Feuerhake, Jörg; Lange, Kerstin; Siegismund, Annelen; Vigneau, Elsa

Article

Kodierung des Geburtsstaats in der Wanderungsstatistik: Ein Vergleich regelbasierter Signierung mit Verfahren des maschinellen Lernens

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Feuerhake, Jörg; Lange, Kerstin; Siegismund, Annelen; Vigneau, Elsa (2020) : Kodierung des Geburtsstaats in der Wanderungsstatistik: Ein Vergleich regelbasierter Signierung mit Verfahren des maschinellen Lernens, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 72, Iss. 3, pp. 98-110

This Version is available at:
https://hdl.handle.net/10419/220347

Jörg Feuerhake

ist Diplom-Volkswirt und Referent im Referat „Maschinelles Lernen und Imputationsverfahren“ des Statistischen Bundesamtes. Er befasst sich unter anderem mit der methodischen Weiterentwicklung und dem Einsatz maschineller Lernverfahren.

Kerstin Lange

hat Statistik studiert und ist seit 2017 im Statistischen Bundesamt im Bereich mathematisch-statistische Methoden tätig. Im Referat „Maschinelles Lernen und Imputationsverfahren“ ist sie schwerpunktmäßig für Imputationsverfahren zuständig.

Annelen Siegismund

hat Volkswirtschaftslehre in Mainz studiert und ist Referentin im Referat „Räumliche Bevölkerungsbewegungen, Gebietsgliederungen“ des Statistischen Bundesamtes. Sie beschäftigt sich insbesondere mit Datenqualitätsaspekten der Wanderungsstatistik.

Elsa Vigneau

hat Politikwissenschaft studiert und ist seit 2018 Wissenschaftliche Mitarbeiterin im Referat „Räumliche Bevölkerungsbewegungen, Gebietsgliederungen“ des Statistischen Bundesamtes. Zu ihren Aufgabenschwerpunkten gehören methodische Fragen der Wanderungsstatistik und europäische Datenlieferungen.

KODIERUNG DES GEBURTSSTAATS IN DER WANDERUNGSSTATISTIK

Ein Vergleich regelbasierter Signierung mit Verfahren des maschinellen Lernens

Jörg Feuerhake, Kerstin Lange, Annelen Siegismund, Elsa Vigneau

↘ **Schlüsselwörter:** Geburtsstaat – Signierung – Wanderungsstatistik – Random Forest – Maschinelles Lernen

ZUSAMMENFASSUNG

Seit dem Berichtsjahr 2008 enthalten die Datenlieferungen zur deutschen amtlichen Wanderungsstatistik Angaben zum Geburtsstaat von zu-, fort- und umziehenden Personen. Wegen unzureichender Qualität wurden diese Daten bislang nur für Schätzungen nach Geburtsstaatengruppen im Rahmen europäischer Lieferverpflichtungen genutzt. Künftig sollen auch Aussagen über die einzelnen Geburtsstaaten der Wandernden möglich sein. Daher wurden im Jahr 2019 verschiedene Methoden untersucht, um ein Verfahren zur automatisierten Plausibilisierung und Signierung des Merkmals zu entwickeln. Der Beitrag stellt zwei Ansätze vor und vergleicht sie miteinander: eine regelbasierte Geburtsstaatssignierung basierend auf Leitdateien und den Einsatz von maschinellen Lernverfahren.

↘ ***Keywords:*** *country of birth – classification – migration statistics – random forest – machine learning*

ABSTRACT

Since reference year 2008, the data provided for German official migration statistics have included information on the country of birth of immigrants, emigrants, or persons moving their main residence inside Germany. Due to insufficient quality, these data have only been used for estimations in a breakdown by country-of-birth groups in order to meet European data transmission obligations. To enable future analysis of the migrants' country of birth, various methodological investigations were carried out in 2019 to develop a method for automatically checking the consistency of and classifying the relevant variable. This article presents two approaches and compares them with each other: on the one hand, a rule-based country-of-birth classification using reference files and, on the other, the use of machine learning methods.

1 Einleitung

Die amtliche Wanderungsstatistik weist die Anzahl der Zu- und Fortzüge über Gemeindegrenzen nach verschiedenen demografischen Merkmalen nach. Das Merkmal „Geburtsstaat“ ist seit 2008 Erhebungsmerkmal der Wanderungsstatistik nach dem Bevölkerungsstatistikgesetz[1]. Allerdings wird diese Angabe bisher aufgrund unzureichender Datenqualität lediglich für Schätzungen nach Staatengruppen für internationale Lieferverpflichtungen genutzt (Mundil/Grobecker, 2011; Carow und andere, 2019). Ein wesentliches Problem für die Datenqualität besteht darin, dass in mehr als 50 % der Lieferdaten aus dem Meldewesen das Merkmal „Geburtsstaat“ nicht befüllt ist. Zur hohen Zahl an fehlenden Angaben trägt vor allem bei, dass gemäß den melderechtlichen Vorgaben der Geburtsstaat nur im Falle einer Geburt im Ausland anzugeben ist. Somit wird das Feld für in Deutschland geborene Personen in der Regel leer gelassen. Hierdurch entfällt die Möglichkeit zu unterscheiden, ob der Geburtsstaat „Deutschland“ ist oder der Staat nicht bekannt ist beziehungsweise die Eintragung fälschlicherweise nicht erfolgte. Die Angabe zum Geburtsstaat kann somit nicht ungeprüft übernommen werden, sondern sie muss plausibilisiert und gegebenenfalls neu kodiert werden. Hierfür bietet sich insbesondere das Merkmal „Geburtsort“ an, das dabei helfen kann, von der Angabe des Ortes auf den Staat zu schließen (zum Beispiel Damaskus → Syrien).

Im Jahr 2019 wurden Methoden zur automatisierten Plausibilisierung und Signierung des Merkmals „Geburtsstaat“ untersucht. Daran beteiligt waren die Arbeitsbereiche im Statistischen Bundesamt, die die Wanderungsstatistik sowie das Maschinelle Lernen und die Imputationsverfahren verantworten. Ziel war, fehlende Angaben auf Mikrodatenebene einzusetzen und eine Veröffentlichung der Wanderungen nach dem Geburtsstaat zu ermöglichen, die den hohen Qualitätsanspruch der amtlichen Statistik erfüllt.

1 Gesetz über die Statistik der Bevölkerungsbewegung und die Fortschreibung des Bevölkerungsstandes (Bevölkerungsstatistikgesetz – BevStatG) vom 20. April 2013 (BGBl. I Seite 826), das zuletzt durch Artikel 9 des Gesetzes vom 18. Dezember 2018 (BGBl. I Seite 2639) geändert worden ist.

Der vorliegende Aufsatz stellt die wesentlichen Ergebnisse dieser Untersuchungen vor. Er zieht zwei ausgewählte Methoden zur Kodierung des Geburtsstaats heran: eine regelbasierte Signierung sowie die Anwendung von maschinellen Lernverfahren. Kapitel 2 gibt einen Überblick über die Rahmenbedingungen für die Signierung. Auf die regelbasierte Signierung des Geburtsstaats geht Kapitel 3 ein, Kapitel 4 behandelt die Signierung mit Verfahren des maschinellen Lernens. Kapitel 5 vergleicht die Ergebnisse beider Methoden miteinander, bewertet die Verfahren im Hinblick auf ihre Einsetzbarkeit und analysiert die Unterschiede in den Ergebnissen. Die Ergebnisse des Vergleichs fasst Kapitel 6 zusammen und gibt einen Ausblick auf das weitere Vorgehen zur Implementierung der Kodierung.

2 Rahmenbedingungen für die Signierung

2.1 Überblick zur Wanderungsstatistik

Die Wanderungsstatistik umfasst alle durch die Meldebehörden registrierten Zu- und Fortzüge mit Verlegung der Haupt- oder alleinigen Wohnung über die Gemeindegrenze. Für das Berichtsjahr 2018 wurden etwa 6,7 Millionen Wanderungsfälle ermittelt (rund 1,6 Millionen Zuzüge aus dem Ausland, 1,2 Millionen Fortzüge in das Ausland und 3,9 Millionen Binnenumzüge). Dies stellt die Größenordnung der jährlich zu signierenden Wanderungsfälle dar.

Der Merkmalsumfang der Wanderungsstatistik begrenzt grundsätzlich die Variablen, die bei der Kodierung berücksichtigt werden können. Darüber hinaus ist allerdings in weiteren Bewegungsstatistiken zu Geburten und Sterbefällen ebenfalls der Einsatz einer Kodierungsfunktion für den Geburtsstaat vorgesehen. Daher wird eine möglichst universelle Einsetzbarkeit der Kodierungsfunktion angestrebt. Die Erhebungsmerkmale der Wanderungsstatistik nach dem Bevölkerungsstatistikgesetz zeigt ↘ Übersicht 1 auf Seite 100.

Übersicht 1

Merkmalsumfang der Wanderungsstatistik nach §4 Bevölkerungsstatistikgesetz

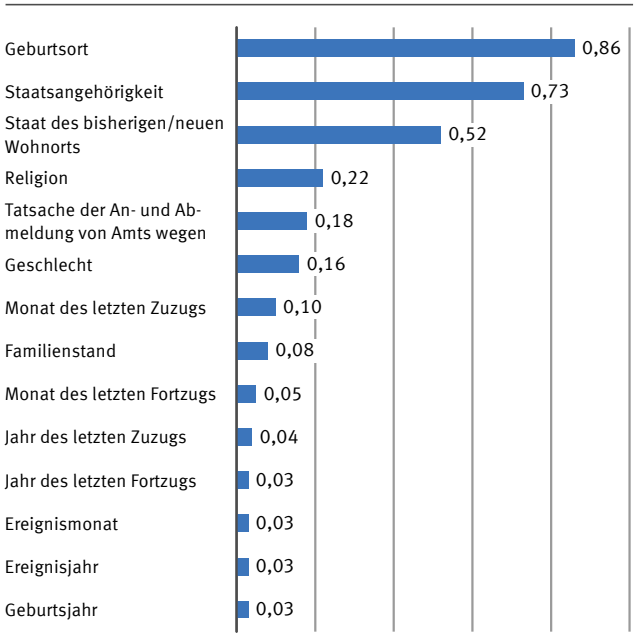
Merkmal	Skalierung
Ereignisdatum (Tag des Einzugs, Auszugs oder Wechsels des Wohnungsstatus)	Datumswert
Bisheriger Wohnort	› Staat (einschließlich unbekannt), 3-stelliger Schlüssel › Amtlicher Gemeindeschlüssel (AGS), 8-stelliger Schlüssel
Neuer Wohnort	› Staat (einschließlich unbekannt), 3-stelliger Schlüssel › Amtlicher Gemeindeschlüssel (AGS), 8-stelliger Schlüssel
Geschlecht	› männlich › weiblich › divers › ohne Angabe
Geburtsdatum	Datumswert
Familienstand	Feste Codeliste
Religion	Feste Codeliste
Staatsangehörigkeit	3-stelliger Schlüssel
Geburtsort	Freitextfeld
Geburtsstaat	3-stelliger Schlüssel
Datum des letzten Zuzugs (bei Fortzügen in das Ausland)	Datumswert
Datum des letzten Fortzugs (bei Zuzügen aus dem Ausland)	Datumswert
Tatsache der An- und Abmeldung von Amts wegen	› nicht zutreffend › Anmeldung von Amts wegen › Abmeldung von Amts wegen

Aus fachlicher Sicht bieten sich für die Signierung neben der Freitextangabe zum Geburtsort insbesondere die Staatsangehörigkeit sowie der Staat des bisherigen Wohnorts und der Staat des neuen Wohnorts an. Dies gilt unter der Annahme, dass viele Personen im Land ihrer Staatsangehörigkeit geboren sind oder aus ihrem Geburtsstaat zuziehen oder dorthin abwandern.

Um einen quantitativen Überblick über Zusammenhänge der möglichen erklärenden Merkmale und des Zielmerkmals zu gewinnen, ist Cramérs V (Cramér, 1946) ein geeignetes Maß. ↘ Grafik 1 stellt die Erhebungsmerkmale der Wanderungsstatistik im Berichtsjahr 2018 und ihren Zusammenhang zum gelieferten Geburtsstaat anhand Cramérs V dar. Hierbei zeigt sich, dass der Geburtsort am stärksten mit dem Geburtsstaat korreliert ist. Erwartungsgemäß weisen die Merkmale Staatsangehörigkeit und Staat des bisherigen und neuen Wohnorts auch einen starken quantitativen Zusammenhang mit dem Geburtsstaat auf, der Zusammenhang mit der Staatsangehörigkeit ist dabei stärker ausgeprägt. Weitere Merkmale (Religion, Tatsache der An- beziehungsweise Abmeldung von Amts wegen, Geschlecht) weisen einen geringeren Zusammenhang mit dem Geburtsstaat auf.

Grafik 1

Zusammenhang zwischen dem Merkmal "Geburtsstaat" und weiteren Erhebungsmerkmalen der Wanderungsstatistik 2018
Cramérs V



Cramérs V liegt immer zwischen 0 und 1.

2020 - 01 - 0202

2.2 Das Merkmal „Geburtsort“

Der Geburtsort ist technisch gesehen ein Pflichtmerkmal in den Datenlieferungen an die Statistik und sollte dementsprechend immer vorhanden sein. Die Nutzung des Merkmals „Geburtsort“ birgt allerdings mehrere Schwierigkeiten: Zum einen liegt keine umfassende Liste über alle Orte der Welt und den jeweils zugehörigen Staat vor. Zum anderen handelt es sich im Datensatz des Meldewesens um eine Freitextangabe, sodass unbrauchbare Angaben nicht ausgeschlossen werden können. Beispielsweise werden verschiedene Varianten von „unbekannt“ und „ohne Angabe“ geliefert. Weiterhin ist die Schreibweise des Geburtsorts nicht standardisiert. Ein Ort kann in vielen verschiedenen Schreibweisen sowie mit Schreibfehlern oder Zusatzinformationen geliefert werden. Es werden auch historische oder ausländische Ortsbezeichnungen geliefert. Eine vollumfassende Auflistung aller möglichen Schreibweisen für alle Orte zu erstellen, ist praktisch unmöglich.

Des Weiteren kann eine Ortsbezeichnung in mehreren Staaten vorkommen. „Neustadt“ beispielsweise gibt es als Gemeindename, Ortsteil oder historische Bezeichnung in vielen Staaten. Anhand der Ortsangabe lässt sich in diesen Fällen der Geburtsstaat nicht eindeutig ermitteln, es müssen weitere Informationen herangezogen werden.

2.3 Technischer Rahmen

Die Signierung soll im Zuge der Datenaufbereitung im Fachverfahren der Wanderungsstatistik durch die Statistischen Ämter der Länder erfolgen. So kann sichergestellt werden, dass plausibilisierte Geburtsstaatsangaben für alle Ämter des Statistischen Verbunds[2] zur Verfügung stehen. Es handelt sich dabei um eine Java-Anwendung mit Zentraler Produktion (das sogenannte OVIS-Anwendungsrahmenwerk; Statistisches Bundesamt, 2020). Meldungen zu Wanderungsfällen sowie Korrekturen werden laufend von den Meldebehörden übermittelt und in den Statistischen Landesämtern plausibilisiert. Das Signierungsverfahren in einem bestehenden Fachverfahren umzusetzen und in die Arbeitsabläufe zu integrieren, stellt eine besondere Herausforderung dar.

2 Das sind die Statistischen Ämter des Bundes und der Länder.

Unter Berücksichtigung dieser Rahmenbedingungen wurden zwei grundsätzliche Ansätze für die Kodierung betrachtet: die regelbasierte Signierung sowie die Signierung mit maschinellen Lernverfahren. Innerhalb dieser groben Kategorien wurden jeweils verschiedene Modelle geprüft und die Ergebnisse des besten regelbasierten Modells dem besten Modell auf Basis maschinellen Lernens gegenübergestellt. Die Signierung wurde für das Berichtsjahr 2018 der Wanderungsstatistik durchgeführt.

3

Regelbasierte Signierung

Als regelbasierte Signierung des Geburtsstaats wird in diesem Aufsatz eine deterministische Signierungsmethode bezeichnet: Sie überprüft die Interplausibilität des Geburtsstaats mit dem Merkmal „Geburtsort“ anhand von Leitdateien und imputiert fehlende oder unplausible Ausprägungen des Merkmals „Geburtsstaat“ nach fest definierten Regeln. Um das Merkmal zuverlässig zu signieren, ist eine möglichst vollständige und korrekte Leitdatei entscheidend.

Folgende Leitdateien wurden für die Untersuchung verwendet:

- Das Ortsverzeichnis des Zensus 2011 deckt weltweit Orte sowie deren Schreibweisen ab, allerdings nur solche, die in den deutschen Melderegistern zum Zeitpunkt des Zensus 2011 enthalten waren.
- Das Gemeindeverzeichnis der Statistischen Ämter des Bundes und der Länder bildet monatlich alle aktuellen deutschen Gemeinden ab.
- Das historische Staatenleitband besteht aus rund 1 100 aktuellen sowie historischen Gebiets- und Staatenbezeichnungen.
- Die Unbekannt-Liste enthält verschiedene Schreibweisen von „unbekannten“ Geburtsortseinträgen (zum Beispiel „nicht bekannt“).

Zusammen decken diese Leitdateien rund 2,2 Millionen unterschiedliche Geburtsort-Staat-Kombinationen ab. Für eine produktive Nutzung eines regelbasierten Verfahrens ist es erforderlich, weitere Ort-Staat-Dateien hinzuzuziehen, die Qualität dieser Dateien zu sichern und sie zu einer Datei zusammenzufügen.

Zusätzlich wurde neben einem direkten Abgleich der gelieferten beziehungsweise standardisierten Geburtsortsangabe mit den Leitdateien noch der Einsatz eines auf dem Jaro-Winkler-Distanzmaß (Jaro, 1989) basierenden Ähnlichkeitsabgleichs untersucht. Ziel war, eine Signierung bei abweichenden Schreibweisen zu erreichen. Dieser Ansatz wurde nicht weiterverfolgt, da die Suche nach gleichen Ortsnamen nach der Standardisierung bereits sehr hohe Zuordnungsquoten lieferte. Die wenigen zu erwartenden zusätzlichen Zuordnungen und die in den Tests relativ hohen Laufzeiten[3] waren dafür ausschlaggebend.

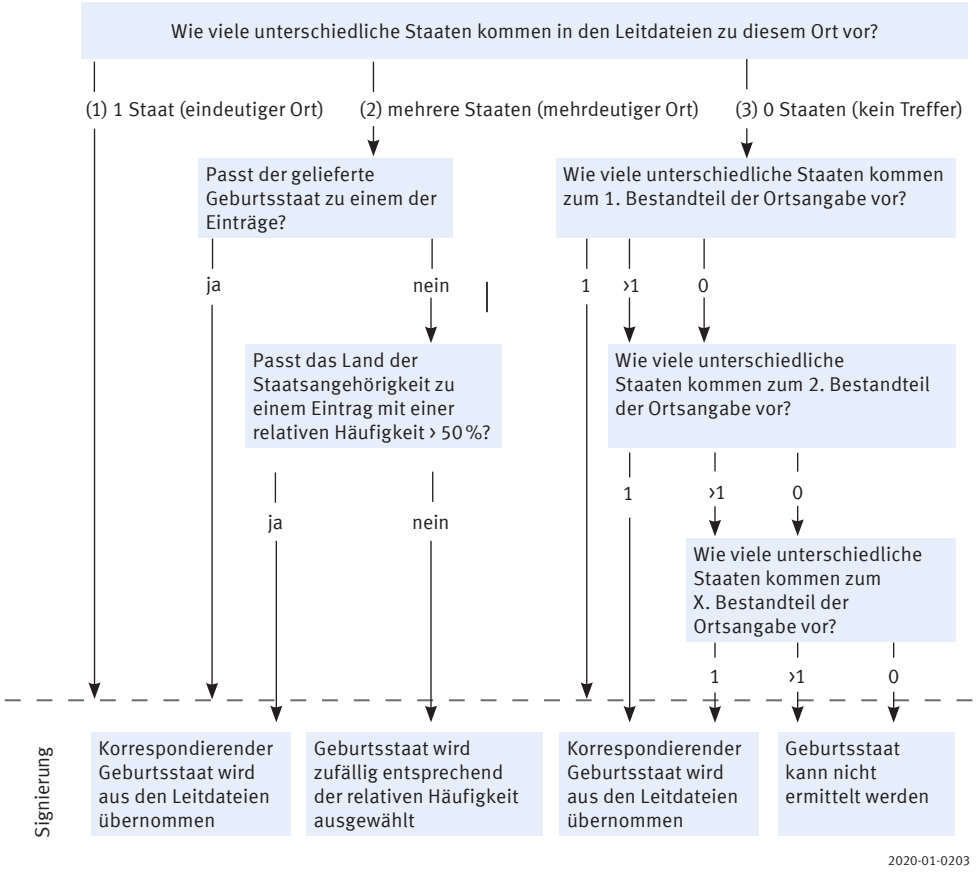
3 In den Tests wurde darauf verzichtet, eine Blocking-Strategie (Blakely/Salmond, 2002) für die Suche in den Leitdateien zu implementieren, was die erheblichen Laufzeiten erklärt. Wäre die Anzahl der zu erwartenden zusätzlichen Zuordnungen größer gewesen, hätte die Umsetzung einer Blocking-Strategie die Laufzeiten erheblich reduziert.

Im ausgewählten Signierungsverfahren werden der Geburtsort, der gelieferte Geburtsstaat, die Staatsangehörigkeit und eine Ort-Staat-Häufigkeitsverteilung berücksichtigt. Somit werden lediglich die zwei am stärksten mit dem Geburtsstaat zusammenhängenden Merkmale verwendet (siehe Abschnitt 2.1). Ein Ziel hierbei ist, das Schätzverfahren relativ einfach und gleichzeitig wenig fehleranfällig zu entwickeln, indem nur starke und intuitiv nachvollziehbare Zusammenhänge einfließen. Ein weiteres Ziel ist, die Kodierungsfunktion möglichst universell einsetzen zu können. Dies soll erreicht werden, indem nur Merkmale verwendet werden, die für alle Datensätze der Wanderungsstatistik sowie weiterer Bevölkerungsstatistiken vorliegen.

Eine Vorstufe der Signierung unterzieht die Geburtsortsangabe einer automatisierten Standardisierung hinsichtlich Groß-/Kleinschreibung, diakritischen Zeichen und Sonderzeichen sowie Leerzeichen. Das Signierungsverfahren setzt unterschiedliche Regeln ein, abhängig davon, ob eine Ortsangabe (1) in den Leitdateien mit einem einzigen Staat vorkommt (= eindeutige Orte), (2) mit mehreren Staaten vorkommt (= mehrdeutige Orte) oder (3) gar nicht vorkommt. ↘ Grafik 2

Ist die Staatszuordnung eindeutig, erfolgt die Signierung mit diesem Geburtsstaat. Ist die Geburtsortsangabe hingegen mehrdeutig, sind weitere Informationen heranzuziehen. Zuerst wird die gelieferte Angabe zum Geburtsstaat überprüft. Wurde der Datensatz bereits mit einer gültigen Angabe zum Geburtsstaat geliefert und ist die Kombination Ort-Staat in den Leitdateien referenziert, wird angenommen, dass der gelieferte Geburtsstaat

Grafik 2
Ablauf der regelbasierten Signierung



korrekt ist. Der Geburtsstaat wird entsprechend signiert. Kann eine solche Signierung nicht erfolgen, wird als Nächstes auf eine Ort-Staat-Häufigkeitsverteilung auf Basis der Lieferdaten früherer Berichtsjahre zugegriffen. Hierbei wird zunächst geprüft, ob das mit der Staatsangehörigkeit übereinstimmende Land mit einer Häufigkeit größer 50 % auftritt. So wird angenommen, dass Wandernde zwar nicht immer die Staatsangehörigkeit des Landes aufweisen, in dem sie geboren sind, dass die Wahrscheinlichkeit dafür jedoch sehr hoch ist, wenn ihre Staatsangehörigkeit zum häufigsten Staat für diesen Geburtsort passt. (Zum Beispiel kann eine Person mit französischer Staatsangehörigkeit außerhalb Frankreichs geboren sein, dies ist aber sehr unwahrscheinlich, sofern ihr Geburtsort Paris ist. Das Gleiche gilt aber nicht für eine Person mit US-amerikanischer Staatsangehörigkeit mit dem Geburtsort Paris – obwohl es eine Stadt namens „Paris“ in den Vereinigten Staaten gibt.) Konnte immer noch kein Geburtsstaat für diesen mehrdeutigen Ort signiert werden, so wird der Staat aus den mit dem Ort referenzierten Staaten zufällig entsprechend der prozentualen Häufigkeit imputiert. So wird eine plausible Verteilung nach Staaten ermittelt, obwohl eine personenscharfe Zuweisung nicht sichergestellt ist.

Kommt die Ortsangabe gar nicht in den Leitdateien vor, wird sie nach der Position von fest definierten Zerlegungstriggern[4] in mehrere Bestandteile aufgeteilt und diese Bestandteile werden mit den Leitdateien abgeglichen. Ein Beispiel dafür ist die Aufteilung von „Frankfurt a. M. (Hessen)“ in „Frankfurt a. M.“ und „Hessen“. Sobald ein eindeutiger Treffer gefunden wird, wird der Geburtsstaat signiert; weitere Ortsnamensbestandteile werden nicht weiter berücksichtigt. Gibt es nach Abgleich aller Bestandteile keinen oder keinen eindeutigen Treffer, so wird der Geburtsstaat als „unbekannt“ signiert. Einen Überblick über die Zuweisungsschritte der regelbasierten Signierung des Geburtsstaats in den Wanderungen 2018 gibt ↘ Tabelle 1.

Tabelle 1
Regelbasierte Signierung der Wanderungen 2018 nach Zuweisungsschritten

Abgleich Leitdateien	Grund der Signierung	Ergebnis der Signierung	Signierte Datensätze	
			Anzahl	%
(1) Eindeutiger Ort	Der Geburtsort kommt eindeutig in den Leitdateien (ohne Unbekannt-Liste) vor.	bekannter Geburtsstaat	4 462 997	66,4
(2) Mehrdeutiger Ort	Die Kombination Geburtsort und gelieferter Geburtsstaat kommt in den Leitdateien vor.	bekannter Geburtsstaat	1 205 491	17,9
(2) Mehrdeutiger Ort	Die Kombination Geburtsort und Land der Staatsangehörigkeit kommt in den Leitdateien mit einer relativen Häufigkeit › 50 % vor.	bekannter Geburtsstaat	546 815	8,1
(2) Mehrdeutiger Ort	Der Geburtsstaat wird zufällig aus mehreren mit dem Geburtsort vorkommenden Staaten ausgewählt.	bekannter Geburtsstaat	80 799	1,2
(3) Unbekannter Ort	Ein Ortsnamensbestandteil kommt eindeutig in den Leitdateien vor.	bekannter Geburtsstaat	99 140	1,5
(1) Eindeutiger Ort oder (3) Unbekannter Ort	Der Geburtsstaat kann nicht ermittelt werden.	Geburtsstaat ist unbekannt	321 177	4,8
Insgesamt			6 716 419	100

4 Maschinelles Lernen

Maschinelle Lernverfahren sind in der Lage, Muster in Datenbeständen zu erkennen. Überwachte Verfahren lernen Muster auf Basis von Eingabedaten, in denen der zu lernende Zusammenhang bereits vorhanden ist. Im hier diskutierten Fall liegen Geburtsstaatskodierungen aus Vorperioden vor; daher wurde geprüft, ob überwachte Lernverfahren geeignet sind, den Geburtsstaat zu signieren.

Bei der Auswahl der zu testenden Verfahren waren Zeit- und Hardware-Restriktionen zu berücksichtigen. Für die Tests in diesem Projekt stand ein Rechner mit 8 Threads und 32 Gigabyte zur Verfügung, als Programmiersprache R.

4 Das sind beispielsweise Schrägstriche, Klammern oder Kommas.

4.1 Der Trainingsdatensatz

Um eine Geburtsstaatssignierung mit maschinellen Lernverfahren vorzubereiten, wurde zunächst ein Trainingsdatensatz aus dem Jahresmaterial 2017 der Wanderungsstatistik erstellt. Ein Trainingsdatensatz sollte sehr zuverlässig sein, damit durch die maschinellen Lernverfahren keine falschen Muster gelernt werden. Um die Zuverlässigkeit der Ort-Staat-Zuordnungen zu verbessern, wurden die gelieferten Geburtsort-Geburtsstaat-Kombinationen anhand der in Kapitel 3 beschriebenen Leitdateien geprüft. Für den Trainingsdatensatz wurden aus dem Jahresmaterial 2017 diejenigen Datensätze entfernt, deren Geburtsort (1) nicht in den Leitdateien gefunden wurde oder (2) zu einer anderen Staatszuweisung als dem gelieferten Geburtsstaat geführt hat. Datensätze mit dem Geburtsstaat „unbekannt“ wurden nur dann beibehalten, wenn die Geburtsortsangabe offensichtlich unbrauchbar war. Das traf beispielsweise zu, wenn der Datensatz gar keine alphanumerischen Zeichen oder eine Variante von „unbekannt“ enthielt. Somit bestand der Trainingsdatensatz aus rund 6,2 Millionen Wanderungsdatensätzen.

4.2 Zeichenkettenähnlichkeit und N-Gramme

Eine wichtige Besonderheit bei Freitextfeldern sind Abweichungen in der Schreibweise (siehe Abschnitt 2.2). Um diese Abweichungen beim Training der Lernverfahren berücksichtigen zu können, wurden die Geburtsorte in N-Gramme zerlegt. N-Gramme (Kondrak, 2005) sind alle Fragmente eines Textes oder einer Zeichenkette mit N aufeinanderfolgenden Elementen. Bigramme bestehen entsprechend aus zwei aufeinanderfolgenden Elementen.[5]

Für die vorliegende Anwendung wurden sämtliche Bigramme[6] ermittelt, die in allen beobachteten Ausprägungen des Geburtsorts vorkommen. Anschließend wurde für jede Beobachtung ermittelt, wie oft jedes N-Gramm in der Ausprägung des Merkmals „Geburtsort“ vorkommt. Das Ergebnis ist eine Matrix mit der Anzahl der Beobachtungen und der Anzahl der Bigramme als Dimension, deren Elemente nichtnegative ganze Zahlen sind. Für das Lernen macht man sich zunutze, dass ähnliche Zeichenketten auch ähnliche Ausprägungen der Matrixelemente haben (↘ Tabelle 2). Um diese Bigramme wurde das Trainingsmaterial ergänzt.

4.3 Maschinelle Lernverfahren

Für die Tests mit maschinellen Lernverfahren kamen in Anbetracht der Hardware-Restriktionen Naïve-Bayes- sowie baumbasierte Klassifikatoren in Betracht. Als erklärende Variablen wurden zunächst alle im Datensatz vorliegenden Merkmale verwendet. Die Verfahren wurden zunächst anhand des Trainingsmaterials von 2017 und eines 15-%-Testdatensatzes getestet. Nachdem der Naïve-Bayes-Klassifikator (Webb, 2011) mit dem Trainingsdatensatz ohne Bigramm-Erweiterung sehr schnell relativ gute Ergebnisse lieferte (rund 87 % korrekte Signierungen), gelang es nicht, mit dem um Bigramme erweiterten Datensatz bessere Ergebnisse zu erzielen.

5 Zum Beispiel sind „Pa“, „ar“, „ri“ und „is“ die Bigramme von „Paris“.

6 Es wurden Bigramme verwendet, weil sie relativ wenige zusätzliche Merkmale erzeugen. Die Bigramme wurden mit dem R-Paket stringdist (van der Loo, 2014) erstellt.

Tabelle 2
Beispielhafte Matrix von Schreibweisen und Bigrammen

Schreibweisen von Mumbai (standardisiert)	Bigramme																		
	MB	BA	MU	UM	AI	_M	AY	BO	I_	OM	M_	MS	SA	_S	AS	BI	BS	IA	Y_
MUMBAI_M_S	1	1	1	1	1	1	0	0	1	0	1	0	0	1	0	0	0	0	0
MUMBAI_MS	1	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0
MUMBAY	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
MUMBIA	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
MUMBSAI_MS	1	0	1	1	1	1	0	0	1	0	0	1	1	0	0	0	1	0	0
BOMBAY	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
BOMBAY_M_S	1	1	0	0	0	1	1	1	0	1	1	0	0	1	0	0	0	0	1
BOMBAI	1	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0

Aus diesem Grund wurden weitere Tests mit baumbasierten Verfahren und speziell mit Random Forest durchgeführt.

Random Forest (Breiman, 2001) ist ein Algorithmus des überwachten Lernens, der auf Basis von Entscheidungsbäumen Klassifikations- und Regressionsmodelle anpasst. Dabei wird im Training mit den Merkmalsausprägungen und anhand der in den Trainingsdaten bekannten Klassen eine Reihe von Entscheidungsbäumen errechnet. Im Klassifikationsschritt werden die ermittelten Bäume verwendet, um anhand der Merkmalsausprägungen einer Beobachtung die unbekannte Klasse zu schätzen. Der Random-Forest-Algorithmus erzielte mit einem Trainingsdatensatz, der um Bigramme erweitert wurde, jedoch zusätzlich nur die Merkmale „Staatsangehörigkeit“ und „Staat des bisherigen Wohnorts“ enthält, erheblich bessere Ergebnisse als der Naïve-Bayes-Klassifikator.[7] Hier wurden bezogen auf den Testdatensatz insgesamt 97 % korrekte Zuordnungen erreicht.[8] Aufgrund dieses Ergebnisses wurde die Signierung des Geburtsstaats in der Wanderungsstatistik 2018 mit diesem Modell geprüft.

7 Auf weitere erklärende Merkmale musste aufgrund der bereits erwähnten Hardware-Beschränkungen verzichtet werden. Die Merkmale „Staatsangehörigkeit“ und „Staat des bisherigen Wohnorts“ wurden ausgewählt, weil sie einen relativ hohen Zusammenhang mit dem Geburtsstaat haben (siehe Grafik 1).

8 Das Modell wurde parallel auf 8 Kernen mit knapp 32 Gigabyte Arbeitsspeicher mit dem R-Paket ranger (Wright/Ziegler, 2017) in gut zwei Tagen angepasst.

5

Ergebnisse und Bewertung

5.1 Ergebnisse

Nachdem das regelbasierte Signierungsverfahren und ein Random-Forest-Modell vorlagen, wurden die Geburtsstaaten in den Wanderungsdaten des Jahres 2018 mit beiden Verfahren signiert und die Ergebnisse verglichen. Bei der Beurteilung einer Übereinstimmung zwischen geliefertem und signiertem Wert wurden fehlende Geburtsstaatseinträge mit „Deutschland“ gleichgesetzt, da gemäß der melderechtlichen Regelungen der Geburtsstaat nur im Falle einer Geburt im Ausland anzugeben ist (siehe Kapitel 1).[9]

↘ Tabelle 3 zeigt, dass beide Verfahren einen Großteil der Wanderungsfälle mit einem bekannten Geburtsstaat signieren. Bei der regelbasierten Signierung entspricht dies rund 95 %, bei der Signierung mit Random Forest 99 % der Wanderungsfälle. Beim Random-Forest-Ansatz verblieb somit in etwa 50 000 Fällen der Geburtsstaat mit „unbekannt“ signiert gegenüber 320 000 Fällen bei der regelbasierten Signierung.

Im Folgenden soll näherungsweise die Zuverlässigkeit der Signierungsergebnisse ermittelt werden. Dazu werden die rund 6,4 Millionen beziehungsweise 6,7 Millionen Fälle mit signierten Geburtsstaaten beider Verfah-

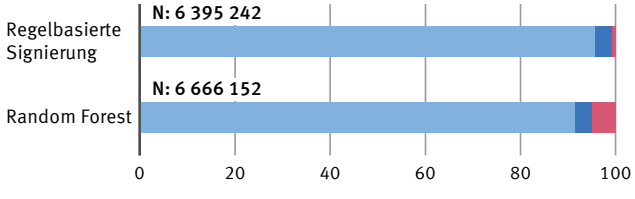
9 Ebenso werden Umkodierungen mit ehemaligen Staaten und deren Nachfolgestaaten (zum Beispiel der gelieferte Wert ist Sowjetunion und der signierte Wert Russland) sowie mit abhängigen Gebieten und deren Mutterstaat als Übereinstimmung ausgewiesen.

Tabelle 3
Zuweisungsquote der regelbasierten Signierung gegenüber Signierung mit Random Forest-Modell zum Geburtsstaat anhand der Wanderungsstatistik 2018

	Regelbasierte Signierung		Random Forest	
	Anzahl	%	Anzahl	%
Insgesamt	6 716 419	100	6 716 419	100
Geburtsstaat signiert	6 395 242	95,2	6 666 152	99,3
Gelieferte und signierte Werte stimmen überein	6 110 490	91,0	6 093 769	90,7
Leer/unbekannt wird geliefert und signierter Wert weicht ab	232 353	3,5	248 857	3,7
Bekannter Staat wird geliefert und signierter Wert weicht ab	52 399	0,8	323 526	4,8
„unbekannt“ signiert	321 177	4,8	50 267	0,7

ren betrachtet. Abgleiche dieser Signierungsergebnisse mit deren gelieferten Geburtsstaaten ergeben eine Größenordnung, in wie vielen Fällen die Signierung korrekt funktioniert. Hierunter fallen für beide Signierungsverfahren rund 6,1 Millionen Datensätze, für die die Signierung den gelieferten Geburtsstaat ergibt. Der signierte Geburtsstaat wird somit als wahrscheinlich korrekt angenommen. Hinzu kommen für beide Verfahren zwischen 230 000 und 250 000 Fälle, in denen der Geburtsstaat „leer“ oder „unbekannt“ geliefert wurde und durch das Signierungsverfahren auf einen gültigen Staat (außer Deutschland bei einem leeren Geburtsstaat) signiert wird. Unter diesen Fällen befinden sich wahrscheinlich viele Fälle, in denen die Eintragung fälschlicherweise nicht vorgenommen wurde (siehe Kapitel 1). Somit sollte in diesen Fällen in der Regel der signierte Geburtsstaat zuverlässiger sein als die Angabe „leer“, die auf eine Geburt in Deutschland hindeutet. Beim regelbasierten Verfahren wird in rund 50 000 Fällen ein anderer als der gelieferte bekannte Staat signiert, im Falle der Signierung mit Random Forest sind es mehr als 320 000 Fälle. Dabei handelt es sich wahrscheinlich meist um Fehlzuweisungen durch das Signierungsverfahren, wobei in Einzelfällen auch der gelieferte Staat falsch sein kann. Bei der regelbasierten Signierung liegt somit eine Fehlzuweisung in der Größenordnung von 0,8 % bezogen auf 6,4 Millionen signierte Fälle vor, beim Random-Forest-Verfahren sind es 4,9 % bezogen auf 6,7 Millionen signierte Fälle. Die höhere Zahl signierter Datensätze des Random-Forest-Modells (99,3 %), die aufgrund von weniger „unbekannt“-Zuweisungen gegenüber der regelbasierten Signierung zustande kommt, geht offenbar mit mehr Fehlzuweisungen in ähnlicher Größenordnung einher. ↘ Grafik 3

Grafik 3
Abgleich gelieferter und signierter Geburtsstaaten bei erfolgter Signierung
in %



2020 - 01 - 0204

Hintergrund der unterschiedlichen Zuweisungsquoten ist, dass das regelbasierte Verfahren bei ergebnisloser Suche in den Leitdateien keine Zuweisung vornimmt, während das maschinelle Lernverfahren den Verteilungen im Trainingsmaterial entsprechend in der Regel einen Staat zuweist.

↘ Tabelle 4 vergleicht nun detailliert die Ergebnisse der regelbasierten Signierung, der Signierung des Random-Forest-Modells und den für das Berichtsjahr 2018 von den Meldeämtern an die Statistik gelieferten Geburtsstaat.

In fast 88 % der Wanderungsfälle 2018 wurde die gelieferte Geburtsstaatsangabe durch beide Signierungsverfahren bestätigt. Die Angabe ist somit mit hoher Wahrscheinlichkeit korrekt. In knapp 4 % der Fälle stimmen beide Signierungsverfahren überein, ermitteln aber einen anderen Geburtsstaat als den gelieferten. Der gelieferte Staat wird somit als unplausibel angenommen. Es handelt sich hierbei insbesondere um „leer“ gelieferte Geburtsstaatsfelder, die nicht auf Deutschland kodiert werden (siehe Zeile B1). Deutlich seltener kommt es vor, dass ein gelieferter Staat von beiden Verfahren als unbekannt kodiert wird (B2) oder dass ein ausländischer Staat von beiden Verfahren auf einen anderen gültigen Staat umgeschlüsselt wird (B3).

Etwa 8 % der Wanderungsfälle weisen eine abweichende Signierung durch das regelbasierte Verfahren und das Random-Forest-Modell auf. Darunter sind etwa 3 % der Wanderungsfälle, für die die regelbasierte Signierung den gelieferten Geburtsstaat ermittelt hat. Das Ergebnis der regelbasierten Signierung wird somit als plausibel und das Ergebnis der Signierung mit dem Random-Forest-Modell als wahrscheinlich falsch angenommen. In weiteren 3 % der Wanderungsfälle ergibt die Signierung mit Random Forest den gelieferten Geburtsstaat. Die regelbasierte Signierung weicht hiervon häufig ab, da das Verfahren keinen Geburtsstaat ermitteln konnte (169 707 Datensätze, das entspricht 83 %). Für rund 2 % der Datensätze liefert der Vergleich unklare Ergebnisse. In den meisten Fällen weichen der gelieferte Staat sowie beide Signierungsergebnisse voneinander ab.[10] Ohne manuelle Prüfung ist in diesen Fällen die Richtigkeit

10 Weiterhin können hierunter Fälle ehemaliger Staaten oder abhängiger Gebiete fallen, beispielsweise wenn der Staat „Jugoslawien“ geliefert wurde und beide Signierungsverfahren unterschiedliche Nachfolgestaaten von Jugoslawien ermitteln.

Tabelle 4
Wanderungsstatistik 2018: Gelieferter Geburtsstaat gegenüber Signierungsergebnissen

Gruppe	Gelieferter Wert = Random Forest	Gelieferter Wert = regelbasierte Signierung	Regelbasierte Signierung = Random Forest	Bedeutung	Anzahl	Anteil in %
Insgesamt					6 716 419	100
A	ja	ja	ja	Übereinstimmung aller drei Werte	5 905 527	87,93
B	nein	nein	ja	Gelieferter Wert wahrscheinlich inkorrekt	256 127	3,81
B1	gelieferter Wert ist leer/unbekannt				204 163	3,04
B2	gelieferter Wert ist bekannt und signierter Wert ist unbekannt				25 686	0,38
B3	gelieferter und signierter Wert entsprechen bekanntem Schlüssel				26 278	0,39
C	nein	ja	nein	Regelbasierte Signierung wahrscheinlich korrekt	222 759	3,32
C1	Regelbasierte Signierung ergibt „Deutschland“				18 672	0,28
C2	Regelbasierte Signierung ergibt ausländischen Staat				202 616	3,02
C3	Regelbasierte Signierung ergibt unbekannten Staat				1 471	0,02
D	ja	nein	nein	Random Forest wahrscheinlich korrekt	204 717	3,05
D1	Random Forest ergibt „Deutschland“				25 415	0,38
D2	Random Forest ergibt ausländischen Staat				179 298	2,67
D3	Random Forest ergibt unbekannten Staat				4	0,00
E	Sonstiges			Keine Übereinstimmung	127 289	1,90

sowohl der erreichten Kodierungen als auch der gelieferten Werte nicht feststellbar.

5.2 Bewertung

Zusammenfassend liefern beide Signierungsverfahren ähnlich gute Ergebnisse. Beide Verfahren sind in der Lage, fälschlicherweise fehlende Geburtsstaatseinträge relativ zuverlässig zu identifizieren und können somit einen systematisch zu hohen Ausweis des Geburtsstaats „Deutschland“ vermeiden. Für beide Verfahren liegt die Zahl an unbekannten Zuordnungen nach der Signierung unter 5 %. Die Verfahren liefern zu einem hohen Teil identische Ergebnisse, aber in rund 8 % der Fälle abweichende Zuordnungen. Dies liegt zum Teil darin begründet, dass das Random-Forest-Modell fast immer eine Staatszuweisung trifft, wenn nicht recht sicher ein unbekannter Staat festgestellt wurde. Im Falle des regelbasierten Verfahrens hingegen wird „unbekannt“ zugewiesen, wenn der Ort beziehungsweise ein Bestandteil nicht in den Leitdateien gefunden wurde. Infolge dieses ersten Tests erweisen sich beide Verfahren als grundsätzlich für die Kodierung aus methodischer Sicht geeignet.

Neben der fachlichen Eignung müssen daher weitere Kriterien für die Entscheidung für ein Verfahren zum Einsatz in der Fachanwendung der Wanderungsstatistik herangezogen werden. ↘ Übersicht 2 auf Seite 108

Beide Verfahren haben Potenzial, noch zuverlässigere Signierungsergebnisse zu erreichen. Die Quote der im regelbasierten Verfahren als „unbekannt“ signierten Fälle kann durch die Erstellung, Qualitätssicherung und laufende Aktualisierung einer umfangreichen Ortsdatei weiter sinken. In den Tests wurde beispielsweise mit dem Ortsverzeichnis des Zensus 2011 eine nicht aktualisierte Datei verwendet. Dies wirkt sich insofern aus, dass es seit 2011 vermehrt Zuwanderungen aus Staaten gegeben haben dürfte, für die bislang wenige Orte in der Datei erfasst sind. Beispielsweise umfasst die Datei nur wenige syrische Orte, während Syrien 2018 der vierthäufigste gemeldete Geburtsstaat für Wandernde ist. Das maschinelle Lernverfahren kann durch Wahrscheinlichkeitsschwellen relativ „unsichere“ Signierungen mit „unbekannt“ anstatt mit einem Geburtsstaat signieren. Auch die Erhöhung des Umfangs des Trainingsmaterials sowie das Training mit weiteren N-Grammen mit $N > 2$ kann geprüft werden. Der Einsatz anderer Lernverfahren, wie etwa Gradient-Boosting, kann ebenfalls getestet werden.

Übersicht 2
Bewertungskriterien für den Einsatz

	Regelbasierte Signierung	Signierung mit Methoden des maschinellen Lernens
Weitere Verbesserung der Verfahren	› Qualitätssicherung und Aktualisierung der Ortsdatei	› Umfang des Trainingsdatensatzes vergrößern › Einsatz von Wahrscheinlichkeitsschwellen › Tests weiterer N-Gramme › Tests mit anderen Verfahren
Einsetzbarkeit für andere Statistiken	› abhängig von Merkmalsverfügbarkeit grundsätzlich gegeben	› abhängig von Merkmalsverfügbarkeit grundsätzlich gegeben
Laufende Aufwände	› Pflege einer Ortsdatei	› Durchführung des Trainings
Risiken bei der technischen Umsetzung	› gering, da erforderliche IT-Methoden grundsätzlich bekannt	› technische Unwägbarkeiten hinsichtlich Performanz › geringe Erfahrungswerte der Implementierung solcher Verfahren in bestehende Fachanwendungen
Fachliche Risiken beim Einsatz der Methoden	› Signierung neuer Orte	› geringere Erklärbarkeit und Nachvollziehbarkeit der Signierung › Signierung neuer Orte

Das Projekt hat zudem gezeigt, dass wegen bestehender Hardware-Restriktionen einige maschinelle Lernverfahren nicht getestet und Modelloptimierungen nicht durchgeführt werden konnten. Dies ist zurzeit ein fachübergreifendes Problem, dessen Lösung erhebliches Potenzial für den Einsatz maschineller Lernverfahren birgt.

Beide Verfahren sind grundsätzlich auf andere Statistiken übertragbar, wobei die Verfügbarkeit von Merkmalen sowie Besonderheiten der jeweiligen Statistik berücksichtigt werden müssen. Für beide Verfahren ist ein laufender Aufwand zu erwarten: Für das regelbasierte Verfahren ist die Ortsdatei zu aktualisieren und hierfür fehlende Zuordnungen oder möglicherweise falsche Zuordnungen sind jährlich auszuwerten. Ein Verfahren auf Basis maschineller Lernverfahren muss regelmäßig neu trainiert werden, um aktuelle Entwicklungen einzubeziehen. Bei der technischen Umsetzung werden aktuell größere Risiken bei der Signierung mit Verfahren des maschinellen Lernens erwartet, da IT-seitig zurzeit noch Erfahrungswerte mit der Implementierung solcher Methoden in bestehende Fachanwendungen fehlen. Darüber hinaus wäre das Zustandekommen der Signierungsergebnisse im Einzelfall bei Random Forest weniger gut nachvollziehbar als im Falle der regelbasierten Signierung. Wegen der genannten Argumente wird für die Wanderungsstatistik das regelbasierte Signierungsverfahren ausgewählt.

6 Fazit und Ausblick

Der vorliegende Beitrag hat zwei Verfahren zur Kodierung des Geburtsstaats in der Wanderungsstatistik verglichen: ein regelbasiertes Verfahren auf Basis von Ort-Staat-Leitdateien sowie ein Random-Forest-Modell. Ziel war, ein Verfahren zu identifizieren, das auf Basis vorliegender Merkmale zuverlässig den Geburtsstaat ermittelt. Die Tests ergaben, dass beide Verfahren für die Kodierung des Geburtsstaats in der Wanderungsstatistik grundsätzlich geeignet sind. Weil die Risiken der IT-seitigen Umsetzung geringer sowie die Signierungsergebnisse besser nachvollziehbar sind, wird der Einsatz des regelbasierten Verfahrens in der Wanderungsstatistik und in weiteren Bevölkerungsstatistiken angestrebt. Eine Umsetzung in der Fachanwendung für die Wanderungsstatistik ist beabsichtigt.

Im Zuge der Analysen wurden für beide Verfahren noch Verbesserungspotenziale erkannt, die in die künftigen Arbeiten einfließen sollen: Das regelbasierte Verfahren signierte seltener einen Geburtsstaat als das Random-Forest-Modell, die erreichten Zuordnungen waren aber in der Regel zuverlässiger. Bis zum produktiven Einsatz der Signierungsfunktion ist geplant, eine Ortsdatei aus mehreren Ort-Staat-Referenzdateien zu erstellen, deren Qualität zu sichern und um fehlende Ort-Staat-Kombinationen zu ergänzen. Darüber hinaus soll die Ortsdatei

laufend gepflegt werden. Hierbei sollen Datensätze, für die bei der Signierung keine oder eine unplausible Zuweisung getroffen wurde, jährlich ausgewertet und in die Ortsdatei eingepflegt werden. Die Ergebnisse gegebenenfalls erfolgter manueller Nachprüfungen und Korrekturen durch die Statistischen Landesämter sollen ebenfalls laufend automatisiert in die Qualitätssicherung einfließen. Mit diesen Verbesserungen ist zu erwarten, dass Zuverlässigkeit und Vollständigkeit der Ortsdatei und damit der Geburtsstaatssignierungen im Zeitverlauf weiter zunehmen. Weiterhin kann untersucht werden, ob maschinelle Lernverfahren auch bei der Qualitätssicherung und Pflege der Ortsdatei eingesetzt werden können. Hiermit können abweichende Schreibweisen von Orten sowie aktuelle Entwicklungen, die in den Trainingsdaten bereits erkennbar sind, sich in der Ortsdatei aber noch nicht widerspiegeln, bei der Verbesserung der Ortsdatei berücksichtigt werden. Mit den vorgestellten Ergebnissen stehen zukunftsfähige und effiziente Werkzeuge für die Geburtsstaatssignierung zur Verfügung, die in der Wanderungsstatistik angewendet und auf weitere Bevölkerungsstatistiken übertragen werden können.

LITERATURVERZEICHNIS

Blakely, Tony/Salmond, Clare. *Probabilistic record linkage and a method to calculate the positive predictive value.* In: International Journal of Epidemiology. Ausgabe 6. Jahrgang 31, 2002, Seite 1246 ff. doi: 10.1093/ije/31.6.1246

Breiman, Leo. *Random forests*. In: Machine Learning. Ausgabe 1/45, 2001, Seite 5 ff., doi: 10.1023/A:1010933404324

Cramér, Harald. *Mathematical Methods of Statistics*. Princeton 1946, hier: Seite 282 ff. ISBN 0-691-08004-6

Carow, Annelen/Mundil-Schwarz, Rabea/Vigneau, Elsa. *Bevölkerung am üblichen Aufenthaltsort und Weiterentwicklung des Schätzverfahrens zur Langzeitmigration*. In: WISTA Wirtschaft und Statistik. Ausgabe 3/2019, Seite 65 ff.

Jaro, Matthew A. *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida*. In: Journal of the American Statistical Association. Ausgabe 406. Jahrgang 84, 1989, Seite 414 ff.

Kondrak, Grzegorz. *N-Gram Similarity and Distance.* In: Consens, Mariano/Navarro, Gonzalo (Herausgeber). SPIRE 2005: String Processing and Information Retrieval. Lecture Notes in Computer Science 3772. Berlin/Heidelberg, 2005.

Mundil, Rabea/Grobecker, Claire. *Schätzverfahren zu Langzeitmigranten in Deutschland 2009, Teil 1: Deutsche Personen*. In: Wirtschaft und Statistik. Ausgabe 10/2011, Seite 967 ff.

Statistisches Bundesamt. *OVIS-Anwendungsrahmenwerk. Datenaufbereitung*. [Zugriff am 14. April 2020]. Verfügbar unter: www.destatis.de

van der Loo, Mark P. J. *The stringdist Package for Approximate String Matching*. In: The R Journal. Jahrgang 6. Ausgabe 1/2014, Seite 111 ff. [Zugriff am 14. April 2020]. Verfügbar unter: https://CRAN.R-project.org/package=stringdist.

Webb, Geoffrey I. *Naïve Bayes*. In: Sammut, Claude/Webb, Geoffrey I. (Herausgeber). Encyclopedia of Machine Learning. Boston 2011.

Wright, Marvin N./Ziegler, Andreas. *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R*. In: Journal of Statistical Software. Band 77. 2017, Seite 1 ff., doi: 10.18637/jss.v077.i01.

RECHTSGRUNDLAGEN

Gesetz über die Statistik der Bevölkerungsbewegung und die Fortschreibung des Bevölkerungsstandes (Bevölkerungsstatistikgesetz – BevStatG) vom 20. April 2013 (BGBl. I Seite 826), das zuletzt durch Artikel 9 des Gesetzes vom 18. Dezember 2018 (BGBl. I Seite 2639) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Juni 2020
Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Artikelnummer: 1010200-20003-4, ISSN 1619-2907