COUNTRY AND PLACE OF BIRTH IN THE CENTRAL REGISTER OF FOREIGNERS – POSSIBILITIES FOR USE BY OFFICIAL STATISTICS

Coşkun Canan, Jan Eberle

↘ ***Keywords:*** *data processing – administrative data – country of birth – migration – Central Register of Foreigners – place of birth*

ABSTRACT

Official statistics use the variables "country of birth" and "place of birth" from the Central Register of Foreigners. In the administrative context, however, the variables are collected incompletely (country of birth) or unstructured as free text (place of birth) and without plausibility checks.

The Federal Statistical Office develops concepts for the standardisation, validation and supplementation of administrative data, so that these can be used for statistical purposes. The article focuses on the processing of the country of birth data from the Central Register of Foreigners for reference day 31 December 2020 and on first exemplary analyses based on that. Also, a concept for the potential processing and geocoding of the place of birth data is presented

Dr. Coşkun Canan

holds a doctorate in sociology from the Humboldt-Universität zu Berlin. He is a member of academic staff in the "Demographic Evaluations and Analyses Based on Microcensus Data" section of the Federal Statistical Office and is primarily concerned with redefining the concept of migration background.

Jan Eberle

is a graduate economist and assistant head of section in the "Intercensal Population Updates, Statistics of Foreigners and Integration" section of the Federal Statistical Office. His work focuses on evaluations of the Central Register of Foreigners with regard to foreigners and persons seeking protection.

1

Introduction

The increased use of administrative data and the resulting reduction in the burden on respondents is a core objective of the Federal Statistical Office's Digital Agenda (Statistisches Bundesamt, 2019). There is a long tradition of using the Central Register of Foreigners (Ausländerzentralregister – AZR) in population statistics for the compilation of central statistics of foreigners on the basis of Section 23 of the Central Register of Foreigners Act (AZR Act). Data from the Central Register of Foreigners have been processed ever since the 1970s in order to obtain information on the foreign population (Fleischer, 1989).

The AZR remains one of the largest administrative registers in Germany at present (March 2022). It contains information on all foreign citizens[1] who are not merely temporary visitors in Germany, i.e. who are usually staying for longer than three months. The AZR brings together information of all local foreigners authorities of all local foreigners authorities and thus serves as a central information platform for a large number of authorities entrusted with administrative tasks relating to asylum seekers and foreigners.

The Federal Statistical Office is as usually only one of many users of administrative data; statistical usability is therefore only one of many criteria which the data must fulfil. Accordingly, the quality of the raw data also differs from that of data from primary surveys. In many cases, careful quality checking and processing is therefore required before the data can be used for statistical purposes.

The AZR has been continuously developed since the migration of refugees seen in 2015 and 2016. In 2019, the legislature extended the possibilities for using the data within official statistics by passing the Second Data Sharing Improvement Act. Since then, the Federal Statistical Office has also received the information on the place and country of birth of foreign citizens for its statistics of foreigners.

Although the AZR contains entries on the place of birth, which is mandatory information, for almost all foreigners, on the place of birth as mandatory information in almost all cases, it is recorded in an unstructured manner in the form of free text entries. Information on the country of birth, on the other hand, is recorded in standardised form as an ISO code[2], however many observations are missing. Both variables are collected for administrative purposes without plausibility checks.

Chapter 2 below provides information on the role of the country and place of birth variables within population statistics; Chapter 3 examines the quality of the data available in the AZR for this purpose. Chapters 4 and 5 describe the processing of the country of birth data from the AZR, and the first evaluations based on data at the end of 2020. A possible solution for processing and geocoding the place of birth variable is outlined in Chapter 6. The article ends with a conclusion.

2

Significance of the country and place of birth variables in population statistics

The country of birth variable increases the potential for analysis in the statistics of foreigners and on persons seeking protection, both of which are based on the Central Register of Foreigners. A person's origin could previously only be analysed in terms of citizenship and a binary indicator of whether or not a person was born in Germany. The country of birth variable now makes it possible to analyse the foreign population in a more differentiated way based on the person's country of origin, in addition to his or her citizenship.

The country of birth is a permanent indicator for immigration. A person's citizenship, on the other hand, can change through naturalisation. Likewise, the country of birth variable allows persons to be assigned unambiguously to a country of origin. Problems such as those arising from the assignment of persons of multiple citizenships to a single country do not arise here. For these reasons, country of birth is also one of the standard demographic variables in international migration and population statistics. In Germany, up to now it

1 Persons with German citizenship who also have foreign citizenship are not recorded in the AZR.

2 The International Organization for Standardization (ISO) lists the country codes in standard 3166-1.

has only been possible to analyse the country of birth on the basis of the survey data from the microcensus sample. This is subject to certain limitations however, such as smaller numbers of cases with regard to small-area analyses and level of detail. In other official population statistics, efforts are therefore also being made to extend the analysis possibilities of the country of birth variable, for example with the aid of machine learning methods (Feuerhake and others, 2020).

The place of birth information plays a double role for the analysis potential of the statistics of foreigners and on persons seeking protection. On the one hand, the place of birth is of great importance for the quality checking and processing of the information on the country of birth. It is mandatory to enter a person's place of birth when registering a person in the AZR. It is entered by default in many international identity documents. Information on the country of birth, by contrast, is often missing from official identity documents. There is also a risk the entry could be confused with a person's citizenship or place of residence. In many cases, a place of birth entry can be used to check the plausibility of existing information on the country of birth, or to supplement missing information.

On the other hand, the place of birth itself represents an variable of interest which is needed to answer important questions such as: From which regions of India do most skilled workers come to Germany? Are more persons seeking protection coming from regions that are worse affected by climate change? The place of birth information can be used as a more detailed (proxy) indicator of origin below the national border level in order to answer these and similar questions.

3

Data quality

The country of birth is recorded as a three-digit alphanumeric ISO code (ISO 3166-1 alpha-3), for example "DEU" for Germany, or "FRA" for France. It is not mandatory to enter a country of birth in the AZR. Rather, the country of birth is often included as part of the compulsory information required for the place of birth. The place of birth is a free text entry based on official documents, where these are available. Otherwise, the entry is made on the basis of oral testimony.

At 33.2 %, there was a high proportion of missing country of birth entries on the reference day 31 December 2020. The place of birth, by contrast, was missing in only 0.4 % of cases. The entries of roughly 7.6 million foreign citizens (66.6 %) registered in the Central Register of Foreigners contain information on both the place and the country of birth. However, the existing entries may contain significant typing and linking errors. These include spelling errors as well as different designations of the same place (for example, "Paris" and "Paris, France"). In addition, places of birth can be linked to incorrect countries of birth (for example, "Paris" linked to "Germany" as country of birth). Quality assurance measures therefore need to be taken here. The entries of just under 3.8 million registered foreign citizens (33.1 %) contain a place but not a country of birth. Accordingly, processing the data in order to assign the names of previously missing countries to the given place of birth holds great potential for statistics. ↘ Table 1

Table 1
Available and missing information on place and country of birth

	Country of birth given		
	yes	no	total
	Number		
Place of birth given	7,639,070	3,793,385	11,432,460
yes	7,613,225	3,779,090	11,392,310
no	25,850	14,300	40,145
	%		
Place of birth given [1]	66.8	33.2	100
yes	66.6	33.1	99.6
no	0.2	0.1	0.4

1 Entries that are evidently without semantic content (e.g. "unknown") were not regarded as entries.

A comparative set of verified links must be consulted in order to identify any incorrect place and country of birth links. This makes it possible not only to check the quality of the country of birth entry in the AZR but also to determine the country of birth if there is no entry for it.

Similar problems and objectives exist in various sets of official statistics (Feuerhake and others, 2020). The Federal Statistical Office has therefore drawn up a central master file (place file) consisting of a comparative set of worldwide place names and the countries to which they are assigned. This place file consists of different reference files. The basis is the gazetteer of the last census. This contains the information on the places and countries of birth of the entire population collected in 2011 from the population registers. This information is supplemented by details from the List of Municipalities of the statistical offices of the Federation and the Länder, and by places of birth from the natural population statistics (births and deaths). The names of all municipalities in European countries are taken from the list of local administrative units maintained by the Statistical Office of the European Union, Eurostat. A historical index of countries covers former spellings of place names. Finally, unknown or missing information in the master file is completed by a catalogue of variant spellings, including incorrect versions, and synonyms.

The success of the project to ensure the quality of the information on places and countries of birth in the statistics of foreigners and in other population statistics depends crucially on the completeness and quality of the place file. Listing a total of 2.2 million combinations, the place file contains an extensive list of places and countries of birth. Quality controls and improvements, both automated and manual, were carried out to ensure the high quality of the place file. For example, web scraping was used to identify potentially incorrect links, and to correct them manually (Feuerhake and others, 2020). It is currently being assessed whether the data from OpenStreetMap can further improve the quality of the place file.

However, the place file may still contain incorrect links or omissions of place names or alternative spellings. The place file is therefore subject to an ongoing quality assurance process which will probably result in further improvements in quality and completeness in the future, especially following the census in 2022.

By comparing the data with the information in the place file, an initial assessment can be made of the quality of the place and country of birth information in the AZR. The AZR contains both a place entry and a country entry for around 7.6 million registered foreign citizens (see Table 1). These combinations can be compared against the information in the place file. A corresponding place-country of birth-link can be found in the place file in roughly 92.7 % of cases. There is no correspondence in the place file for the remaining 7.3 %.

The cases with no corresponding entry in the place file show that German places of birth are often wrongly linked to foreign countries in the AZR. This could be due to confusion between a person's country of birth and citizenship at the time of registration. Furthermore, some deviations from the place file in the AZR are caused by places of birth being assigned to countries that no longer exist or to which they no longer belong (for example, in former Yugoslavia). Ultimately, there are no restrictions on what is entered in these free-text fields. In some cases this can result in place of birth information that is

Overview 1

Quality evaluation of the place and country of birth variables

Quality criterion	Place of birth	Country of birth
Mandatory information	Yes	No
Completeness	Complete (0.4 % missing)	Incomplete (33.2 % missing)
Standardisation	Unstructured entries as free text without plausibility check	Structured entries as ISO 3166 ALPHA-3 code
Correctness of content	Some entries with no semantic content, with spelling mistakes, alternative spellings or different languages	Usually correct three digit ISO codes but in some cases two digit ISO codes were entered
Unambiguity	Problematic because of ambiguous place names, for example "Neustadt"	Yes, unique ISO codes
Consistency of combinations	Inconsistent combinations (7.3 %): › German places linked to foreign country, › Spelling not included in place file, › Assignments to former countries	

correct but the specific spelling or language is currently unknown to the place file. The quality evaluation of the place and country of birth information in the AZR can be found in ↘ Overview 1.

4

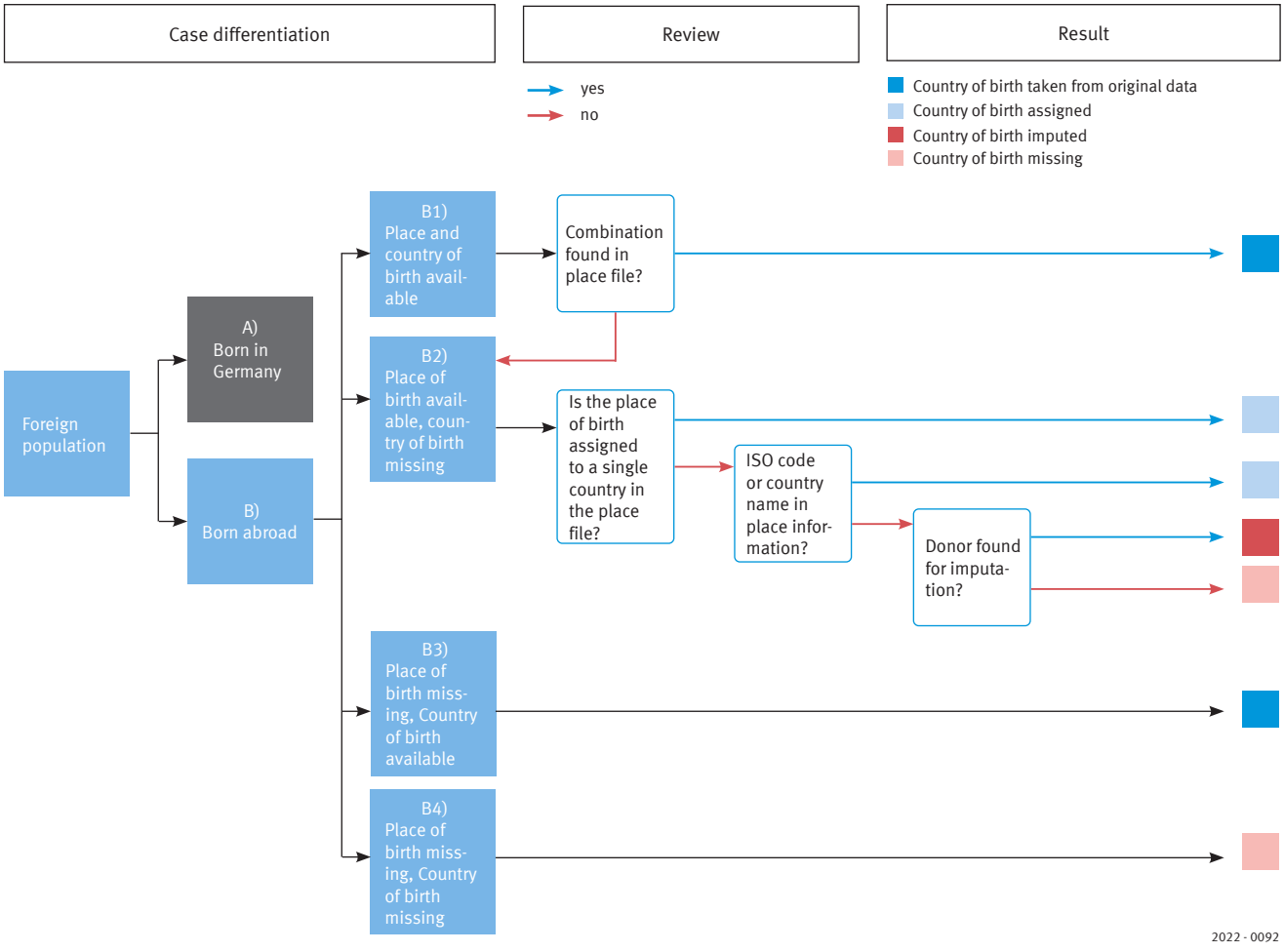
Processing of country of birth data

With the help of suitable data linkage and plausibility methods, it is possible to produce reliable data from the information on the place and country of birth of foreign citizens provided in the Central Register of Foreigners. The quality of the processed data depends largely on the completeness and quality of the information in the place file. With a complete and correct place file as a master file, the new variables can be integrated into the data processing of the statistics of foreigners. The place file is therefore also instrumental in processing the data.

A standardisation process is first carried out in order to facilitate comparison with the place file. In this process, the given ISO country code is converted to the code used in the Classification of States and Territories in official statistics (Statistisches Bundesamt, 2022). The free text information on the place of birth is standardised[3] and any entries with no semantic content are rejected. In the subsequent comparison with the place file, only place-country combinations that are contained in the place file

3 For example, UTF-8 special characters are converted into Latin1-compatible characters and common abbreviations (such as USA) are resolved.

Figure 1
Data processing procedure for the place and country of birth variables from the Central Register of Foreigners



are deemed plausible. If a different country is assigned to a place in the place file, the assignment is corrected in the AZR. "Minimally invasive" processing ensures that country of birth imputations are limited to cases in which the place entry in the place file is assigned to more than one country and no hint about the actual country of birth is given in the place of birth entry. The data processing procedure is described in more detail below and is shown in ↘ Figure 1.

The group of foreign citizens born in Germany is first distinguished from that of foreign citizens born abroad. This distinction is based on a binary indicator in the AZR data that shows whether a person with foreign citizenship was born in Germany or immigrated here.[4] This assignment is thus independent of the information on the place and country of birth. According to this indicator, at the end of 2020 around 1.5 million of the 11.4 million registered foreign citizens were born in Germany (Table 2: QS identifier 1). The further processing steps concern the remaining 9.9 million foreign-born citizens. The sequence of these steps is based on whether the data sets contain potentially usable information on the place and country of birth or not.

4 This indicator derives from a comparison of the date of birth and the date of first entry to Germany. In the case of foreign children born in Germany, the date of birth is registered as the date of first entry in the AZR. For immigrant foreign citizens, the date of first entry does not correspond to the date of birth.

Entry for place and country of birth

For foreign citizens born abroad, if information on both the place and country of birth is available, the combination is first checked for plausibility by comparing it against the place file. If the combination of place and country of birth corresponds to an entry in the place file, it is classified as plausible (Table 2: QS identifier 2). If the combination is not found in the place file, further steps are taken: First, it is checked whether the place of birth entry is listed in the place file in combination with another country of birth. It is assumed that an existing place of birth entry is more reliable than the country of birth entered along with it. This is because there is a

Table 2
Results of the data processing

QS identifier	Description	Frequency	Percent
1 to 14	Foreign population	11,432,460	100
1	A) Born in Germany[1]	1,509,335	13.2
2 to 14	B) Born abroad	9,923,125	86.8
2 to 7	B1) Place of birth given: yes, Country of birth: yes	6,819,120	59.6
2	Combination plausible	6,327,680	55.3
3 to 7	Combination not plausible	491,440	4.3
3	Assignment via place of birth from place file	80,415	0.7
4	Name of country in place name entry	6,130	0.1
5	Country code in place name entry	10,870	0.1
6	Country of birth could be imputed	30,205	0.3
7	Country of birth could not be imputed	363,820	3.2
8 to 12	B2) Place of birth given: yes, Country of birth: no	3,064,250	26.8
8	Assignment via place of birth from place file	1,914,245	16.7
9	Name of country in place name entry	12,670	0.1
10	Country code in place name entry	6,805	0.1
11	Country of birth could be imputed	818,195	7.2
12	Country of birth could not be imputed	312,335	2.7
13	B3)Place of birth given: no, Country of birth: yes	25,530	0.2
14	B4) Place of birth given: no, Country of birth: no	14,225	0.1

68.8 % Information from the original data
17.8 % Assignments based on master files
7.4 % Imputed information
6.0 % Missing entries

1 The assignment is made by comparing the date of birth and the date of first entry.
When foreign children born in Germany are registered, the date of birth is given in the Central Register of Foreigners as the date of first entry.

higher probability of an error being made in the country of birth entry due to potential confusion over the citizenship. In these cases, the country of birth is therefore taken from the place file (Table 2: QS identifier 3).

If the registered place of birth is not found in the place file or if there are several potential countries for a particular place name, it is checked whether the free text information on the place of birth contains a reference to a specific country of birth. In many cases, the name or ISO code of the country is also recorded in the free text for the place of birth. If an ISO code or name from the official Classification of States and Territories (Statistisches Bundesamt, 2022) is found in the free text, this is adopted as the country of birth (Table 2: QS identifiers 4 and 5).

If this comparison provides no indication of a specific country of birth, the data set is considered to contain information on the place of birth but none on the country of birth. The country of birth is imputed (if possible) as described in the following section (Table 2: QS identifiers 6 and 7).

Entry for place, but not country of birth

If there is no entry for the country of birth but one for the place of birth in the AZR, the processing procedure first checks whether there is a unique link to a country for the place name in the place file.If this is the case, the corresponding country of birth is adopted (Table 2: QS identifier 8). Otherwise, it is deemed possible that the name (Table 2: QS identifier 9) or ISO code (Table 2: QS identifier 10) of the country of birth might have been entered in the free text entry for the place of birth. A probabilistic random hot-deck procedure is used if no country of birth can be found based on these rule-based approaches.

For each data record with no country of birth (receiver), this imputation procedure is used to search for a similar (donor) record that does contain country of birth information. The procedure makes use of the fact that the missing country of birth assignment for one person can potentially be taken from that of other similar persons. In this context, persons with the same place of birth and the same citizenship are considered similar. All complete and plausible data records are potential donor data records.

The procedure is thus oriented towards the distributions already observed in the plausibility-checked AZR data. It preserves these when assigning previously unknown countries of birth. If, for example, a country of birth is to be imputed for a person of French citizenship (person A in Table 3) whose place of birth is Paris, the distribution of the countries of birth of French men and women born in Paris provides the required probability distribution. Thus, in the example in Table 3, four of the five French citizens with Paris as their place of birth and whose information contains the country of birth (donor) were born in France. Accordingly, person A would be assigned to France with a probability of 80 % and to the United States with a probability of 20 %. If this person holds US citizenship, as in the case of person B, France would be assigned as the country of birth with 33 % probability, and the United States with 67 % probability, based on the observed distribution. ↘ Table 3

Table 3
Probabilistic imputation

	Citizenship	Place of birth	Country of birth
	Receiver		
Person A	french	Paris	N/A
Person B	US-american	Paris	N/A
	Donor		
Person C	french	Paris	USA
Person D	french	Paris	FRA
Person E	french	Paris	FRA
Person F	french	Paris	FRA
Person G	french	Paris	FRA
Person H	US-american	Paris	USA
Person I	US-american	Paris	USA
Person J	US-american	Paris	FRA

If such a probability distribution can be derived from suitable donor data records, the imputation is successful (Table 2: QA identifier 11). If no matching donor data records are found, no country of birth can be entered and the data record counts as missing (Table 2: QS identifier 12).

No place of birth entry

If there is no place of birth but a valid ISO code is contained as the country of birth, this is considered plausible and is used (Table 2: QS identifier 13).

Data records with no entry on the place and country of birth are considered unit non-responses and are not

processed. Probability-based assignment by citizenship is not used (Table 2: QS identifier 14), due to the small number of cases involved (0.1 %).

Overview of case constellations

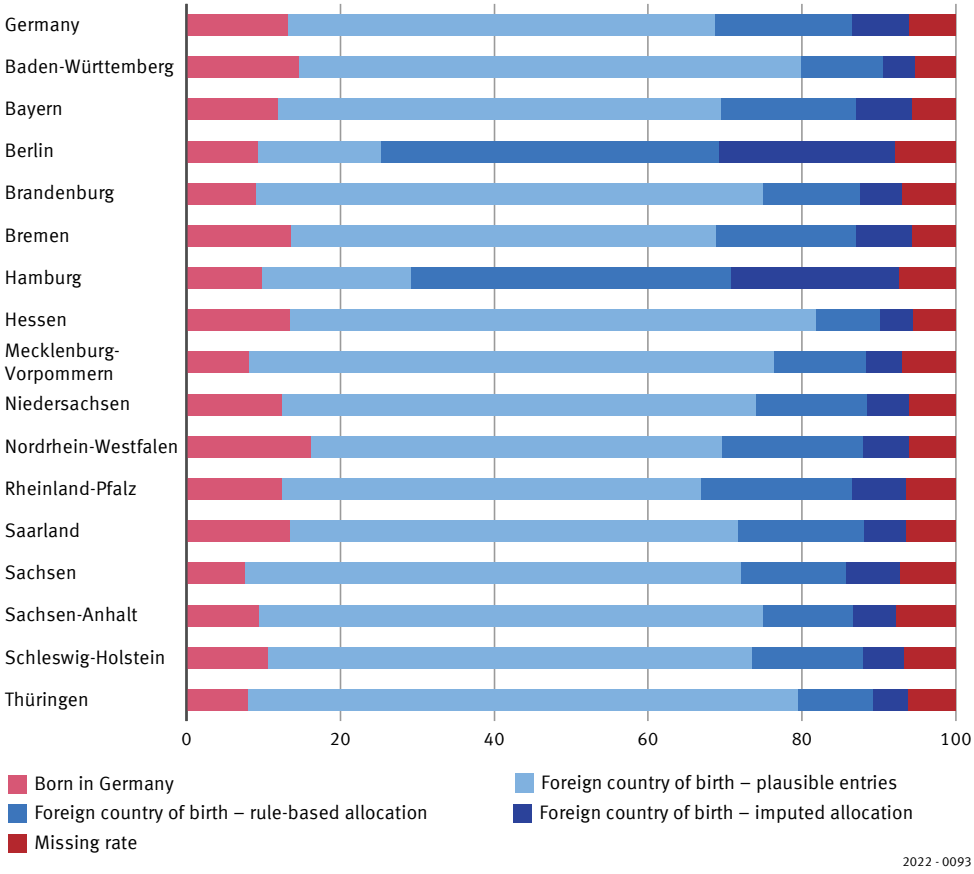
Table 2 summarises the case constellations as described, thus providing an overview of the AZR data processing result as of 31 December 2020. In total, 68.8 % of all country of birth data were taken from the original data.[5] In 17.8 % of cases, a missing country of birth could be deduced from the place of birth information contained in a master file. For 7.4 % of the foreign population, the country of birth was imputed based on the observed distribution among persons with the same place of birth and citizenship (imputation rate).[6] Yielding a final missing rate of roughly 6.0 %, the processing has led to a significant overall increase in quality.

The results for Germany as a whole, however, mask significant differences between the different Länder. Berlin and Hamburg in particular are well above the national average, with imputation rates of 22.9 and 21.8 % respectively. Excluding these two outliers, the average imputation rate in the remaining Länder is 5.7 %. The reason for this significant deviation is that only in rela-

5 The country of birth was Germany for 13.20 % of these (based on a comparison of date of birth and date of first entry); there was a foreign country of birth for 55.35 % based on plausible combinations of the place and country of birth variables; and a foreign country of birth for 0.22 % based on an entry for the country, but none for the place of birth.

6 Of these, 0.26 % cases had an implausible combination of place and country of birth, and the country of birth was missing in 7.16 %.

Figure 2
Results of processing the country of birth data from the Central Register of Foreigners, as of 31 December 2020
Percent



tively rare cases do Hamburg and Berlin record the country of birth when registering foreign citizens in the Central Register of Foreigners (AZR): in Berlin, only 16.0 % of registered foreign citizens have a plausible entry for the country of birth, in Hamburg the figure is 19.3 %. The corresponding average for the other Länder is 59.9 %. The fact that the missing rates for Berlin and Hamburg are still relatively low can be attributed to the fact that the mandatory place of birth information can be used to assign the missing country of birth during the processing. The examples of correct place-country combinations necessary for this are taken from the plausibility-checked data of all Länder. ↘ Figure 2

5

Analysing country of birth entries

Evaluations based on citizenship and those based on country of birth can be compared using the processed data. When it reformed the Nationality Act at the beginning of 2000, the German legislature introduced the principle of ius soli in addition to the principle of ius sanguinis. Since then, children born in Germany to foreign parents have been able to acquire German citizenship under certain conditions. One such requirement is that one of the parents has been legally resident in Germany for at least eight years and has a permanent right of residence (Section 4 (3) Nationality Act). The introduction of the principle of ius soli saw a significant initial decline in the number of children born in Germany without German citizenship (Bundesamt für Migration und Flüchtlinge, 2020). Since 2016, significantly more children without German citizenship have been born in Germany again. One reason for this is that the many people seeking protection who arrived between 2014 and 2016 do not yet fulfil the minimum period of residence of eight years required for parents.

↘ Table 4 compares the ten most frequent citizenships with the ten most frequently reported countries of birth. This shows that Germany was the most frequent country of birth of foreign persons at the end of 2020 (13 %). It is also noticeable that the order of the countries changes depending on whether citizenship or country of birth is used as the indicator of a person's origin. Croatian citizenship, for example, is the sixth most common foreign citizenship, but Croatia only ranks tenth among the most common countries of birth. This is due to the fact that only slightly more than half of the Croatians registered in the AZR were also born in Croatia. Just under 13 % of Croatian citizens were born in Germany, around 25 % in Bosnia and Herzegovina and 11 % in other countries.

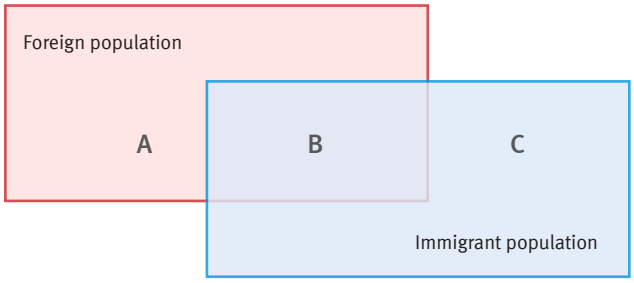
The results can be compared against microcensus evaluations in order to provide external validation of the quality of the processed data. It should be noted that only foreign citizens can be covered in the evaluations

Table 4

Most frequent countries of birth and most frequent citizenships of foreign citizens in the Central Register of Foreigners as of 31 December 2020

Position	Country of birth	Foreign citizens		Position	Citizenship	Foreign citizens		Of whom born in:		
								Germany	country of citizenship	other country
		number	%			number	%			
Total		11,432,460	100	Total		11,432,460	100	X	X	X
1	Germany	1,509,335	13.2	1	Turkey	1,461,910	12.8	26.8	69.2	4.0
2	Turkey	1,028,560	9.0	2	Poland	866,690	7.6	6.1	90.6	3.3
3	Poland	792,765	6.9	3	Syria	818,460	7.2	12.3	77.7	10.0
4	Syria	687,430	6.0	4	Romania	799,180	7.0	7.3	77.3	15.4
5	Romania	628,240	5.5	5	Italy	648,360	5.7	24.1	62.4	13.5
6	Italy	427,860	3.7	6	Croatia	426,845	3.7	12.9	51.3	35.7
7	Bulgaria	324,620	2.8	7	Bulgaria	388,700	3.4	7.3	82.7	10.0
8	Bosnia and Herzegovina	286,150	2.5	8	Greece	364,285	3.2	20.3	61.3	18.4
9	Greece	232,540	2.0	9	Afghanistan	271,805	2.4	9.9	74.3	15.8
10	Croatia	231,585	2.0	10	Russian Federation	263,300	2.3	5.3	73.6	21.1

Figure 3
Population groups by country of birth in Central Register of Foreigners and in microcensus



Group A: Domestic-born foreign population
Group B: Immigrated foreign citizens
Group C: Immigrated German citizens

2022 - 0094

on immigration and country of birth based on the AZR. German citizens (including persons with both German and a foreign citizenship) are not included in the AZR data. ↘ Figure 3 Thus the AZR analyses do not take into account persons who were born abroad but who then became naturalised after immigrating (group C). The microcensus, on the other hand, as a representative survey of the entire population, also includes this group.

Foreigners born in Germany (group A) and foreigners born abroad (group B) can therefore be used to compare the results on the country of birth in the AZR and in the microcensus. The Covid-19 pandemic lowered the quality of the 2020 microcensus with regard to the usual depth of statistical analysis (Statistisches Bundesamt, 2021a). Data from the 2019 microcensus were used for the comparison for this reason. ↘ Table 5 compares the shares of foreigners born in Germany and those born abroad, as well as the relative distribution of individual countries of birth from both data sources.

Table 5
Results of the AZR-microcensus comparison

	Evaluation of Central Register of Foreigners as of 31 December 2020	Result of 2019 microcensus	Difference
	%		Percentage points
Foreign population, total	100	100	0.0
Born in Germany	13.3	15.5	– 2.2
Born abroad	86.7	84.5	2.2
of whom:			
With information on country of birth	93.0	99.8	– 6.8
Turkey	11.1	11.4	– 0.3
Poland	8.6	8.3	0.3
Syria	7.4	7.9	– 0.5
Romania	6.8	5.6	1.2
Italy	4.6	5.7	– 1.0
Bulgaria	3.5	2.9	0.6
Bosnia and Herzegovina	3.1	3.1	0.0
Greece	2.5	3.2	– 0.7
Croatia	2.5	3.0	– 0.5
Russian Federation	2.5	2.8	– 0.3
Kosovo	2.4	2.5	0.0
Iraq	2.4	2.3	0.0
Afghanistan	2.2	2.1	0.1
Serbia	2.0	2.1	– 0.1
Hungary	1.8	1.8	0.0
Austria	1.6	1.9	– 0.3
China	1.5	1.4	0.1
Iran, Islamic Republic of	1.5	1.2	0.2
Ukraine	1.4	1.6	– 0.1
Spain	1.3	1.4	– 0.1
Other countries	29.1	27.7	1.4

↘ Representation of country of birth in the AZR and in the microcensus

The following qualifications must be borne in mind when comparing results from the microcensus and the AZR: The microcensus is a representative population survey based on a sample. The figures from the AZR, on the other hand, are based on administrative data for all foreigners in Germany. The microcensus extrapolation is carried out using the results of the intercensal population updates. There are significant discrepancies between the intercensal population updates and the AZR with regard to the absolute size of the foreign population (Statistisches Bundesamt, 2021b). Furthermore, the microcensus only asks the population living in private households about their country of birth, whereas the AZR also includes people in collective living quarters.

The share of immigrants among the foreign population in the microcensus is 84.5 %, 2.2 percentage points below the result from the AZR data. The microcensus lists almost the same countries as the AZR as the most frequent countries of birth. The order and the distributions are also comparable. This concordance – despite the major differences in the survey methodology – is a powerful indicator of the validity of the data processed from the AZR.

6

Outlook: Possible geocoding of places of birth with OpenStreetMap

Place of birth entries for foreign citizens in the AZR provide an indication of a person's precise origin within a country's national borders. They offer a variety of possibilities for increasing the evaluation potential of the statistics of foreigners and of those seeking protection. Reliable determination of the correct country of birth is a necessary prerequisite for unambiguous identification of a place of birth. Three further conceptual steps must be taken for the geocoding of location information: gazetteer matching, toponym resolution and geocoding. Leidner (2008) provides a good overview here.

Gazetteer matching identifies the potential geographical locations which could be meant by a place name (Goldberg and others, 2007). Gazetteer matching is thus comparable to searching for place names in an index of places. This index of places can be taken directly from OpenStreetMap (OSM), a worldwide gazetteer, or from a place file previously geocoded using OSM data.

The OpenStreetMap community has been collecting worldwide geoinformation since 2004 and making it available as open data. Based on a system comparable to that used for the free encyclopaedia Wikipedia, an open community of volunteers with no commercial interest collects the information and continuously updates it.

The data quality varies from region to region and depends to a great extent on how active the local editors are. Seto and others (2020) have identified the most active communities for the period 2013 to 2019, namely those in Germany, the Russian Federation and the United States. There has been an increased focus on regional differences in access to geospatial data since the devastating earthquake in Haiti in 2010. Since then, the OSM community, which also includes numerous aid organisations such as the Red Cross, has been filling geodata gaps in less developed regions of the world in support of the "humanitarian mapping" effort (Herfort and others, 2020).

The result of the gazetteer matching is a list of potential geographical locations that could be meant by a given place name (toponyms). Toponym resolution then involves establishing a probability distribution for these candidates (DeLozier and others, 2015). The frequency distributions in the existing data can be used for this purpose. The ambiguity of place names can be resolved from additional name identifiers, such as a federal state or a river. For example, the probability that "Frankfurt" means "Frankfurt am Main" as the place of birth would be estimated from the relative frequency of "Frankfurt am Main".

$$P(\text{Place=Frankfurt am Main, DE} \mid \text{Name=Frankfurt, Country=DE})$$
$$= h(\text{Frankfurt am Main, DE} \mid \text{Country=DE})$$

Subsequent geocoding can then be performed online using an application programming interface (API). A significant advantage of using APIs is that some include autocomplete functions which automatically correct minor spelling mistakes and recognise alternative spellings of a place. The autocomplete function thus replaces the need for time-consuming preprocessing by the user. The solution is yet to be tested in a feasibility study.

7

Conclusion

This paper presents the processing of the "country of birth" and "place of birth" variables from the Central Register of Foreigners for their use in official statistics. Although information on the place of birth is recorded in almost all cases in the AZR, it is recorded as free text in an unstructured form. Information on the country of birth, by contrast, is recorded in standardised form as an ISO code, however the information is not provided in a considerable number of cases.

In the first step the Federal Statistical Office developed a solution for processing the country of birth data. The comparative data set of worldwide place names and assigned countries held by the Federal Statistical Office plays a central role in this solution. The quality of the processed data depends to a significant extent on the quality of this master file.

The results of the processing of the Central Register of Foreigners data at the end of 2020 are promising. For foreign citizens born abroad, the proportion of missing (26.8 %) or implausible (4.3 %) information on the country of birth was reduced from a total of 31.1 % in the raw data to 6.0 % in the processed data. This significantly improved the quality of the administrative data for statistical use. The good comparability with evaluations of the country of birth from the microcensus indicates the reliability of the processed data.

The information on places of birth might be geocoded in the second step. Geocoding allows the local origin of a person to be defined more effectively, thereby opening up new evaluation possibilities on a more detailed small-area level. A potential processing solution could be based on OpenStreetMap geodata.

BIBLIOGRAPHY

Bundesamt für Migration und Flüchtlinge. *Migrationsbericht 2020*. [retrieved on 1 March 2022]. Available at: www.bamf.de

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

Feuerhake, Jörg/Lange, Kerstin/Siegismund, Annelen/Vigneau, Elsa. *Kodierung des Geburtsstaats in der Wanderungsstatistik*. In: WISTA Wirtschaft und Statistik. Issue 3/2020, p. 98 ff.

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

Goldberg, Daniel W./Wilson, John P./Knoblock, Craig A. *From text to geographic coordinates: the current state of geocoding*. In: URISA Journal. Volume 19, Issue 1/2007, p. 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. Volume 11. Article 3037/ 2021.

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

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

Statistische Ämter des Bundes und der Länder. *Gemeindeverzeichnis-Online*. [retrieved on 10 March 2022]. Available at: www.statistikportal.de

Statistisches Bundesamt. *Digitale Agenda des Statistischen Bundesamtes*. 03/2019. [retrieved on 8 March 2022]. Available at: www.destatis.de

Statistisches Bundesamt. *Qualitätsbericht Ausländerstatistik*. 2021a. [retrieved on 1 March 2022]. Available at: www.destatis.de

Statistisches Bundesamt. *Ausländische Bevölkerung*. 2021b. [retrieved on 1 March 2022]. Available at: www.destatis.de

Statistisches Bundesamt. *Staats- und Gebietssystematik*. [retrieved on 1 March 2022]. Available at: www.destatis.de

LEGAL BASIS

Central Register of Foreigners Act (AZR Act) of 2 September 1994 (Federal Law Gazette I page 2265), last amended by Section 8 of the Act of 9 July 2021 (Federal Law Gazette I page 2467).

Nationality Act (StAG) in the adjusted version published in the Federal Law Gazette Part III, number 102-1, as last amended by Section 1 of the Act of 12 August 2021 (Federal Law Gazette I page 3538).

Second Act to Improve Registration and Data Exchange for Purposes of Residence and Asylum (Second Data Exchange Improvement Act - 2nd DAVG) of 4 August 2019 (Federal Law Gazette I page 1131).

Extract from the journal WISTA Wirtschaft und Statistik

Published by:
Statistisches Bundesamt (Federal Statistical Office)

www.destatis.de

You may contact us at
www.destatis.de/kontakt

Abbreviations

WISTA	=	Wirtschaft und Statistik
JD	=	annual average
D	=	average (for values which cannot be added up)
Vj	=	quarter of a year
Hj	=	half-year
a. n. g.	=	not elsewhere classified
o. a. S.	=	no main economic activity
St	=	piece
Mill.	=	million
Mrd.	=	billion

Explanation of symbols

–	=	no figures or magnitude zero
0	=	less than half of 1 in the last digit occupied, but more than zero
.	=	numerical value unknown or not to be disclosed
. . .	=	data will be available later
X	=	cell blocked for logical reasons
I or —	=	fundamental change within a series affecting comparisons over time
/	=	no data because the numerical value is not sufficiently reliable
( )	=	limited informational value because numerical value is of limited statistical reliability

© Statistisches Bundesamt 2022

Reproduction and distribution, also of parts, are permitted provides that the source is mentioned.

Figures have in general been rounded without taking account of the totals, so that there may be an apparent slight discrepancy between the sum of the constituent items and the total as shown.