AUFBEREITUNG VON ANGABEN ZU GEBURTSORTEN IM AUSLÄNDERZENTRALREGISTER MIT OPENSTREETMAP

Jan Eberle

Schlüsselwörter: Ausländerzentralregister – Schutzsuchende – Geburtsort – Geokodierung – OpenStreetMap

ZUSAMMENFASSUNG

Angaben zu Geburtsorten von Ausländerinnen und Ausländern ermöglichen eine Vielzahl neuer Auswertungsmöglichkeiten für amtliche Statistiken. Kommen somalische Schutzsuchende vermehrt aus Regionen, die durch Dürre und Nahrungsmittelknappheit besonders stark vom Klimawandel betroffen sind? In welchen US-amerikanischen und indischen Regionen gelingt das Anwerben von Fachkräften für den heimischen Arbeitsmarkt? Für die Beantwortung dieser und ähnlicher Fragen kann die Angabe zum Geburtsort als Indikator für die regionale Herkunft einer Person innerhalb eines Staates herangezogen werden. Der Artikel liefert eine Machbarkeitsstudie zur Aufbereitung und Geokodierung von Angaben zum Geburtsort im Ausländerzentralregister mithilfe der Geodaten von OpenStreetMap. Die Ergebnisse stehen auch als interaktives Kartenangebot zur Verfügung.

Keywords: *Central Register of Foreigners – people seeking protection – birthplace – geocoding – OpenStreetMap*

ABSTRACT

Data on the birthplaces of foreigners create a multitude of new options for analysis in official statistics. Do Somali people seeking protection mostly come from regions that are particularly affected by climate change because of droughts and food shortages? From which U.S. and Indian regions can Germany successfully attract skilled workers for the domestic labour market? To answer these and similar questions, information on the birthplace is useful as an indicator of a person's regional origin within a country. This article presents a feasibility study on the processing and geocoding of birthplace data held in the Central Register of Foreigners, using OpenStreetMap geodata. The results are also available as an interactive map offer.

Jan Eberle

ist Volkswirt und angehender Data Scientist. Im Statistischen Bundesamt arbeitet er als Referent im Bereich der laufenden Bevölkerungsstatistiken. Zu seinen Arbeitsschwerpunkten gehören dabei auch Auswertungen des Ausländerzentralregisters zur ausländischen Bevölkerung sowie zu Schutzsuchenden.

Interaktive Karte

Die Ergebnisse können Sie auf einer interaktiven Karte erkunden.

Datenschutzhinweis:
Durch die Nutzung der interaktiven Karte erklären Sie sich mit der Verarbeitung personenbezogener Daten wie der IP-Adresse und technisch notwendiger Cookies, beispielsweise zum Zweck einer aggregierten Nutzungsstatistik, einverstanden.

Zur interaktiven Karte

1

Einleitung

Die Nutzung von Verwaltungsdaten und die damit einhergehende Entlastung der Auskunftspflichtigen ist ein Kernziel der Digitalen Agenda des Statistischen Bundesamtes (Statistisches Bundesamt, 2019) und auch in seiner Hausstrategie fest verankert. Die Nutzung des Ausländerzentralregisters (AZR) hat dabei eine lange Tradition in der Bevölkerungsstatistik. Bereits seit den 1970er-Jahren werden dessen Daten aufbereitet, um Informationen zur ausländischen Bevölkerung in Deutschland zu gewinnen (Fleischer, 1989).

Das AZR ist eines der größten Verwaltungsregister in Deutschland. Es enthält Informationen zu allen ausländischen Personen, die sich nicht nur vorübergehend in Deutschland aufhalten. In der Regel sind das Personen, deren Aufenthalt länger als drei Monate andauert.[1] Das AZR führt die Datenbestände aller Behörden zusammen, die mit asyl- und aufenthaltsrechtlichen Aufgaben betraut sind, und dient diesen als zentrale Informationsplattform.

Wie bei Verwaltungsdaten üblich, ist die statistische Verwertbarkeit der Daten damit nur einer von vielen Verwendungszwecken. Entsprechend unterscheidet sich die Datenqualität auch von jener von Daten aus Primärerhebungen. Vor einer Verwertung der Daten für statistische Zwecke steht daher eine sorgfältige Qualitätsprüfung und Aufbereitung.

Infolge der Fluchtmigration der Jahre 2015 und 2016 hat der Gesetzgeber das AZR kontinuierlich weiterentwickelt und dabei auch die Nutzungsmöglichkeiten für die amtliche Statistik weiter verbessert.[2] Seit Anfang 2020 erhält das Statistische Bundesamt zusätzlich Angaben zu Geburtsstaaten und Geburtsorten aus dem Ausländerzentralregister für die Verwendung in der amtlichen Statistik (§ 23 AZR-Gesetz). Die beiden Merkmale werden im Verwaltungskontext von Ausländerbehörden, Erstaufnahmeeinrichtungen oder Polizeibehörden ohne Plausibilitätskontrollen erfasst. Die Erfassung des Geburtsstaats erfolgt strukturiert als ISO-Code[3], ist als freiwillige Angabe jedoch in erheblichem Umfang unvollständig. Angaben zum Geburtsort werden in dem Register als Pflichtangabe weitgehend vollständig erfasst, allerdings erfolgt die Erfassung unstrukturiert als Freitextangabe.

Vor diesem Hintergrund hat das Statistische Bundesamt zunächst ein Verfahren zur Aufbereitung der Angaben zum Geburtsstaat entwickelt (Canan/Eberle, 2022). Der vorliegende Aufsatz stellt darauf aufbauend ein Konzept zur Aufbereitung und Geokodierung der Angaben zum Geburtsort vor.

Lösungsansätze einer vergleichbaren Problemstellung liefern Šimbera und andere (2021) für administrative Daten aus Tschechien sowie Conti und Cimbelli (2018) für Italien. Šimbera und andere (2021) entwickeln einen eigenen Matching-Algorithmus mit aufwendigem Preprocessing zur Standardisierung der Ortsangaben. Conti und Cimbelli (2018) hingegen nutzen mit OpenRefine ein bestehendes Open-Source-Werkzeug für den gesamten Aufbereitungsprozess. Hierbei verbleiben gewisse Details, wie der Umgang mit mehrdeutigen Ortsnamen, in der Black Box. Das in diesem Artikel vorgeschlagene Aufbereitungskonzept versucht, einen Mittelweg zwischen Aufwand und Transparenz zu gehen.

Als konkretes Anwendungsbeispiel sind die Geburtsorte von Ende 2020 im AZR registrierten Schutzsuchenden im Zuge einer Machbarkeitsstudie aufbereitet und geokodiert worden.[4] Eine zentrale Rolle bei der Aufbereitung spielen die Geodaten von OpenStreetMap (OSM), anhand derer die unstrukturierten Angaben in den administrativen Daten abgeglichen, standardisiert und geokodiert werden. Damit liefert dieser Beitrag auch ein Beispiel dafür, wie Open Web Data effektiv in der amtlichen Statistik genutzt werden kann.

1 Personen mit deutscher und zusätzlich einer ausländischen Staatsangehörigkeit sind nicht im AZR registriert.

2 Durch das Datenaustauschverbesserungsgesetz 2016, das Zweite Datenaustauschverbesserungsgesetz 2019 sowie das Gesetz zur Weiterentwicklung des Ausländerzentralregisters 2021.

3 Ländercodes der International Organization for Standardization: www.iso.org

4 Schutzsuchende sind Ausländerinnen und Ausländer, die sich nach den Angaben im AZR aufenthaltsrechtlich unter Berufung auf humanitäre Gründe in Deutschland aufhalten (Eberle, 2019).

2

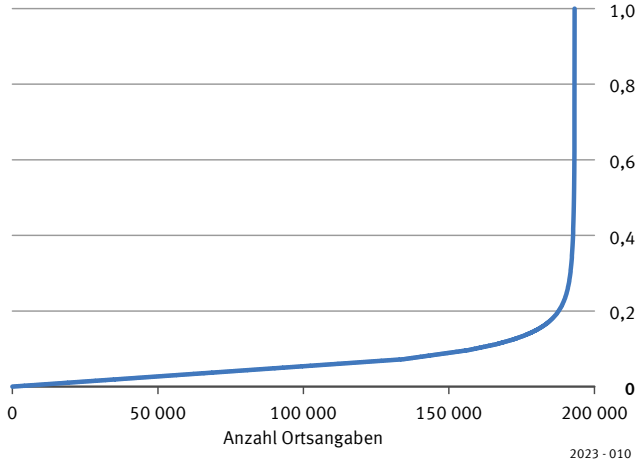
Daten zum Geburtsort von Schutzsuchenden aus dem AZR

Der Geburtsort ist standardmäßig in vielen internationalen Ausweisdokumenten eingetragen und Pflichtangabe bei der Registrierung im AZR. Die Erfassung erfolgt durch Ausländerbehörden, Erstaufnahmeeinrichtungen sowie Polizeibehörden im Zuge der Erstregistrierung von Ausländerinnen und Ausländern.

Ende 2020 waren 1 857 000 Schutzsuchende im AZR registriert. Insgesamt finden sich in den Daten 193 000 unterschiedliche Angaben zu Geburtsorten in 176 Staaten. Die Häufigkeit der Orte in den Daten folgt einer stark linksschiefen Verteilung. Knapp 70 % der Ortsangaben kommen genau einmal vor. Dies steht auch in Zusammenhang mit der unstrukturierten Erfassung der Geburtsorte. Die Herkunftsländer von Schutzsuchenden liegen zum großen Teil im arabischen Sprachraum und oft können Schutzsuchende bei ihrer Registrierung keine offiziellen Dokumente vorlegen. Daraus ergibt sich eine Vielzahl von unterschiedlichen Schreibweisen und Schreibfehlern in den Daten, die im Aufbereitungsprozess aufzulösen sind. Die 70 % einzigartigen Ortsangaben repräsentieren allerdings nur rund 7 % aller Schutzsuchenden. Gleichzeitig stehen die zehn häufigsten Ortsangaben für 22 % aller Schutzsuchenden. Fehler bei der Geokodierung dieser Ortsangaben fallen demnach besonders ins Gewicht. Positiv formuliert ist es möglich, einem Großteil der Schutzsuchenden den richtigen Geburtsort zuzuweisen, wenn nur ein geringer Anteil der Ortsangaben korrekt aufbereitet wird. ↘ Grafik 1

In einigen Fällen werden höhere administrative Verwaltungseinheiten, wie Provinzen oder Distrikte, als Geburtsorte erfasst. Diese ermöglichen immerhin eine Eingrenzung des Geburtsortes und werden daher weiterverarbeitet.[5] Darüber hinaus finden sich Angaben ohne verwertbaren Inhalt. Hierzu zählen Angaben von Staatennamen, ISO-Codes und eine Vielzahl unterschiedlicher Einträge, die darauf verweisen, dass der Geburtsort nicht bekannt ist. Diese nicht verwertbaren Angaben werden nicht weiter aufbereitet.

Grafik 1
Kumulative Häufigkeitsverteilung der Ortsangaben im Ausländerzentralregister vor der Aufbereitung
Anteil an den Schutzsuchenden in %



3

Geodaten von OpenStreetMap

Das OSM-Projekt sammelt seit Mitte der 2000er-Jahre weltweit raumbezogene Informationen (Geodaten) und stellt diese als Open Data unter der Open Database License zur Verfügung. Ebenso wie in der freien Enzyklopädie Wikipedia werden die Informationen nach dem Prinzip des Crowdsourcing gesammelt, das heißt die Geodaten werden von Freiwilligen und ohne kommerzielle Interessen gesammelt und laufend aktualisiert. Im Jahr 2021 zählte diese OSM-Community rund 1,6 Millionen aktive Editorinnen und Editoren und 7,8 Millionen registrierte Nutzerinnen und Nutzer. Editoren sammeln und validieren Geodaten aus öffentlich zugänglichen Datenquellen und durch manuelle Eingaben. Mit insgesamt 7 Milliarden Einträgen und rund 4,5 Millionen Aktualisierungen am Tag ist OSM eine dynamische Geodatenbank. Mitte 2022 zählte OSM weltweit rund 4,7 Millionen Einträge, die als bewohnte Siedlungen (englisch: populated settlements) annotiert waren.[6] ↘ Tabelle 1

5 In Staaten, in denen dies vermehrt vorkommt, sollte eine Auswertung nicht unterhalb der Ebene dieser Verwaltungseinheiten erfolgen (siehe Kapitel 5).

6 https://taginfo.openstreetmap.org

Tabelle 1
Siedlungsgebiete (populated settlements) in OpenStreetMap

Tag (Schlüssel)	Anzahl	Beschreibung
city (Hauptstadt/Großstadt)	10 978	Die größte städtische/urbane Siedlung innerhalb eines Verwaltungsgebiets.
town (Stadt/Kleinstadt)	106 369	Eine städtische Siedlung von lokaler bis regionaler Bedeutung mit üblicherweise mehr als 10 000 Einwohnerinnen und Einwohnern.
village (Dorf)	1 417 498	Eine kleine ländliche Siedlung mit wenigen zentralörtlichen Einrichtungen und mit weniger als 10 000 Einwohnerinnen und Einwohnern.
hamlet (Weiler)	1 700 548	Eine kleinere, ländliche Ansammlung von Haushalten, typischerweise mit weniger als hundert Einwohnerinnen und Einwohnern und minimaler Infrastruktur, meist gänzlich gewerbelos und ohne Anbindung öffentlichen Nahverkehrs.
Sonstige	1 486 399	
Insgesamt	4 721 792	

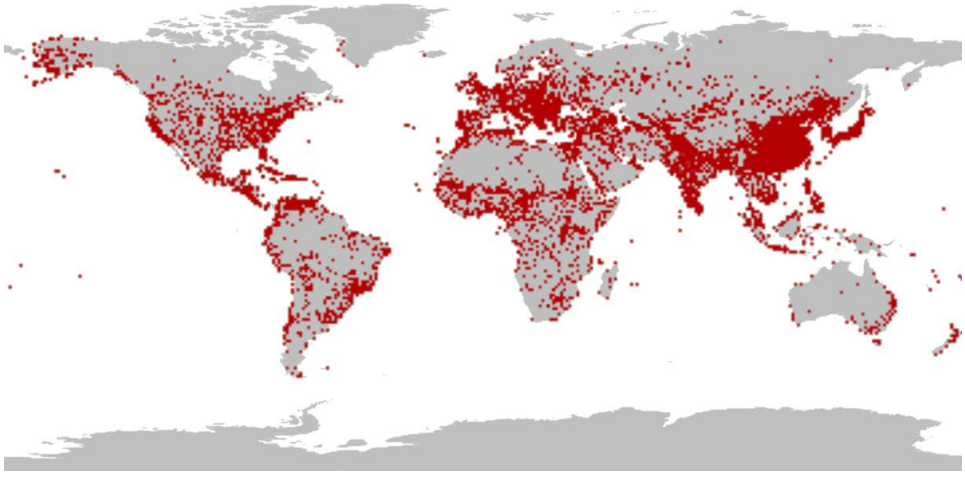
Quelle: wiki.openstreetmap.org/wiki/Key:place [Zugriff am 11. Januar 2023].

Die Datenqualität ist regional unterschiedlich und hängt stark davon ab, wie aktiv die jeweiligen Editoren sind. Für den Zeitraum 2013 bis 2019 identifizieren Seto und andere (2020) die aktivsten Communitys in Deutschland, in der Russischen Föderation und in den Vereinigten Staaten. Auch Herfort und andere (2021) attestieren, dass die meisten Einträge in entwickelten Ländern mit einem sehr hohen Human Development Index beigetragen werden. Nach dem verheerenden Erdbeben in Haiti im Jahr 2010 sind diese Unterschiede im Hinblick auf den Zugang zu Geodaten vermehrt in den Fokus gerückt. Seitdem schließt die OSM-Community, darunter zahlreiche Hilfsorganisationen wie Rotes Kreuz und Roter Halbmond, durch humanitäre Aktionen gezielt Lücken in den Geodaten in weniger entwickelten Regionen der Welt. Beispielsweise wurden mit dem Missing Maps Project seit 2014 rund 70 Millionen Geodaten vornehmlich in Subsahara-Afrika, Südostasien sowie Mittel- und Südamerika gesammelt. Herfort und andere (2021) berechneten, dass zwischen Januar 2008 und Mai 2020 rund 14 % aller in OSM hinzugefügten Gebäude auf humanitäre Mapping-Aktivitäten zurückzuführen sind. Die weltweite geografische Abdeckung am Beispiel der Annotationen Stadt und Dorf (englisch: city beziehungsweise village) zeigt ↘ Grafik 2.

Lizenzrechtlich ist jegliche Art der Nutzung von OSM-Daten zulässig, sofern die Nutzungsbedingungen der Open Database Licence eingehalten werden. Die Nutzungsbedingungen verlangen, dass durch die Nutzung entstandene abgeleitete Datensätze ebenso frei zugänglich als Open Data zur Verfügung gestellt werden. Ein abgeleiteter Datensatz entsteht, wenn OSM-Daten mit externen Daten weiterverarbeitet werden oder die gesamten OSM-Daten beziehungsweise ein substanzieller Teil der OSM-Datenbasis mit einer anderen Datenbasis zusammengeführt werden. Den OSM Licence Guidelines (2017) folgend entsteht durch Geokodierung, also durch die Ergänzung bestehender Datensätze mit Namen, Adressen und Geokoordinaten, kein abgeleiteter Datensatz, solange nicht systematisch OSM-Datenbestände in ihrer Gesamtheit extrahiert werden.

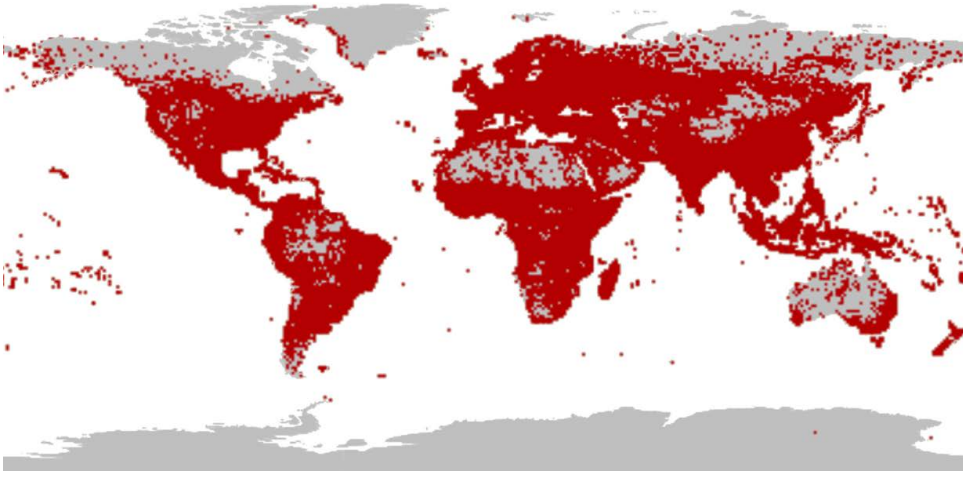
Grafik 2
Weltweite Abdeckung der OSM-Geodaten

OSM Tag Place=city (Stadt)



Quelle: https://taginfo.openstreetmap.org/tags/place=city#map

OSM Tag Place=village (Dorf)



Quelle: https://taginfo.openstreetmap.org/tags/place=village#map

2023 - 0011

4

Aufbereitung des Geburtsortes mit OSM-Geodaten

Die Aufbereitung der Angaben zu den Geburtsorten von Schutzsuchenden aus dem AZR anhand der Geodaten von OpenStreetMap erfolgt konzeptionell in drei Schritten: Potenzielle geografische Orte werden identifiziert, Mehrdeutigkeit aufgelöst und die Geodaten des zugewiesenen Ortes übernommen.

4.1 Potenzielle geografische Orte identifizieren

In einem ersten Schritt wird festgestellt, welche geografischen Orte mit einer Ortsangabe im AZR gemeint sein könnten. Dieser auch als Gazetteer Matching bezeichnete Vorgang (Goldberg und andere, 2007) ist

vergleichbar mit dem Suchen von Ortsnamen in einem Ortsverzeichnis (englisch: gazetteer). Bei der Auswahl potenzieller geografischer Orte treten unterschiedliche Probleme auf. Einerseits können mit unterschiedlichen textlichen Repräsentationen die gleichen geografischen Orte gemeint sein, zum Beispiel ‚Frankfurt a. M.' und ‚Frankfurt am Main' (Synonyme). Aufgrund der Erfassung als Freitext enthalten die Daten unterschiedliche Schreibweisen, unterschiedliche Sprachen und auch Schreibfehler. Synonyme Ortsangaben können mit Werkzeugen aus dem Bereich des Natural Language Processing identifiziert werden. Hierbei werden über textliche oder phonetische Distanzmaße unterschiedliche Schreibweisen und Schreibfehler erkannt.

Andererseits können mit einem Ortsnamen unterschiedliche geografische Orte gemeint sein, zum Beispiel könnte ‚Frankfurt' sowohl ‚Frankfurt am Main' als auch ‚Frankfurt an der Oder' betreffen (Homonyme). Um homonyme beziehungsweise mehrdeutige Ortsangaben aufzulösen, ist zunächst eine möglichst komplette Auswahl aller möglichen Bedeutungen zu erstellen. Grundlage ist ein weitgehend vollständiges Verzeichnis aller weltweit bewohnten Siedlungen. Als ein solches weltweites Ortsverzeichnis dient OSM.

Technisch erfolgt der Zugriff auf die OSM-Daten über eine Online-Programmierschnittstelle (englisch: Application Programming Interface, kurz API). APIs ermöglichen eine vollständige Automatisierung und damit die Verarbeitung großer Datenmengen. Die Ortsangaben aus dem AZR werden dabei in einer lokalen PostgreSQL-Datenbank gehalten. Über ein Python-Skript werden diese blockweise aus der Datenbank extrahiert und in einzelnen HTTP-Anfragen an die API geschickt. Die Antwort der API erfolgt in Form eines JSON-Dokuments, aus dem die potenziellen geografischen Orte extrahiert werden.[7]

APIs bieten einen komfortablen Datenzugang, allerdings bestehen unterschiedliche Zugangsbeschränkungen. Kostenlosen OSM-Datenzugang für die Geokodierung bietet beispielsweise Nominatim. Beschränkungen bestehen hier nur im Hinblick auf die Anzahl der Anfragen, die innerhalb einer gewissen Zeit gestellt werden können. Andere Anbieter bieten neben der Geokodierung weitere, zumeist kostenpflichtige Dienstleistungen. Für die Aufbereitung der Angaben zum Geburtsort von Schutzsuchenden nutzt das Statistische Bundesamt die API von LocationIQ. Dieser Anbieter stellt eine Autocomplete-Funktion zur Verfügung, die automatisch kleinere Schreibfehler und unterschiedliche Schreibweisen eines Ortes erkennt. So erkennt die API in den Originaldaten beispielsweise rund 50 unterschiedliche Repräsentationen der äthiopischen Hauptstadt Addis Abeba und ordnet diese korrekt dem gleichen geografischen Ort zu. Damit ersetzt die Autocomplete-Funktion aufwendiges nutzerseitiges Preprocessing.[8] ↘ Grafik 3

Das Preprocessing beschränkt sich daher im Wesentlichen darauf, Angaben ohne verwertbaren Inhalt (ISO-Codes, Staatennamen und unterschiedliche Repräsentationen für unbekannte oder fehlende Angaben) zu identifizieren, um diese von der weiteren Verarbeitung auszuschließen.

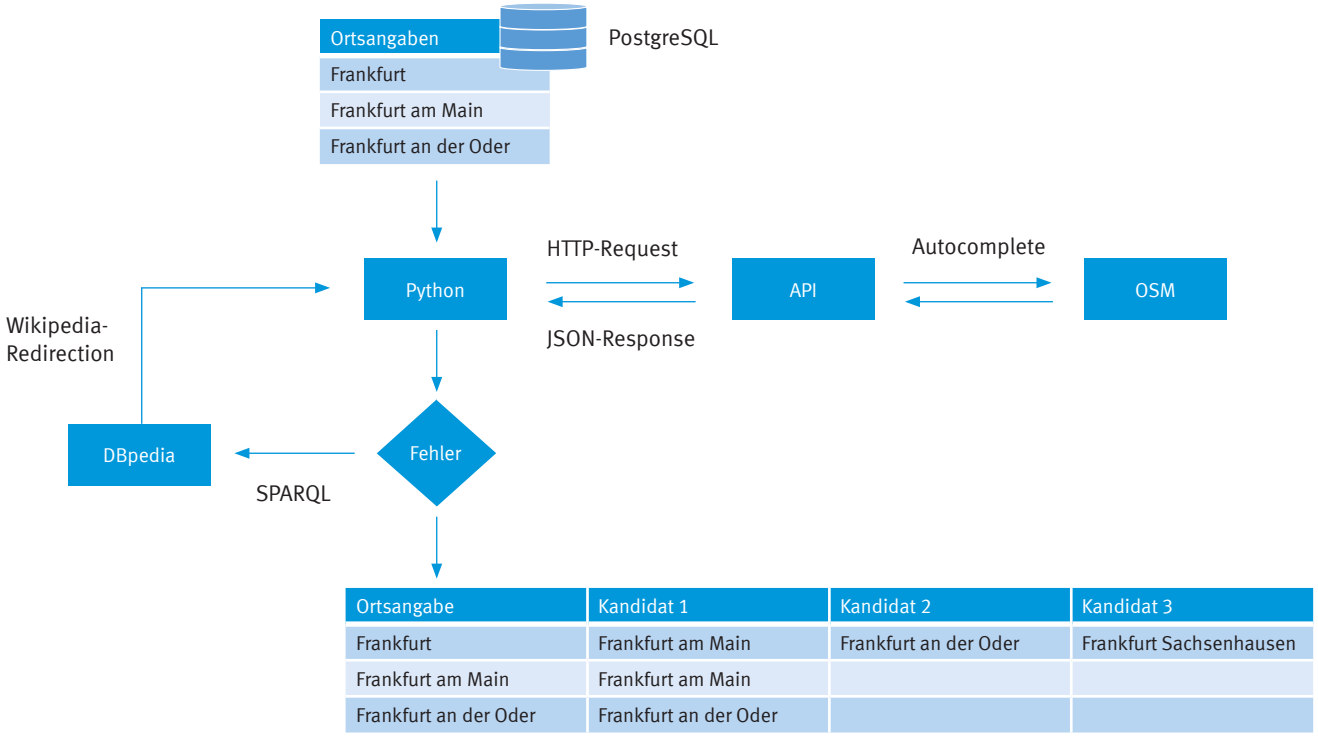
Findet die API für eine Geburtsortsangabe keinen passenden Eintrag in den OSM-Daten, erfolgen Korrekturversuche. Zunächst wird überprüft, ob der nicht auffindbare Ort umbenannt wurde oder der Ortsname in einer anderen Sprache zu einem Treffer führt. Hierfür werden Informationen aus Wikipedia genutzt. Beispielsweise lautete der offizielle Ortsname der ukrainischen Stadt Dnipro im Ukrainischen bis Mai 2016 ‚Dnipropetrovsk'. Ein großer Teil der Bevölkerung von Dnipro spricht russisch und würde als Geburtsort vermutlich ‚Dnepropetrovsk' angeben. In Wikipedia führen sowohl Suchanfragen nach ‚Dnipropetrovsk' als auch nach ‚Dnepropetrovsk' auf den Artikel des heutigen Dnipro. Ebenso verweisen Suchanfragen nach ‚Leningrad' auf das heutige Sankt Petersburg.

Die Prüfung, ob für eine Ortsbezeichnung in Wikipedia eine Weiterleitung besteht, erfolgt automatisiert über den SPARQL-Endpunkt von DBpedia. Das DBpedia-Projekt extrahiert regelmäßig die Informationen aus Wikipedia und stellt diese in strukturierter Form (RDF-Graph) als Linked Open Data zur Verfügung (Lehmann und andere, 2015). Über die Abfragesprache SPARQL können die Informationen zu Weiterleitungen von Suchanfragen automatisiert abgefragt werden. Führt auch dieser Korrekturschritt nicht zu einem Treffer in

7 Für jede Ortsangabe werden zehn Kandidaten extrahiert. Die Reihenfolge der Kandidaten legt ein in den OSM-Daten enthaltener Importance Score fest. Dieser Wert wird auf Basis des Wikipedia-Artikels eines Orts berechnet. Dabei wird Orten, auf deren Artikel innerhalb von Wikipedia häufiger verlinkt wird, ein höherer Wert zugeordnet.

8 Die API ermöglicht täglich 25 000 Anfragen unter Verwendung der Autocomplete-Funktion.

Grafik 3
Ablauf der Identifizierung potenzieller geografischer Orte



Ortsangabe	Kandidat 1	Kandidat 2	Kandidat 3
Frankfurt	Frankfurt am Main	Frankfurt an der Oder	Frankfurt Sachsenhausen
Frankfurt am Main	Frankfurt am Main		
Frankfurt an der Oder	Frankfurt an der Oder		

2023 - 0012

den OSM-Daten, erfolgen bei häufig vorkommenden Angaben manuelle Korrekturen.

4.2 Mehrdeutigkeit von Ortsangaben auflösen

Im zweiten Schritt werden mehrdeutige Ortsangaben identifiziert und anschließend einem geografischen Ort zugewiesen. In der Literatur wird dieser Schritt auch als Toponym Resolution bezeichnet (DeLozier und andere, 2015). Ausgangspunkt ist die Liste potenzieller geografischer Orte, die mit einem Ortsnamen (Toponym) gemeint sein könnten, aus dem vorherigen Aufbereitungsschritt. Im Zuge der Toponym Resolution wird für diese Kandidaten eine Wahrscheinlichkeitsverteilung gesucht, anhand derer im Anschluss eine stochastische Zuordnung erfolgen kann.

Mehrdeutige Ortsangaben identifizieren

Zunächst besteht die Herausforderung, mehrdeutige Ortsangaben automatisiert zu identifizieren. Hierfür erfolgt ein Vergleich der ursprünglichen Ortsangabe mit den Rückmeldungen (Kandidaten) aus den OSM-Daten. Eine Ortsangabe gilt dabei als mehrdeutig, wenn alle einzelnen Wörter, beginnend vom ersten Wort in der richtigen Reihenfolge, in mehreren Kandidaten vorkommt. Dahinter steht die Annahme, dass Namenszusätze, die die Mehrdeutigkeit auflösen, immer am Ende eines Ortsnamens angehängt werden. Weiterhin werden Kandidaten, die in OSM nicht als bewohnte Siedlungen (englisch: populated places) annotiert sind, ausgeschlossen. Dieses Vorgehen verhindert die Verwechslung von Städten mit gleichnamigen Distrikten oder Provinzen. Ebenso sollen Stadtteile, beispielsweise ‚Frankfurt Sachsenhausen', nicht mit mehrdeutigen Ortsangaben wie ‚Frankfurt am Main' und ‚Frankfurt an der Oder' verwechselt werden. Daher werden Kandidaten ausgeschlossen, die eine Distanz von weniger als zehn Kilometer zu einem bereits identifizierten Kandidaten aufweisen. Letztlich wird nicht von mehrdeutigen Ortsangaben ausgegan-

gen, wenn Ortsangabe und Kandidat auf dem ersten Rang, also der Kandidat mit dem höchsten OSM-Importance-Score, identisch sind.[9]

Mehrdeutige Ortsangaben auflösen

Durch die bereits aufbereiteten Angaben zum Geburtsstaat (Canan/Eberle, 2022) ist die potenzielle internationale Mehrdeutigkeit der Ortsangaben (Paris in Texas oder Paris in Frankreich) bereits aufgelöst. Bestehen bleibt die Mehrdeutigkeit von Ortsnamen innerhalb eines Staates. Für die in Deutschland mehrdeutige Ortsangabe Frankfurt ergibt sich beispielsweise folgende unbekannte Wahrscheinlichkeitsverteilung:

(1) P(Geburtsort = Frankfurt am Main | Ortsangabe = ‚Frankfurt') +
P(Geburtsort = Frankfurt an der Oder | Ortsangabe = ‚Frankfurt') = 1

Zur Auflösung der Mehrdeutigkeit von Ortsnamen innerhalb von Staaten gibt es unterschiedliche Strategien. Einen Überblick bietet Leidner (2008). Beispielsweise können die Wahrscheinlichkeiten (P) als Funktion (F) der Bevölkerungsgröße dargestellt werden.

(2) P(Geburtsort = Frankfurt am Main | Ortsangabe = ‚Frankfurt') +
P(Geburtsort = Frankfurt an der Oder | Ortsangabe = ‚Frankfurt')
= F(Bevölkerung Frankfurt am Main) +
F(Bevölkerung Frankfurt an der Oder)

Im Zusammenhang mit Geburtsorten von Schutzsuchenden ist diese Vorgehensweise aus verschiedenen Gründen nicht geeignet: Zum einen stehen für Geburtsorte von Schutzsuchenden oftmals keine aktuellen und verlässlichen Bevölkerungszahlen zur Verfügung. Zum anderen ist fraglich, ob die Größe der Gesamtbevölkerung eines Ortes auch die gesuchte Wahrscheinlichkeitsverteilung für Schutzsuchende in Deutschland approximiert.

Alternativ ist es möglich, die beobachteten Häufigkeitsverteilungen in den vorhandenen Daten zu nutzen. Dabei wird sich zunutze gemacht, dass die Mehrdeutigkeit von Ortsangaben durch identifizierende Namenszusätze, wie die zusätzliche Nennung des Bundes-

9 Bei der Ortsangabe ‚Berlin' wird beispielsweise unmittelbar von der gleichnamigen deutschen Hauptstadt ausgegangen, obwohl auch eine Ortschaft ‚Berlin Seedorf' in Schleswig-Holstein existiert.

Grafik 4
Schritte zur Auflösung der Mehrdeutigkeit von Ortsangaben

Ortsangabe	Kandidat 1	Kandidat 2	Kandidat 3
Frankfurt	Frankfurt am Main	Frankfurt an der Oder	Frankfurt Sachsenhausen
Frankfurt am Main	Frankfurt am Main		
Frankfurt an der Oder	Frankfurt an der Oder		

↓ mehrdeutige Ortsangaben und valide Kandidaten identifizieren

Ortsangabe	Kandidat 1	Kandidat 2
Frankfurt	Frankfurt am Main	Frankfurt an der Oder

↓ Wahrscheinlichkeitsverteilung aus beobachteten Daten ableiten

Ortsangabe	Kandidat 1	Anzahl Kandidat 1	Kandidat 2	Anzahl Kandidat 2
Frankfurt	Frankfurt am Main	90	Frankfurt an der Oder	10

↓ stochastische Zuordnung

Ortsangabe	Zufallszahl	Prob(Kandidat 1)	Prob(Kandidat 2)	Ort
Frankfurt	0,71	0,90	0,10	Frankfurt am Main

2023 - 0013

staates oder eines Flusses, aufgelöst werden kann. In großen Datensätzen ist davon auszugehen, dass für die meisten mehrdeutigen Ortsangaben auch ausreichend eindeutige Ortsangaben mit identifizierenden Zusätzen enthalten sind, um eine Wahrscheinlichkeitsverteilung über relative Häufigkeiten zu approximieren. ↘ Grafik 4

(3) P(Geburtsort = Frankfurt am Main | Ortsangabe = ‚Frankfurt') +
P(Geburtsort = Frankfurt an der Oder | Ortsangabe = ‚Frankfurt')
= h(Frankfurt am Main) + h(Frankfurt an der Oder)

4.3 Geokodierung

Sobald auch den mehrdeutigen Ortsangaben ein geografischer Ort zugewiesen wurde, können die Geokoordinaten aus OSM angespielt werden. Geokoordinaten können in unterschiedliche Koordinatensysteme eingebettet sein. OpenStreetMap nutzt das weitverbreitete World Geodetic System 1984 (WGS 84) als Grundlage für die Positionsangaben. Das WGS 84 teilt die Erde in 360 Längengrade von – 180 bis 180 Grad und 180 Breitengrade von – 90 bis 90 Grad. Die Geokoordinaten werden als Dezimalgrad angegeben. Der Standort Wiesbaden des Statistischen Bundesamtes wird in diesem Koordinatensystem beispielsweise mit Breitengrad 50,0708 und Längengrad 8,2520 verortet.

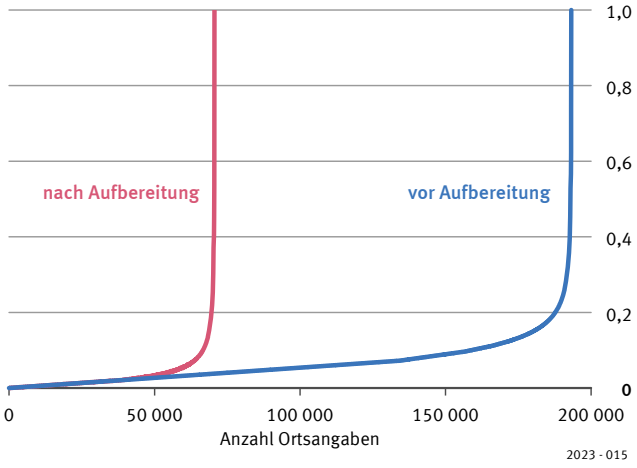
5

Qualitätsbewertung der aufbereiteten Daten

Vor der Aufbereitung befanden sich rund 193 000 unterschiedliche Angaben zu Geburtsorten in den AZR-Daten. Nach der Aufbereitung sind es noch knapp 71 000. Der Anteil an einzigartigen Ortsangaben sinkt von 70 % in den Originaldaten auf 55 % in den aufbereiteten Daten. Gleichzeitig ist der Anteil der zehn häufigsten Angaben von 22 auf 30 % gestiegen. ↘ Grafik 5

Die gestiegene Konzentration spiegelt einerseits die Auflösung vieler unterschiedlicher Schreibweisen in den Originalangaben wider. Andererseits ist in einigen Regionen eine systematische Verzerrung von ländlichen

Grafik 5
Kumulative Häufigkeitsverteilung der Ortsangaben im Ausländerzentralregister vor und nach der Aufbereitung
Anteil an den Schutzsuchenden in %



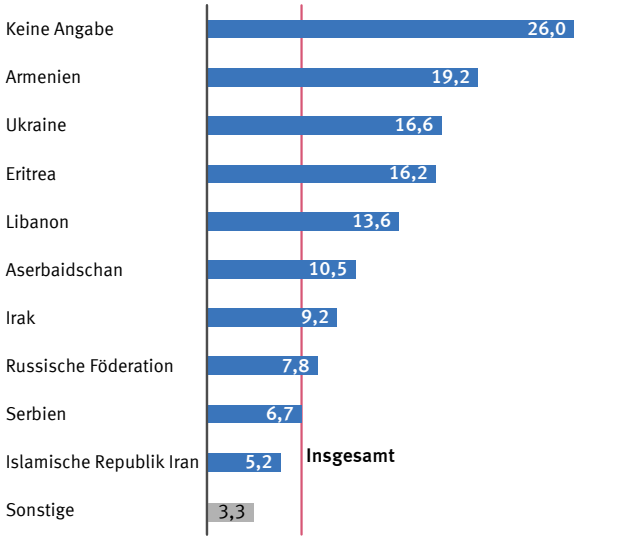
Gebieten in Richtung von Provinz- beziehungsweise Distrikthauptstädten zu beobachten. Ursache hierfür ist, dass in einigen Staaten administrative Verwaltungseinheiten die gleichen Namen wie deren Hauptstädte tragen. Syrien ist beispielsweise in 14 Gouvernements unterteilt, die alle den gleichen Namen tragen wie ihre Hauptstädte. Aleppo ist zum Beispiel gleichzeitig der Name des Gouvernements im Norden Syriens und der Name der Hauptstadt dieses Gouvernements. Auch in Afghanistan, dem Irak und der Türkei tragen die höchsten administrativen Gebietsgliederungen teilweise die gleichen Namen wie ihre Hauptstädte. Wurde neben dem Ortsnamen auch der Name der administrativen Verwaltungseinheit registriert, zum Beispiel ‚Afrin in Aleppo', wird teilweise fälschlich die Hauptstadt zugewiesen. Tragen administrative Verwaltungseinheiten die gleichen Namen wie ihre Hauptstädte, sollte die regionale Verteilung deshalb nicht unterhalb der Ebene dieser Verwaltungseinheiten ausgewertet werden.[10]

Insgesamt konnten 42 000 Ortsangaben keinem geografischen Ort zugeordnet werden. Die nicht zugeordneten Angaben repräsentieren 124 000 Schutzsuchende.

10 In diesen Fällen können den Ortsangaben die falschen Geokoordinaten zugeordnet sein – der Ortsangabe „Afrin in Aleppo" beispielsweise die Geokoordinaten der Stadt Aleppo. Neben den Geokoordinaten kann den OSM-Adressdaten aber auch eine Zuordnung zu höheren administrativen Ebenen entnommen werden. Bei einer Analyse auf Ebene der Gouvernements stimmt die Zuordnung wieder, da sowohl Afrin als auch Aleppo im Gouvernement Aleppo liegen.

Daraus ergibt sich ein Anteil fehlender Werte von 6,7 %. Höhere Anteile zeigen sich für die Geburtsstaaten, in denen nicht lateinische Alphabete genutzt werden. ↘ Grafik 6

Grafik 6
Fehlende Werte nach Geburtsstaaten nach der Aufbereitung
in %



2023 - 016

Eine weitere Qualitätseinschätzung liefert ein Vergleich der Ergebnisse mit einer Geokodierung deutscher Geburtsorte über den Geokodierungsdienst des Bundesamtes für Kartographie und Geodäsie (BKG; BKG Geocoder Version 1.1). Um eine möglichst große Vergleichsgruppe zu erhalten, werden für diesen Vergleich die Geburtsorte aller am 31. Dezember 2020 im AZR registrierten und in Deutschland geborenen Ausländerinnen und Ausländer herangezogen.

Für 95 % der rund 188 000 in Deutschland geborenen Schutzsuchenden ergeben beide Vorgehensweisen eine übereinstimmende geografische Ortsangabe.[11] Die Analyse der abweichend kodierten Zuweisungen zeigt, dass bei mehrdeutigen Ortsangaben unterschiedliche Zuweisungen erfolgen. Die meisten Abweichungen entstehen durch die mehrdeutige Ortsangabe ‚Oberhausen'. Bei dieser Angabe ergibt sich aus den OSM-Daten als wahrscheinlichster Kandidat die kreisfreie Großstadt Oberhausen in Nordrhein-Westfalen. Der Geokodierungsdienst des BKG hingegen entscheidet sich für die Gemeinde Oberhausen bei Kirn in Rheinland-Pfalz.

6

Ergebnisse

Die Ergebnisse stehen auch als interaktives Kartenangebot zur Verfügung. Die meisten Menschen, die sich Ende 2020 als Schutzsuchende in Deutschland aufhielten, kamen aus Syrien, genauer gesagt aus den Gouvernements Aleppo, Damaskus und Al-Hasaka.[12] Mit Nordrhein-Westfalen ist auch ein deutsches Bundesland unter den häufigsten Herkunftsregionen vertreten. Aufgrund der jungen Altersstruktur der 2015 und 2016 zugewanderten Schutzsuchenden werden seitdem vermehrt Kinder als Schutzsuchende in Deutschland geboren (Statistisches Bundesamt, 2021a). ↘ Tabelle 2

Tabelle 2
Häufigste Geburtsorte von Schutzsuchenden

Geburtsort	Geburtsstaat	Anzahl	Anteil in %
Aleppo	Syrien	131 260	7,1
Damaskus	Syrien	119 235	6,4
Al-Hasaka	Syrien	91 310	4,9
Nordrhein-Westfalen	Deutschland	53 615	2,9
Kabul	Afghanistan	36 600	2,0
Ninawa	Irak	36 260	2,0
Teheran	Iran	33 280	1,8
Homs	Syrien	28 325	1,5
Bagdad	Irak	26 770	1,4
Idlib	Syrien	26 745	1,4
Sonstige		1 278 491	68,6
Insgesamt		1 856 785	100

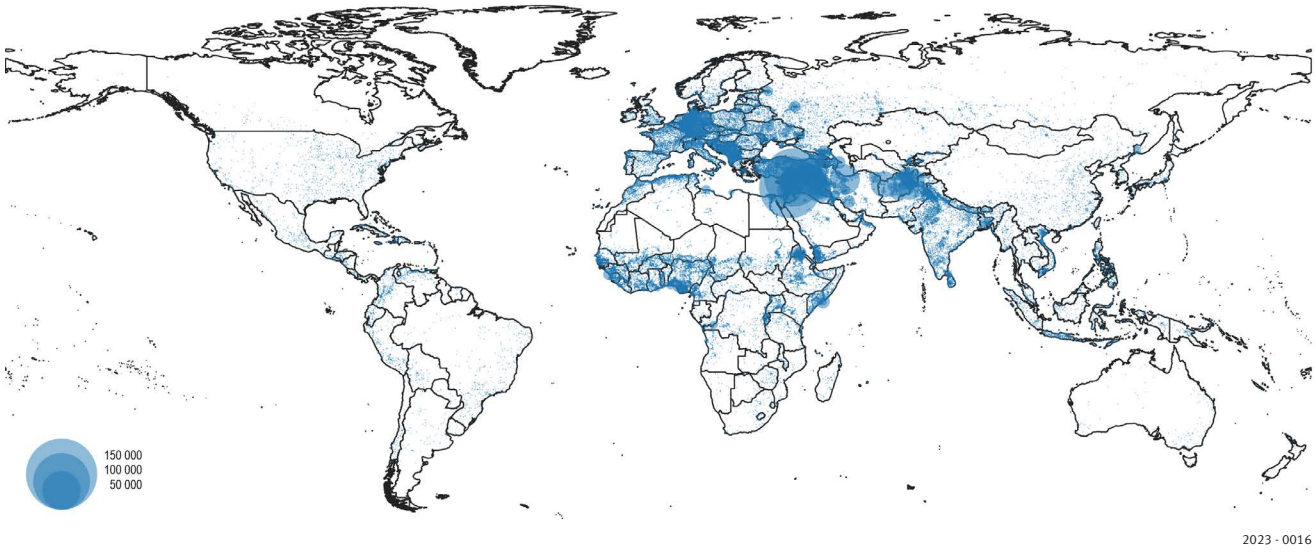
Auswertung auf der höchsten innerstaatlichen administrativen Ebene.

Die weltweite Verteilung der Geburtsorte von Schutzsuchenden ist Ende 2020 geprägt von der hohen Anzahl an Schutzsuchenden aus Syrien, Afghanistan und dem Irak. Dennoch sind auch weitere Herkunftsregionen wie die Balkanstaaten, Nordafrika und die Sahelzone sichtbar. Die Geburtsorte liegen aber auch in Deutsch-

11 Als Übereinstimmung gelten Fälle, deren Angaben zum Breitengrad nicht mehr als 0,045 Grad (5 km) und deren Angaben zum Längengrad nicht mehr als 0,065 Grad (zwischen 4,1 km und 4,9 km) abweichen.

12 Der Geburtsort ist nur ein Indikator für die Herkunft und muss nicht mit dem Ort übereinstimmen, an dem die Person aufgewachsen ist oder ihren Lebensmittelpunkt hat oder hatte.

Grafik 7
Weltweite Verteilung der Geburtsorte von Schutzsuchenden



sichtbar. Die Geburtsorte liegen aber auch in Deutschland und anderen EU-Mitgliedstaaten. Dies ist ein Beleg dafür, dass der Geburtsstaat für viele Schutzsuchende nicht mehr der Staat ist, aus dem die Eltern ursprünglich geflohen sind. ↘ Grafik 7

Die aufbereiteten Daten zu den Geburtsorten geben erstmals Einblicke in die regionale Herkunft dieser Menschen innerhalb einzelner Staaten. Beispielsweise kommt knapp die Hälfte aller in Syrien geborenen Schutzsuchenden aus dem nordsyrischen Gouvernement Aleppo (25 %) oder dem Gouvernement Damaskus (23 %). ↘ Grafik 8 auf Seite 38

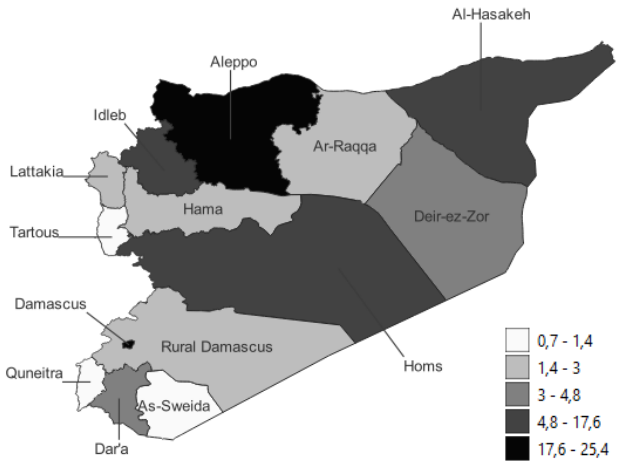
Die regionale Verteilung innerhalb der Staaten kann Hinweise auf unterschiedliche Fluchtursachen enthalten. Dies wird am Beispiel der Türkei deutlich. Nach dem gescheiterten Putschversuch im Juli 2016 folgte in der Türkei ein zweijähriger Ausnahmezustand, in dem Staatspräsident Erdoğan staatliche und gesellschaftliche Institutionen ohne parlamentarische Beteiligung gezielt umstrukturierte (Gieler, 2021). Im Jahr 2017 wurden die Befugnisse des Staatspräsidenten per Verfassungsreferendum mit knapper Mehrheit auch langfristig erheblich erweitert. Seit 2016 steigt die Zahl türkischer Schutzsuchender in Deutschland stetig an, nachdem sie zuvor rückläufig war. Der umstrittene und polarisierende Umbau von Politik, Verwaltung und Gesellschaft spiegelt sich auch in der regionalen Herkunft von türkischen Schutzsuchenden in Deutschland wider. Schutzsuchende, die vor 2016 nach Deutschland kamen, sind großteils in den kurdisch geprägten Gebieten im Südosten der Türkei geboren. Seit 2016 kommen türkische Schutzsuchende hingegen häufiger aus Istanbul, Ankara und Izmir, und damit aus nordwestlichen Provinzen, die 2017 mehrheitlich gegen die Verfassungsänderungen gestimmt hatten. ↘ Grafik 9 auf Seite 39

Angaben zu Geburtsorten ermöglichen darüber hinaus, Kettenmigration, Migrationsnetzwerke und Cluster in der regionalen Verteilung der ausländischen Bevölkerung detaillierter zu analysieren. Beispielsweise liefern die Daten Hinweise dafür, dass Schutzsuchende aus unterschiedlichen türkischen Provinzen unterschiedliche Präferenzen im Hinblick auf ihren Wohnort in Deutschland haben. So lebte mit 4,7 % der größte Teil der Schutzsuchenden aus Istanbul Ende 2020 in Berlin. Türkische Schutzsuchende aus der kurdisch geprägten Provinz Mardin hingegen waren zumeist in Bremen registriert (4,9 %). ↘ Grafik 10 auf Seite 40

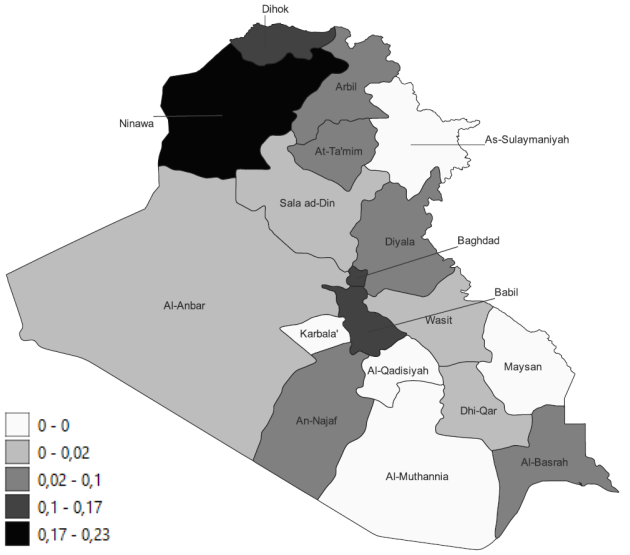
Grafik 8

Verteilung der Geburtsorte in Syrien, dem Irak und Afghanistan

Syrien
Regionale Ebene: 14 Gouvernements

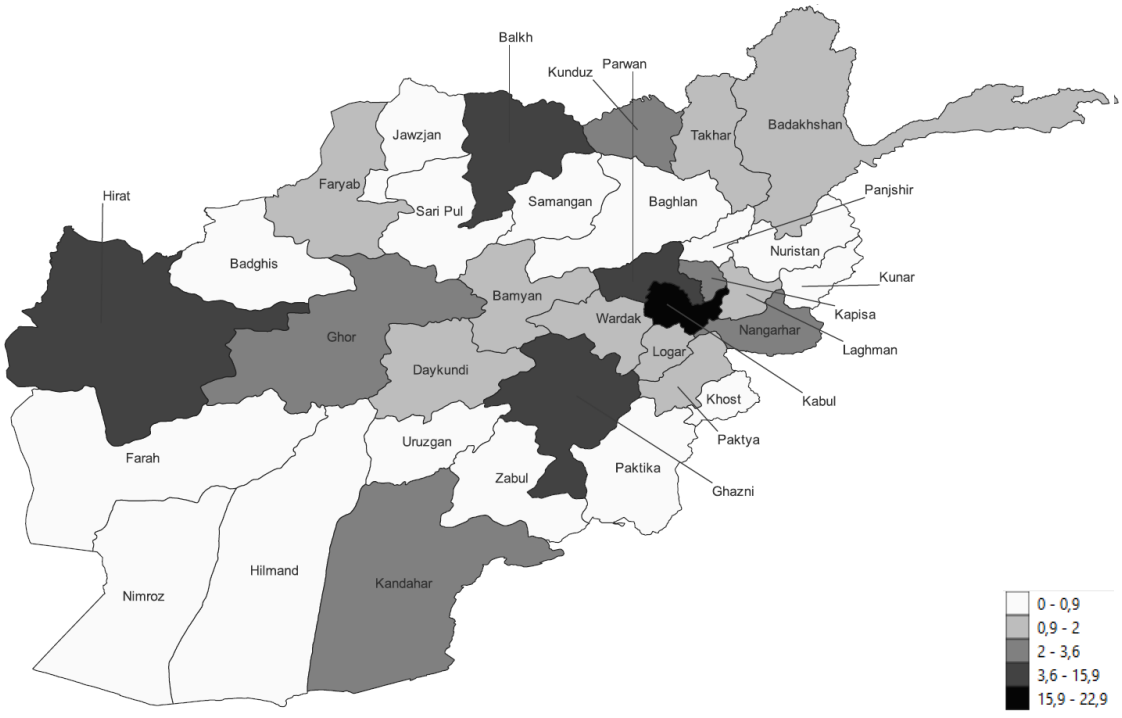


Irak
Regionale Ebene: 18 Gouvernements[1]



1 5 900 Angaben zum Geburtsort autonome Region Kurdistan wurden anteilsmäßig auf die Gouvernements Erbil, Dahuk und Halabdscha aufgeteilt.

Afghanistan
Regionale Ebene: 34 Distrikte



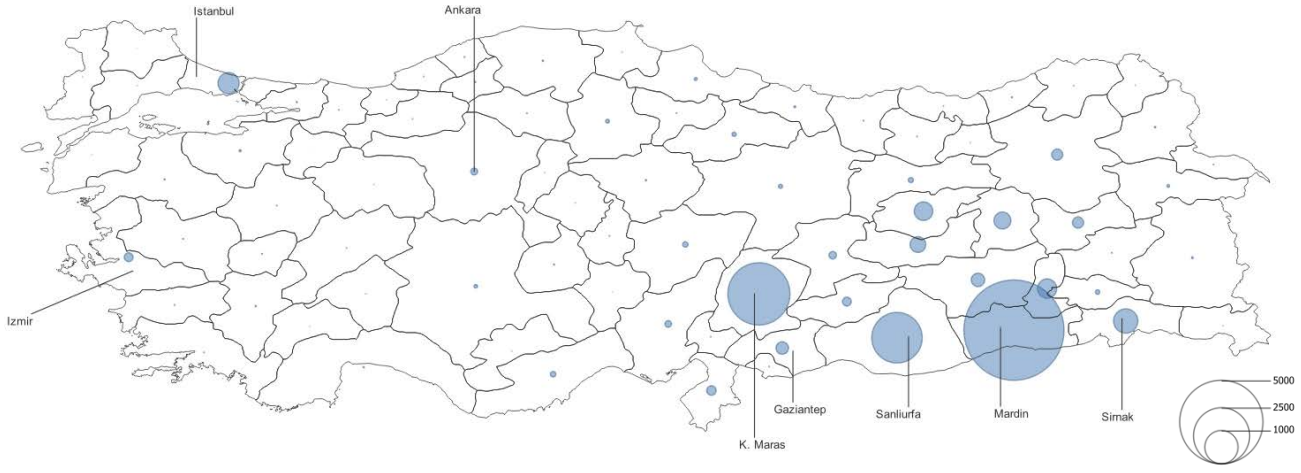
2023 - 0017

Grafik 9

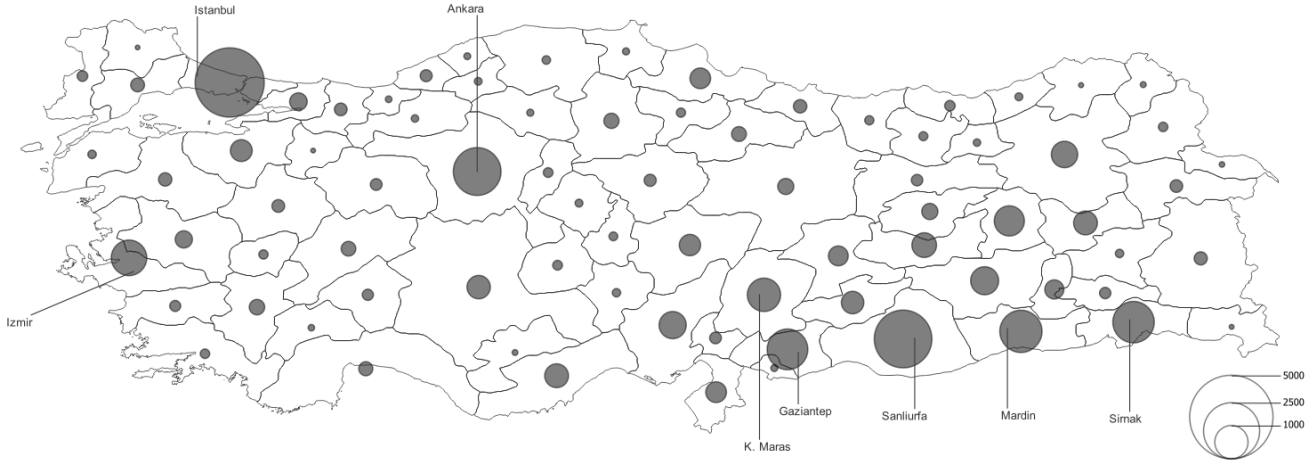
Regionale Verteilung von in der Türkei geborenen Schutzsuchenden

Regionale Ebene: 81 Provinzen

mit Ersteinreise vor 2016

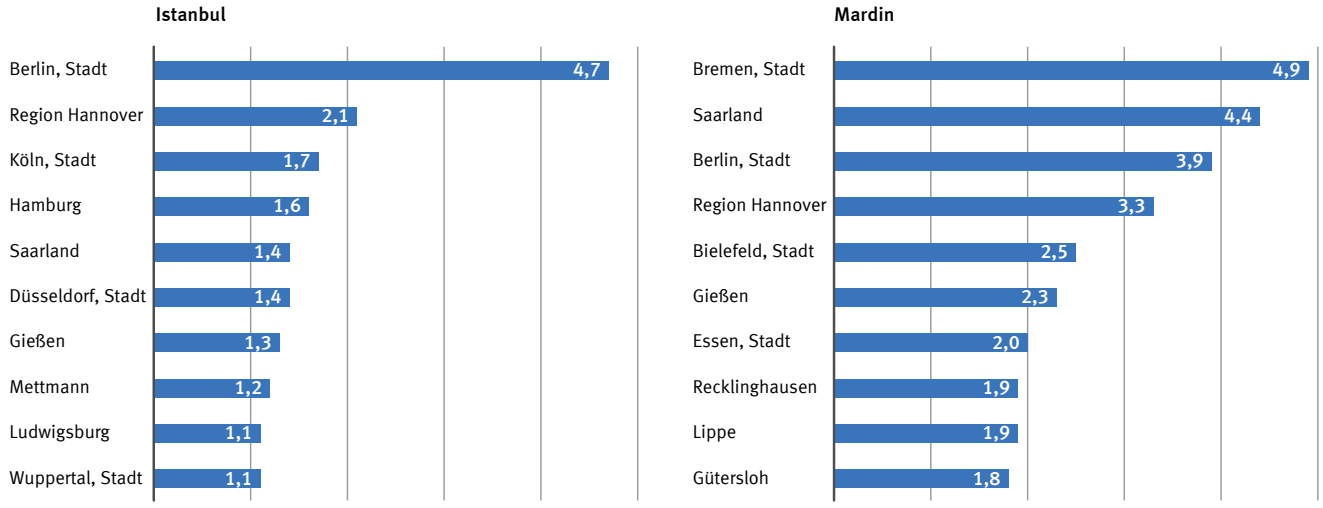


mit Ersteinreise seit 2016



2023 - 0018

Grafik 10
Regionale Verteilung von türkischen Schutzsuchenden aus Istanbul und Mardin nach Wohnorten in Deutschland Ende 2020
in %



2023 - 020

7

Fazit

Informationen zum Geburtsort im AZR bieten vielfältige neue Auswertungsmöglichkeiten in der Ausländerstatistik und der Statistik über Schutzsuchende. Analysen zur Herkunft können damit in einem kleinräumigeren regionaleren Kontext innerhalb der Staatengrenzen durchgeführt werden. Über die Geokoordinaten können die AZR-Daten beispielsweise mit Klimadaten kombiniert werden und Beiträge zur Klimafolgenforschung leisten.

Voraussetzung für die statistische Nutzung ist eine sorgfältige Aufbereitung und Qualitätssicherung der unstrukturiert als Freitext erhobenen Angaben. Im Zuge der Aufbereitung werden unterschiedliche Repräsentationen der gleichen geografischen Orte (synonyme Ortsbezeichnungen) und gleiche Repräsentationen unterschiedlicher geografischer Orte (homonyme beziehungsweise mehrdeutige Ortsbezeichnungen) aufgelöst. Von zentraler Bedeutung für den Aufbereitungsprozess ist eine möglichst vollständige Vergleichsdatenbasis. Mit rund 4,7 Millionen Einträgen zu bewohnten Siedlungen bietet OpenStreetMap eine der größten weltweiten Geodatensammlungen und stellt diese unter der Open Database License zur freien Verfügung.

Getestet wurde der Aufbereitungsprozess anhand der Angaben zu Geburtsorten von knapp 1,9 Millionen Schutzsuchenden, die am 31. Dezember 2020 im AZR registriert waren. Dieser Anwendungsfall entspricht einem Härtetest, da die Hauptherkunftsländer im arabischen Sprachraum liegen und bei der Registrierung von Schutzsuchenden offizielle Dokumente häufig fehlen. Nach der Aufbereitung verblieb mit 6,7 % dennoch nur ein geringer Anteil an Angaben, die keinem geografischen Ort zugewiesen werden konnten. Erwartungsgemäß sind die Anteile an nicht zugeordneten Ortsangaben in Staaten höher, in denen ein nicht lateinisches Alphabet genutzt wird. Dabei enthalten die Angaben, die keinem geografischen Ort zugeordnet werden konnten, zumeist auch keinen verwertbaren Hinweis. Oftmals handelt es sich hierbei um ISO-Codes oder Angaben ohne jeden semantischen Inhalt. Probleme entstehen dort, wo administrative Verwaltungseinheiten (Regionen, Provinzen, Distrikte) den gleichen Namen tragen wie ihre Hauptstädte. Hier sollte die regionale Verteilung nicht unterhalb der Ebene dieser administrativen Einheiten ausgewertet werden.

Unter den genannten Einschränkungen zeigt die Machbarkeitsstudie: Die administrativen Daten zu Geburtsorten können mithilfe der OSM-Geodaten aufbereitet und für Analysen genutzt werden.

LITERATURVERZEICHNIS

Canan, Coşkun/Eberle, Jan. *Geburtsstaat und Geburtsort im Ausländerzentralregister – Nutzungsmöglichkeiten für die amtliche Statistik.* In: WISTA Wirtschaft und Statistik. Ausgabe 2/2022, Seite 53 ff.

Conti, Cinzia/Cimbelli, Alessandro. *Exploring international migration at subnational scale: the Italian case.* In: UNECE Session on Migration Statistics Working Paper. 2018. [Zugriff am 15. Dezember 2022]. Verfügbar unter: unece.org

DeLozier, Grant/Baldridge, Jason/London, Loretta. *Gazetteer-independent toponym resolution using geographic word profiles.* In: Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015. Seite 2382 ff.

Eberle, Jan. *Schutzsuchende*. In: WISTA Wirtschaft und Statistik. Ausgabe 1/2019, Seite 19 ff.

Fleischer, Henning. *Entwicklung der Ausländerzahl seit 1987*. In: Wirtschaft und Statistik. Ausgabe 9/1989, Seite 594 ff.

Gieler, Wolfgang. *Politische Entwicklungen in der Türkei seit dem Putschversuch*. In: Henrich, Christian Johannes/Gieler, Wolfgang (Herausgeber). Die Außenpolitik der Türkei im Mittleren Osten. Wiesbaden 2021, Seite 27 ff.

Goldberg, Daniel W./Wilson, John P./Knoblock, Craig A. *From text to geographic coordinates: the current state of geocoding*. In: URISA Journal. Jahrgang 19. Ausgabe1/2007, Seite 33 ff.

Herfort, Benjamin/Lautenbach, Sven/Porto de Albuquerque, João/Anderson, Jennings/Zipf, Alexander. *The evolution of humanitarian mapping within the OpenStreetMap community*. In: Scientific Reports. Jahrgang 11. Artikel 3037/2021.

Lehmann, Jens/Isele, Robert/Jakob, Max/Jentzsch, Anja und weitere. *DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia*. In: Semantic web 6.2.2014. Seite :167 ff. DOI: 10.3233/SW-140134

Leidner, Jochen L. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. 2007.

OpenStreetMap Foundation. *Licence Community Guideline Geocoding*. 2017. [Zugriff am 15. Dezember 2022]. Verfügbar unter: wiki.osmfoundation.org

Seto, Toshikazu/Kanasugi, Hiroshi/Nishimura, Yuichiro. *Quality verification of volunteered geographic information using osm notes data in a global context*. In: International Journal of Geo-Information. Jahrgang 9. Ausgabe 6/2020, Seite 372 ff.

Šimbera, Jan/Drbohlav, Dušan/Štych, Přemysl. *Geocoding Freeform Placenames: An Example of Deciphering the Czech National Immigration Database*. In: ISPRS International Journal of Geo-Information. Ausgabe 10.5/2021, Seite 335 ff.

Statistisches Bundesamt. *Digitale Agenda des Statistischen Bundesamtes. Version 2.1. 03/2019*. [Zugriff am 15. Dezember 2022]. Verfügbar unter: www.destatis.de

LITERATURVERZEICHNIS

Statistisches Bundesamt. *Seit 2015 werden mehr Schutzsuchende in Deutschland geboren*. Pressemitteilung Nr. N 39 vom 17. Juni 2021. 2021a. [Zugriff am 15. Dezember 2022]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Qualitätsbericht Ausländerstatistik*. 2021b. [Zugriff am 15. Dezember 2022]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Methodische Hinweise zur Ausländerstatistik*. 2021c. [Zugriff am 15. Dezember 2022]. Verfügbar unter: www.destatis.de

Statistisches Bundesamt. *Staats- und Gebietssystematik*. 2021d. [Zugriff am 15. Dezember 2022]. Verfügbar unter: www.destatis.de

RECHTSGRUNDLAGEN

Gesetz über das Ausländerzentralregister (AZR-Gesetz) vom 2. September 1994 (BGBl. I Seite 2265), das zuletzt durch Artikel 5c des Gesetzes vom 23. Mai 2022 (BGBl. I Seite 760) geändert worden ist.

Gesetz zur Verbesserung der Registrierung und des Datenaustausches zu aufenthalts- und asylrechtlichen Zwecken (Datenaustauschverbesserungsgesetz) vom 2. Februar 2016 (BGBl. I Seite 130).

Gesetz zur Weiterentwicklung des Ausländerzentralregisters vom 9. Juli 2021 (BGBl. I Seite 2467).

Zweites Gesetz zur Verbesserung der Registrierung und des Datenaustausches zu aufenthalts- und asylrechtlichen Zwecken (Zweites Datenaustauschverbesserungsgesetz – 2. DAVG) vom 4. August 2019 (BGBl. I Seite 1131).

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2023, Seite 27 und Seite 36 aktualisiert am 13. Juli 2023
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-23001-4, ISSN 1619-2907