennungsquote eingesetzt. Das 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 „Karteileichen" erwiesen haben. Mit den zusätzlichen zehn Jahren an Informationen wurde das bisherige Modell weiterentwickelt und mit den neuesten Daten rekalibriert.

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 und 2021 erfolgreich angewendet.

Dabei wird ausgehend vom Zentralen Melderegister (ZMR) versucht, jede Person mit einer gültigen Hauptwohnsitzmeldung durch andere Register zu bestätigen. Diese als sogenannte Lebenszeichen titulierten Einträge in Registern, wie beispielsweise jenem des Dachverbandes 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 Teil – aber eben nicht vollständig – Löschfälle dar, welche den betreffenden Gemeinden als Wohnbevölkerung abzuerkennen sind.[3]

Im Zuge der Probezählung 2006 sowie den Registerzählungen 2011 und 2021 wurde von Statistik Austria an diese Personenmenge per RSb-Brief[4] eine Benachrichtigung versendet, in welcher um eine Bestätigung des fraglichen Hauptwohnsitzes ersucht wurde. Hierdurch konnten die Löschfälle unmittelbar festgestellt werden.[5]

Für die außerhalb einer Registerzählung stattfindenden jährlichen Bevölkerungsfeststellungen für den Finanzausgleich kann die Menge an fraglichen Klärfällen alljährlich ermittelt werden; jedoch ist aus diesen die Menge an Löschfällen direkt – also ohne die Möglichkeit einer Briefbefragung –möglichst exakt abzuleiten. Dies geschieht – in Fortschreibung der Methodik zur Festlegung der Formel zu Nichtanerkennungen von Hauptwohnsitzen nach der Registerzählung 2011 – mittels Schätzung der Wahrscheinlichkeit einer Löschung durch statistische Klassifikation (logistische Regression). Daraus wird dann die gemäß Finanzausgleichsgesetz zu ermittelnde Bevölkerungszahl[6] berechnet. Da mit der Registerzählung 2021 erstmals seit zehn Jahren neue Daten bezüglich der tatsächlichen Zähl- und Löschfälle zur Verfügung stehen, wurde das bisherige Modell überprüft. Im Ergebnis wurde die grundsätzliche Struktur des Modells nicht verändert, lediglich Variablen und Parameter wurden angepasst.

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 des Registerzählungsgesetzes fallen, werden nicht anerkannt. Dies sind z. B. 90-Tage-Fälle, KIT-Fälle[7], 180-Tage-Fälle.

1) Siehe *Statistik Austria* (2009): „Bericht über die Probezählung 2006".
2) Dabei handelt es sich um Bestimmungen des Registerzählungsgesetzes: 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) Registerzählungsgesetz (BGBL I Nr. 33/2006) i. d. g. F., § 7 Abs. 1 in Verbindung mit § 5 Abs. 6.
4) Ein RSb-Brief („Rückscheinbrief") ist ein behördliches Schriftstück, das einer Bestätigung durch eine:n Empfänger:in bedarf.
5) Aufgrund des Registerzählungsgesetzes musste den betroffenen Gemeinden auch die Möglichkeit gegeben werden, für die so ausgelisteten zweifelhaften Hauptwohnsitze noch andere Lebenszeichen zu erbringen (schriftliche Bestätigungen der Bürger:innen). Der Vorgang kann folglich als erschöpfend bezeichnet werden.
6) Ergebnisse für 31.10.2022 siehe Webseite: www.statistik.at > Statistiken > Bevölkerung und Soziales > Bevölkerung > Bevölkerungsstand > Bevölkerung gemäß Finanzausgleichsgesetz.
7) Darunter versteht man Personen, die fälschlicherweise im ZMR mehrfach mit Hauptwohnsitz angelegt wurden.

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.

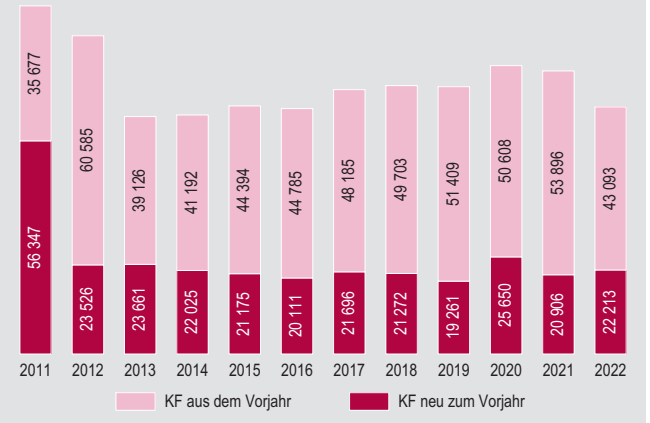
Nichtanerkennungen (NA): Personen aus der Menge der Klärfälle, für die keine Bestätigung des Hauptwohnsitzes vorliegt bzw. die nach Anwendung des statistischen Modells nicht anerkannt werden.

Ausgangslage

Für die Identifikation der Klärfälle steht mit der Lebenszeichenanalyse alljährlich eine Menge an Datensätzen aus dem ZMR mit exakten Adressdaten (und damit verbunden, in aller Regel einer recht genauen Erfassung der Wohnsituation) sowie einer Reihe von demographischen Personenmerkmalen zur Verfügung, zu denen das registerbasierte Verfahren keine weiteren Informationen anreichern konnte. Somit entspricht die Qualität der demographischen Merkmale einzig den im Zentralen Melderegister enthaltenen Informationen.

Allein aus diesem jährlich erstellten 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 zeigt sich eine recht große Menge an unveränderten, alljährlich wiederkehrenden Verdachtsfällen *(Grafik)*. Die Information, seit wann und wie oft eine Person als ein Klärfall gewertet wurde, steht mittlerweile in einer Zeitreihe von zwölf Jahren zur Verfügung und wurde dementsprechend in das neue Modell mit aufgenommen. Dies war aufgrund der kurzen Zeitreihe für das bisherige Modell nicht möglich.

Entwicklung der Klärfälle (KF) sowie des Anteils an im Vergleich zum Vorjahr neuen Meldungen in den Jahren 2011–2022 Grafik



Q: STATISTIK AUSTRIA, Registerzählung.

Ergebnisse eines Population Stability Index (PSI)[8] ergaben eine signifikante Veränderung der Verteilung für einige der verwendeten Inputvariablen des bisherigen Modells. Aufgrund dieser strukturellen Veränderungen wurde das Jahr 2006 aus der Datenbasis für die Neuentwicklung entfernt und nur die Jahre 2011 und 2021 herangezogen. Umfangreiche Tests ergaben, dass die gewählte Datenbasis für alle drei möglichen Validierungsjahre (2006, 2011 und 2021) die stabilsten Ergebnisse hinsichtlich Treffergenauigkeit und Präzision erzielte.

Mit Hilfe des über zwölf Jahre kumulierten Datenbestandes der Klärfälle, welcher etwa 167 000 Einzeldatensätze umfasst, lassen sich Rückschlüsse auf die Struktur der Löschfälle ziehen. Dies geschieht, indem alle im Datensatz enthaltenen Merkmale auf ihre Eigenschaft hin untersucht werden, ob sie Anerkennungen von Nichtanerkennungen zu trennen vermögen.

Logistisches Regressionsmodell

Die Entscheidung für ein Regressionsmodell fußte auf einer Reihe von Gründen. Zum einen eignete sich die vorliegende historische Datenbasis in der Struktur ihrer Ausprägungen gut für einen derartigen Ansatz.

So konnte eine Reihe von Merkmalen ausgemacht werden, welche direkt mit dem Löschverhältnis korrelieren (also beispielsweise mit höherem Wert zu einer niedrigeren Löschquote führen) und auch die Anzahl der kategorialen Merkmale (Geburtsort, Staatsbürgerschaft etc.) ist nicht überbordend groß im Vergleich zu der Zahl an direkten (Alter bei Zuzug etc.) und indirekten numerischen Variablen (durchschnittliche Quadratmeterzahl pro Einwohner:in am Wohnort etc.).

Zum anderen stellen Nachvollziehbarkeit und Transparenz im Rahmen der Registerzählung bei allen Überlegungen stets wichtige Leitgedanken dar: Das Verfahren eines logistischen Modells stellt ein Standardwerkzeug aus dem statistischen Methodeninventar dar und ist einem breiten Publikum bekannt.

Das umgesetzte logistische Modell hat die Form:

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

Wobei α die Schätzkonstante und der Vektor $\boldsymbol{\beta}$ der Form

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

die gesuchten Koeffizienten für die potentiellen Trennvariablen darstellen.

Erhält man nun in den Folgejahren einen Datenbestand mit allen benötigten Trennvariablen, wie dieser im Zuge der registerbasierten Bevölkerungsfeststellung für den Finanzaus-

8) Ein Konzept zur Prüfung ob sich die aktuelle Population gegenüber der Population zum Zeitpunkt der Modellentwicklung verändert hat.

gleich alljährlich vorliegt, so lassen sich die entsprechenden Wahrscheinlichkeiten auf Einzeldatensatzebene zuordnen.

Historische Datenbasis

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 entsprechen. Statt lediglich ein einziges Regressionsmodell für Gesamtösterreich abzuleiten, wurden die Gemeinden entsprechend der Anzahl ihrer Klärfälle in unterschiedliche Gruppen eingeteilt. Wien wurde aufgrund der hohen Klärfallanzahl als gesonderte Gruppe kategorisiert, weitere Landeshauptstädte wie Graz, Linz, Innsbruck und Salzburg bilden die zweite Gruppe. Die restlichen Gemeinden/Städte teilen sich in anhand ihrer Klärfallanzahl in die übrigen drei Gruppen. Somit sind die Modelle abgesichert gegen Änderungen von Gemeindegrenzen, die Zugehörigkeit zu einer Gruppe hängt also nur 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 Löschfälle zu identifizieren. Die Datenbasis wurde demnach in fünf Gemeindegruppen aufgeteilt und zur separaten Modellbildung übergeben *(Tabelle 1)*.

Historische Datenbasis 2011 und 2021 Tabelle 1

Gemeinden mit	Gruppe	Anzahl Gemeinden	Anzahl Klärfälle	Anzahl Löschfälle
10 000 KF und mehr	1	1	72 097	52 956
999 bis 9 999 KF	2	4	26 953	22 091
100 bis 999 KF	3	32	13 344	8 403
15 bis 99 KF	4	435	32 586	19 073
0 bis 14 KF	5	1 590	21 846	12 678
Gesamt		**2 062**	**166 826**	**115 201**

Q: STATISTIK AUSTRIA, Registerzählung. – Anm: Einteilung der Gemeinden zur Behandlung mit unterschiedlichen Regressionsmodellen nach der Anzahl ihrer Klärfälle entsprechend der Datenbasis 2011 und 2021. – KF = Klärfall.

Die zweite Modifikation trägt dem Umstand Rechnung, dass die Ergebnisse der Wohnsitzanalyse 2006, 2011 und 2021 auf Personenebene vorliegen.

Sind diese Personen in einem Folgejahr ebenfalls Klärfälle, dann ist es nicht notwendig, über statistische Wahrscheinlichkeiten zu argumentieren. Stattdessen kann einfach das Ergebnis der Briefbefragung herangezogen werden. Dies wurde technisch so realisiert, dass nach der Modellanwendung die Löschwahrscheinlichkeiten all jener Datensätze auf 1 gesetzt werden, die in der letzten Briefbefragung 2021 als „Karteileichen" identifiziert wurden und neuerlich als fragliche Hauptwohnsitzmeldungen aufscheinen. Zur Bewertung mit der geschätzten Löschwahrscheinlichkeit verbleiben also nur vollständig neue Klärfälle.

Diese Anpassung bringt zudem noch den Vorteil einer größeren Modellstabilität in Bezug auf die historische Variabilität der gemeindespezifischen Löschquoten.

Ermittlung der Schwellenwerte zur Nichtanerkennung von Klärfällen

Nachdem im Durchlauf der Regressionsmodelle jedem Klärfall eine Löschwahrscheinlichkeit zugeordnet wurde, stellt sich die Frage, welche Datensätze tatsächlich als Löschfälle zu klassifizieren sind. Die Datensätze der historischen Datenbasis werden pro Gemeindegruppe nach den erhaltenen Löschwahrscheinlichkeiten absteigend sortiert, sowie die Anzahl der tatsächlich beobachteten Nichtanerkennungen N ausgezählt. Die Schwelle, ab der künftig ein Löschfall angenommen wird, stellt die dem an N-ter Stelle stehenden Datensatz zugewiesene Wahrscheinlichkeit dar und wird für jede der fünf Gruppe festgehalten.

Als Alternative zu den individuellen Schwellenwert der Löschwahrscheinlichkeiten pro Gemeindegruppe wurde ein allgemeiner Wert für alle Modelle untersucht. Bisher lagen die einzelnen Schwellenwerte zwischen 53,9 % und 69,7 % bzw. 59,1 % und 71,5 % für das neue Modell. Die untersuchten Werte waren: der Mittelwert der neu kalkulierten Löschwahrscheinlichkeiten, der Schwellenwert der gesamten absteigend sortierten Klärfälle sowie ein von Expert:innen definierter Wert von 70 %. Alle diese Alternativen führten allerdings dazu, dass in manchen Gemeindegruppen zu viele Fälle anerkannt wurden, dafür in anderen zu wenige. Zudem wurde bei vielen dieser Fälle die Modellentscheidung durch das Ergebnis der Briefbefragung 2021 „überschrieben". Insgesamt ging die Differenz der Ergebnisse für die allgemeinen Löschwahrscheinlichkeiten stark auseinander, sodass aus Stabilitätsgründen individuellen Schwellenwerten der Vorzug gegeben wurde.

An dieser Stelle sei vermerkt, dass ein Anstieg des Schwellenwertes ab dem gelöscht wird, nicht zwangsläufig dazu führt, dass weniger Klärfälle gelöscht werden. In den Gemeindegruppen mit 100 bis 999 bzw. 1 000 bis 9 999 Klärfällen werden im Vergleich zum alten Modell, trotz höherer Löschwahrscheinlichkeit, mehr Personen nicht anerkannt. Insgesamt über alle fünf Gemeindegruppen werden mit dem Modellupdate jedoch weniger Fälle gelöscht als mit dem bisherigen Modell.

Validierung des Regressionsmodells

Ziel des Regressionsmodells ist es, Löschfälle bzw. „Karteileichen" in einer Masse von Klärfällen möglichst gut zu identifizieren. Entscheidend dafür sind die verfügbaren Inputvariablen.

Verfügbare/Verwendete Variable

Wie detailliert in der Dokumentation zur Ermittlung der Nichtanerkennungsquote nach der Registerzählung 2011[9] ausgeführt, ist bei der Variablenauswahl darauf zu achten, dass diese auch für den Bestand der Klärfälle zur Verfügung stehen. Da ein Klärfall per Definition ausschließlich im

9) Siehe *Statistik Austria* (2013): „Berechnungsmethode zur Ermittlung der Nichtanerkennungsquote".

Variablenauswahl der potentiellen Trennvariablen Übersicht

Variable	Erläuterung
ALTER	Alter zum Stichtag
ALTER_5	Alter in Fünfjahresgruppen
ALTER_GR	Altersgruppen (Kind, Erwerbsfähig, Pensionist:in)
ALTER_SQR	Alter zum Quadrat
ALTER_ZUZUG	Alter zum Zeitpunkt des Zuzugs nach Österreich
ANT_KF_GKZ	Anteil der Klärfälle in der Gemeinde
ANT_UEBER60	Anteil der über 60-Jährigen in der Gemeinde
ANT_UNTER15	Anteil der unter 15-Jährigen in der Gemeinde
ANZ_KF_JAHRE	Anzahl der Jahre in denen die Person ein Klärfall war
ANZ_MELDAEND	Anzahl Meldeänderungen der letzten fünf Jahre = Übersiedlungen
ANZ_NTZ_OBJ	Anzahl Nutzungseinheiten = Wohnungen in einem Objekt = Gebäude
ANZ_WS	Anzahl Wohnsitze (inkl. Nebenwohnsitze)
ANZAHL_BEZIEHUNGEN	Nachweis/Anzahl von weiteren Beziehungen in der Beziehungstabelle[1]
AZ_KF_GEM	Anzahl Klärfälle in der Gemeinde
BDL	Bundesland
DEM_GESCHL	Geschlecht
FAMST	Familienstand
GEBDAT	Geburtsdatum
GEBDAT_YEAR	Geburtsjahr
GEBSTAAT	Geburtsstaat
GEBSTAAT_GR	Aggregierte Kategorie Geburtsstaat: Österreich, Deutschland, EU-13 (vor 2004), EU-13 (ab 2004), EU-Rest, Kontinente
GEBSTAAT_GR2	Aggregierte Kategorie Geburtsstaat: Österreich, Deutschland, EU, Rest der Welt
GEBSTAAT_GR3	Aggregierte Kategorie Geburtsstaat: Österreich und Deutschland, Rest der Welt
KF_SEIT_JAHR	Jahr seitdem die Person ununterbrochen ein Klärfall ist
MAX_KF_JAHR	letztes Vorkommen als Klärfall
MIN_KF_JAHR	erstes Vorkommen als Klärfall
NUTS3	EU Klassifizierung – kleinere Regionen, teils schon Großstädte
PARTNER	Nachweis eines:einer Partner:in in der Beziehungstabelle[1]
PERS_M2_NTZ	Anzahl der Personen pro m² bezogen auf Wohnung excl. Nebenwohnsitze
PERS_M2_NTZ_NWS	Anzahl der Personen pro m² bezogen auf Wohnung inkl. Nebenwohnsitze
PERS_M2_OBJ	Kategorie Objekt: Personen pro m² im Gebäude
STAATB	Staatsbürgerschaft
STAAT_GR	Aggregierte Kategorie Staatsbürgerschaft Österreich, Deutschland, EU-13 (vor 2004), EU-13 (ab 2004), EU-Rest, Kontinente

Q: STATISTIK AUSTRIA, Registerzählung.
1) Die Beziehungstabelle enthält alle (seit 2006) registrierten Familienbeziehungen zwischen je zwei Personen, wie z. B. Ehepartner:in.

Melderegister und sonst in keinem anderen Register vorkommt, ist hier also für den jeweiligen Stichtag eine Einschränkung auf die Merkmale des Melderegisters gegeben. Außerdem gibt es Merkmale, die nicht die Person betreffen, sondern z. B. auf Ebene der Gemeinden aggregiert sind. Diese können auch aus anderen Quellen stammen und sind für alle Fälle einer Gemeinde gleich, wie z. B. der Anteil der über 60-Jährigen in der Gemeinde

Zusätzlich zu den Variablen, die bei der ursprünglichen Modellentwicklung getestet wurden, kamen nun auch Merkmale bezüglich der Klärfallvergangenheit der letzten Jahre hinzu. Diese wurden für die Variablenauswahl der einzelnen Modelle auf ihre Signifikanz untersucht.

Nach Evaluierung und Auswahl der unterschiedlichen Regressionsmodelle je Gemeindegruppe blieben im Vergleich zu den bisherigen fünf Modellen 30 Variablen gleich, neun Variablen ähnlich im Sinne ihrer Bedeutung (z. B. „Alter“ und „Alter in 5-Jahresgruppen“), fünf fielen weg, und insgesamt 20 Variablen kamen neu hinzu. Darunter war pro Gemeindegruppe mindestens eine Variable zur Klärfallhistorie.

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. Zusätzlich zu den Modellen, die bei der initialen Entwicklung des Modells analysiert wurden (Boosting und Entscheidungsbäume), wurden diesmal als Alternative zur logistischen Regression ein Gradient-Boosting-Verfahren, sowie unterschiedliche Gruppierungsmethoden der Gemeinden getestet.

Da die unterschiedlichen Verfahren zu sehr ähnlichen, aber weniger akkuraten Ergebnissen führten, wurde entschieden, weiterhin eine logistische Regression zu verwenden.

Logistische Regression

In der Dokumentation zur Ermittlung der Nichtanerkennungsquote nach der Registerzählung 2011[10] sowie im obigen Abschnitt „Logistisches Regressionsmodell“ wurde bereits das Verfahren einer logistischen Regression behandelt.

10) Siehe *Statistik Austria* (2013): „Berechnungsmethode zur Ermittlung der Nichtanerkennungsquote“.

Für die fünf Gemeinde- bzw. Klärfallgruppen wurden wieder in einem schrittweisen Verfahren für eine 70 %-ige Trainingsstichprobe jene Variablen identifiziert, die zur Erklärung beitragen. Das jeweilige Modell wurde anschließend anhand der 30 %-igen Validierungsstichprobe geprüft. Für alle fünf Gruppen konnten akzeptable Modelle ermittelt werden. Die niedrigste Treffergenauigkeit mit 67,7 % findet sich in Gemeinden mit 0 bis 15 Klärfällen, was auf die vielen unterschiedlichen Gemeinden mit relativ wenigen Klärfällen zurückzuführen ist. Die höchste Treffergenauigkeit hat dagegen mit 84,8 % die Gemeindegruppe mit 1 000 bis 9 999 Klärfällen. Insgesamt konnten etwa drei Viertel der Trainingsstichprobe richtig identifiziert werden *(Tabelle 2)*.

Treffergenauigkeit der Regression der Trainingsstichprobe nach Gemeindegruppen Tabelle 2

Gemeindegruppe	Treffergenauigkeit in Prozent
0 bis weniger als 15 Klärfälle	*67,7*
15 bis weniger als 100 Klärfälle	*69,5*
100 bis weniger als 999 Klärfälle	*74,3*
1 000 bis weniger als 9 999 Klärfälle	*84,8*
10 000 Klärfälle und mehr (Wien)	*75,1*
Insgesamt	***74,3***

Q: STATISTIK AUSTRIA, Registerzählung.

Zur Beurteilung der Modellgüte wurden auch die ROC-Kurven (Receiver Operating Characteristic)[11] einer genaueren Betrachtung unterzogen. Die Abweichungen bzw. Übereinstimmungen (Prozent konkordant) sind alle im akzeptablen Bereich *(Tabelle 3)*.

Anteil der konkordanten Paare nach Gemeindegruppen Tabelle 3

Gemeindegruppe	in Prozent
0 bis weniger als 15 Klärfälle	*69,9*
15 bis weniger als 100 Klärfälle	*71,2*
100 bis weniger als 999 Klärfälle	*74,3*
1 000 bis weniger als 9 999 Klärfälle	*77,0*
10 000 Klärfälle und mehr (Wien)	*71,6*

Q: STATISTIK AUSTRIA, Registerzählung.

Gruppierung der Gemeinden

Zusätzlich zu der erweiterten Variablenauswahl wurden alternative Einteilungen in Gemeinde- bzw. Klärfallgruppen betrachtet, ohne allerdings etwas an der Zahl der Gruppen zu verändern.

Zunächst wurde getestet, die Gemeinden mittels k-Means-Algorithmus zu gruppieren. Dabei wird aus einer Menge von ähnlichen Objekten eine vorher bekannte Anzahl von k Gruppen gebildet. Der Algorithmus ist eine der am häufigsten verwendeten Techniken zur Gruppierung von Objekten, da er schnell die Zentren der Cluster findet. Dabei bevorzugt der Algorithmus Gruppen mit geringer Varianz und ähnlicher Größe. Die entstandenen Cluster lieferten insgesamt keine Verbesserung der Trefferquote im Vergleich zu der aktuell verwendeten Einteilung.

11) Siehe *Facett* (2003): „ROC Graphs".

Als nächstes wurden die Gemeinden anhand ihres Klärfallanteils in fünf Gruppen unterteilt, wobei Wien aufgrund der hohen Anzahl an Klärfällen weiterhin eine eigene Gruppe bildete. Die restlichen Gemeinden wurden wie folgt, je nach KF-Anteil eingeteilt:

- weniger als 0,01 %
- 0,025 % bis 0,01 %
- 0,07 % bis 0,025 %
- mehr als 0,07 %
- Wien

Die meisten Gemeinden fielen in die Kategorie mit weniger als 0,01 % an Klärfällen, die meisten Klärfälle nach Wien (27 424) landeten in Gruppe 3. Die Trefferquote dieser Variante war allerdings geringer als für das oben erwähnte k-Means-Clustering und wurde deshalb wieder verworfen.

Schließlich wurde untersucht, ob die Grenzen der bisherigen Klärfallgruppierung etwas angehoben werden sollten. Da die Performance der Modelle für die Gemeindegruppen mit mehr Klärfällen bisher recht gut war, sollte dadurch mehr Information zur Entwicklung in den Gruppen mit relativ wenigen Klärfällen zur Verfügung stehen. Die oberste Schranke anzuheben hätte keinen Effekt, da ohnehin nur Wien die geforderten Klärfälle vorweist. Folgende neue Klärfallgruppen wurden gebildet:

- weniger als 25 Klärfälle
- 25 bis 149 Klärfälle
- 150 bis 1 999 Klärfälle
- 2 000 bis 9 999 Klärfälle
- mehr als 10 000 Klärfälle

Wie erwartet landeten mehr Gemeinden in den niedrigeren Klärfallgruppen. Die Trefferquote fiel hier ein wenig besser aus im Vergleich zu dem Ansatz mit Klärfallanteilen. Jedoch lieferte die ursprüngliche oder die k-Means-Gruppierung weiterhin die besseren Ergebnisse, so dass die ursprüngliche Einteilung nach der Klärfallanzahl letztlich angewendet wurde.

Gradient-Boosting-Verfahren

Zur Modellierung der Nichtanerkennungsquote wurde alternativ versucht, ein geeignetes Modell mittels Gradient-Boosting zu finden. Das Gradientenverfahren wird in der Numerik eingesetzt, um allgemeine Optimierungsprobleme zu lösen. Es basiert auf dem Konzept des sequentiellen Zusammenfügens von Modellen, um die Vorhersagegenauigkeit zu erhöhen. Umgesetzt wurde dieses Modell mittels Extreme Gradient Boosting (XGBoost), einer Open Source-Softwarebibliothek für die Programmiersprache R.

Mit dem XGBoost-Modell wird zwar oft eine höhere Genauigkeit als mit einem einzelnen Entscheidungsbaum erreicht, aber die eigentliche Interpretierbarkeit von Entscheidungsbäumen geht dabei verloren. So ist es beispielsweise

trivial, den Weg eines Entscheidungsbaums zu verfolgen, aber es ist viel schwieriger, die Wege von hunderten oder tausenden von Bäumen zu verfolgen.

Zu den wichtigsten Merkmalen von XGBoost, die es von anderen Gradient-Boosting-Algorithmen unterscheiden, gehören:

- clevere „Bestrafung" von Bäumen
- proportionale Schrumpfung der Blattknoten
- Newton-Boosting
- zusätzlicher Randomisierungsparameter
- Implementierung auf einzelnen, verteilten Systemen und Out-of-Core-Berechnungen
- automatische Auswahl von Merkmalen

Bei der Ergebnisbetrachtung der richtig beziehungsweise falsch vorhergesagten Löschungen erkennt man, dass diese Methode etwa gleich gut ist wie die einer logistischen Regression. Für die Gemeinden mit vielen Klärfällen lieferte sie etwas schlechtere Resultate, für jene mit wenigen Klärfällen geringfügig bessere. Aufgrund der insgesamt ähnlichen Ergebnisse und der einfacheren Interpretierbarkeit wurde einer logistischen Regression der Vorzug gegeben.

Fazit

Wie bereits oben angeführt wurde das bisherige Modell überprüft und mit neuen Daten aus der Wohnsitzanalyse 2021 erweitert. Die österreichischen Gemeinden werden anhand der Anzahl ihrer Klärfälle in fünf Gruppen eingeteilt. Nach ausführlichen Tests zeigt eine logistische Regression für jede dieser fünf Gemeindegruppen die besten Ergebnisse, um die Löschfälle zu bestimmen

Literatur

Fawcett, T. (Palo Alto 2003): „ROC Graphs: Notes and Practical Considerations for Data Mining Researchers", Researchers", Website HP.

Statistik Austria (Wien 2009): „Bericht über die Probezählung 2006".

Statistik Austria (Wien 2013): „Berechnungsmethode zur Ermittlung der Nichtanerkennungsquote nach der Registerzählung 2011", Website Statistik Austria.

Rechtliche Grundlage

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

Summary

As a result of the so-called residence analysis of the 2021 register census, the previous model for estimating the proportion of nominal members in the registered population from the mass of suspected cases on the basis of a standard statistical procedure, was reviewed and adapted.

This updated model will be used, by means of a logistic regression procedure, to calculate the so-called non-recognition rate for the resident population at municipal level starting with the financial year 2024. The model assigns high non-recognition probabilities to all those suspected cases that are very similar in their demographic characteristics to those data records that turned out to be only nominal registrations in past residence analyses. With the input of ten additional years of information, the previous model was further developed and recalibrated with the latest data. Methodologic alternatives to the previous model (gradient boosting, alternative clustering of municipality) were compared, but no significant improvement could be found.