Gutachten zur Bestimmung der Nichtanerkennungsquote im Rahmen der Registerzählung

Univ. Prof. Dr. Wilfried Grossmann, Universität Wien

Bei der Registerzählung tritt das Problem auf, dass bei bestimmten Personen in den Registern keine gesicherte Evidenz über ihren tatsächlichen Aufenthalt in Österreich ist. Dies betrifft insbesondere Personen von denen es neben dem Melderegister keinen Nachweis in anderen Registern gibt. Da eine direkte Nachforschung aus Zeit- und Kostengründen bei der jährlichen Fortschreibung des Bevölkerungsstandes nicht möglich ist, muss für diese Personen mittels statistischer Methoden ermittelt werden, ob sie zu der Wohnbevölkerung zu zählen sind oder nicht. Diese Berechnung muss für die einzelnen Gemeinden gute Schätzungen über die Nichtanerkennungsquote liefern.

Der Bereich Registerzählung hat für die Berechnung der Nichtanerkennungsquote ein neues Modell entwickelt, das statistischen Klassifikationsmethoden verwendet. Gemäß den im Anbot genannten Fragestellungen möchte ich das Verfahren hinsichtlich der Daten und deren Qualität, der Datenaufbereitung, der Berechnungsmethode, der Qualität der Schätzungen und der Berücksichtigung von Vertraulichkeit beurteilen.

Daten und Datenqualität

Die Berechnung der Nichtanerkennungsquote beruht auf den Daten der Registerzählungen 2006 und 2011. Für diese Daten wurden im Rahmen der Registerzählung selbst umfangreiche Qualitätsüberprüfungen durchgeführt, daher kann nicht an deren Qualität hinsichtlich ihrer Vollständigkeit und Korrektheit der Merkmalsausprägungen gezweifelt werden.

Wesentlich für die Berechnung ist die Unterteilung der Daten in Zählfälle, Klärfälle und Technische Nichtanerkennungen. Diese Unterteilung wurde entsprechend den formalen Kriterien (Definitionen) ordnungsgemäß durchgeführt. Innerhalb der Klärfälle wurden durch Befragung und Rücksprache mit lokalen Behörden die zu löschenden Fälle und die zu anerkennenden Fälle vollständig bestimmt. Es liegt also für diese Daten eine vollständige Einteilung in Nichtanerkennungsfälle (Löschfälle) und Anerkennungsfälle vor.

Der Datenkörper, der zur Entwicklung und Validierung des Modells verwendet wird, ist also nach den Kriterien der Datenqualität vollständig und korrekt.

Da die gleichen Methoden auch für die Erstellung der Daten angewendet wurden, für die dann mit dem Modell eine Schätzung der Nichtanerkennungsquote durchgeführt werden soll, ist auch von einer hohen Datenqualität bei diesen Fällen auszugehen.

Methoden der Datenvorverarbeitung für die Berechnungsmodelle

Die Vorverarbeitung der Daten für die Modellerstellung umfasst einerseits die Auswahl der in der Modellierung verwendeten Merkmale, andererseits die Stratifizierung der Daten. In beiden Fällen wurden verschiedene Varianten getestet, so dass der gewählte Prozess der Vorverarbeitung eigentlich das Ergebnis eines umfangreichen Analyseprozesses ist. Diese Vorgangsweise entspricht der üblichen Praxis in der angewandten Statistik.

Die Datenvorverarbeitung für die Modellentwicklung umfasste die folgenden Punkte:

1. Erstellung eines Merkmalssatzes
   Da im Modell nur Merkmale verwendet werden sollen, welche vollständig vorhanden sind, kommen nur jene Merkmale in Frage, welche für alle Personen vorhanden sind. Neben dem ZMR sind dies noch das GWR und die Beziehungstabelle. Die Vorselektion der Merkmale und die Berechnung von abgeleiteten Merkmalen (z.B. Altersgruppen) erfolgten mittels traditionellen Methoden der deskriptiven Statistik und komplexeren Methoden zur Datenanalyse (Hauptkomponenten und Clusteranalyse) und scheinen auch vom inhaltlichen Standpunkt plausibel.
2. Stratifizierung der Daten
   Es erscheint vernünftig das Schätzmodell für unterschiedliche Gemeindegrößen getrennt zu entwickeln. Daher scheint auf den ersten Blick ein derartiges Kriterium als sinnvoll. Dass anstelle dieses Kriteriums die Anzahl der Klärfälle verwendet werden, ist Ergebnis der Voranalysen. Da die Anzahl der Klärfälle aber mit der Gemeindegröße korreliert, kann man diese Stratifizierung als eine hinsichtlich des Auftretens von Klärfällen adjustierte Gemeindegröße interpretieren. Dies erscheint mir sachlich sehr vernünftig zu sein.

Statistische Modelle

Zur Modellierung der Nichtanaerkennungsquote wurde ein Klassifikationsansatz gewählt. Die Lernstichprobe besteht aus den Klärfällen der Daten der Registerzählungen 2006 und 2011, die einwandfrei hinsichtlich der Nichtanerkennung und Anerkennung zugeordnet werden können. Diese werden entsprechend des oben beschriebenen Kriteriums stratifiziert und für jedes Stratum werden Klassifikationsverfahren angewendet. Dabei werden entsprechend der allgemeinen Methodologie in der Klassifikation die Daten in eine Trainingsstichprobe und eine Teststichprobe unterteilt. Das Verhältnis 70% Trainingsdaten und 30% Testdaten ist die heute übliche Aufteilung bei Daten dieser Größenordnung.

Zur Klassifikation selbst wurden unterschiedliche Verfahren angewendet, die für die gegebenen Eigenschaften der verfügbaren Variablen heute in der Literatur allgemein empfohlen werden: Logistische Regression, Baumverfahren und Kombinationsverfahren (Boosting). Alle Verfahren brachten ähnliche Ergebnisse und es scheint mir vernünftig, dass man sich als endgültige Methode für die logistische Regression entschieden hat. Dies vor

allem deshalb, weil die Ergebnisse vom inhaltlichen Standpunkt am einfachsten zu interpretieren sind.

Die Modelle liefern für jeden Klärfall (Personen) eine Wahrscheinlichkeit dafür, dass diese Person ein Löschfall ist. Die Entscheidung erfolgt durch Wahl eines Tresholds für die Wahrscheinlichkeit. Dieser wurde empirisch aus den Daten festgelegt und scheint mir nach Analysen der ROC-Kurven (siehe nächster Punkt: Qualität der Schätzungen) vernünftig.

In der Produktion werden dann diese Modelle für die einzelnen Gemeindeschichten nur für jene Klärfälle angewendet, die bisher noch nicht zugeordnet werden konnten. Diese Vorgangsweise ist sinnvoll, da man davon ausgehen kann, dass eine Person, für die bereits einmal eine positive Klärung erfolgt ist auch weiterhin zur Bevölkerung gezählt werden soll, solange bis eine technische Löschung erfolgt.

Das Modell ist also aus meiner Sicht eine statistisch gut begründbare Vorgangsweise, die für Fragestellungen ähnlicher Art in der angewandten Statistik heute üblich ist. Ich möchte aber trotzdem noch kurz auf zwei mögliche Alternativen eingehen.

1. Zunächst einmal könnte man einwenden, dass es auch möglich wäre ein Modell auf Gemeindeebene zu entwickeln. Dies wäre zwar möglich, hätte aber meiner Meinung nach eine Reihe von Nachteilen. Zunächst einmal ist die Gemeindeeinteilung einem Wandel unterworfen. In den letzten Jahren hat es eine Reihe von Zusammenlegungen gegeben (Verwaltungsreform) und dieser Prozess ist sicher noch nicht abgeschlossen. Ein weiterer Grund der gegen das Gemeindemodell spricht, sind die zur Verfügung stehenden Variablen. Es müssten erst für Gemeinden charakteristische Merkmale berechnet werden, zum Beispiel Anteile in den einzelnen Staatsbürgerschaften, oder in den Zuzugsaltersgruppen. Für kleine Gemeinden können solche Schätzungen sehr instabil werden. Der wesentlichste Grund scheint mir aber das grundsätzlich andere Problemverständnis: Die Wahl eines Landes als Lebensmittelpunkt scheint mir eine persönliche Entscheidung, eventuell auch eines Haushaltes, zu sein. Daher sollten primär persönliche Variablen zur Charakterisierung herangezogen werden und nicht Gemeindevariablen.
2. Es wäre auch denkbar ein Übergangsmodell zu definieren, das für jede Person den Wechsel zwischen den Zuständen Zählfall, Technische Nichtanerkennung, Anerkennung und Löschfall beschreibt. Von diesen Zuständen sind aber die zwei Zustände Anerkennung und Nichtanerkennung nicht beobachtbar, nur der Zustand Klärfall, der die beiden subsumiert. Es wäre also eine keine reine Markovkette zu modellieren, sondern eine sogenannte Hidden Markov Chain. Derartige Modelle sind weit komplexer und die Einbeziehung von personenspezifischen Variablen wäre sehr aufwändig. Es wurde daher eine derartige Modellierung nach intensiver Diskussion verworfen.

Qualität der Schätzungen nach statistischen Kriterien

Bei prädiktiven Modellen sind neben den üblichen statistischen Kriterien wie Tests und Konfidenzintervallen für die einzelnen Parameter sowie Tests über die Anpassungsgüte des Modells vor allem Aussagen über die Vorhersagegüte von Interesse. Im Bereich der Qualität der Schätzungen wurden die üblichen Verfahren angewendet und die Ergebnisse geprüft. Wesentlicher scheint mir hier die Vorhersagegüte zu sein. Hier wurden die beiden wesentlichen Kriterien Fehlklassifikationsmatrix (Confusion matrix) und ROC-Kurve berechnet. Die ROC-Kurve erlaubt eine gesamtheitliche Beurteilung des Verfahrens und aus ihr lässt sich auch ein Güteindex berechnen. Mit einem Wert um 0.7 liegt er doch deutlich über 0.5, was einer naiven Klassifikation entsprechen würde. Mehr wird wohl aufgrund des doch eher geringen verfügbaren Merkmalssatzes für die Personen nicht möglich sein. Die Vorhersagequalität wurde auch noch auf Gemeindeebene überprüft und die Ergebnisse der relativen Fehler scheinen mir akzeptabel zu sein.

Neben den formalen Kriterien spielen für die Qualität auch noch die Voraussetzungen für die Anwendbarkeit des Verfahrens eine zentrale Rolle. Hierzu ist zu bemerken, dass alle gewählten Klassifikationsverfahren praktisch keine Voraussetzungen hinsichtlich der Verteilung der Variablen erfordern. Für die Abhängigkeit der Parameterschätzungen von einzelnen Datenpunkten wurden die üblichen Diagnosewerkzeuge angewendet und es ergaben sich keine besonders auffälligen Ergebnisse.

Ein Punkt bei der Anwendung der Modelle soll aber noch kurz erwähnt werden: Der Ansatz setzt implizit voraus, dass sich die Abhängigkeit der Klassifikationswahrscheinlichkeit von den persönlichen Merkmalen im Laufe der Zeit nicht ändert. Diese Annahme scheint auf Grund der Analyse der Datensätze zu zwei Zeitperioden (2006 und 2011) gerechtfertigt zu sein. Es sollte bei der Anwendung aber doch immer überprüft werden, ob diese Annahme weiterhin gültig ist.

Einhaltung der Richtlinien der Vertraulichkeit

Alle persönlichen Daten werden anonymisiert nur innerhalb von Statistik Austria verwendet und es werden auch keine Einzelergebnisse veröffentlicht. Damit sind aus meiner Sicht die Richtlinien der Vertraulichkeit erfüllt.

Zusammenfassend möchte ich feststellen, dass die gewählte Vorgehensweise für das Problem sehr gut geeignet ist und die Umsetzung methodisch einwandfrei durchgeführt wurde.

Wien, 13.9. 2013