International Journal of
Geo-Information

Article

Geocoding Freeform Placenames: An Example of Deciphering the Czech National Immigration Database

Jan Šimbera [1], Dušan Drbohlav [2] and Přemysl Štych [1,*]

[1] Department of Applied Geoinformatics and Cartography, Faculty of Science, Charles University, Albertov 6, 128 43 Prague, Czech Republic; simberaj@natur.cuni.cz

[2] Department of Social Geography and Regional Development, Faculty of Science, Charles University, Albertov 6, 128 43 Prague, Czech Republic; dusan.drbohlav@natur.cuni.cz

* Correspondence: stych@natur.cuni.cz

Abstract: The growth of international migration and its societal and political impacts bring a greater need for accurate data to measure, understand and control migration flows. However, in the Czech immigration database, the birthplaces of immigrants are only kept in freeform text fields, a substantial obstacle to their further processing due to numerous errors in transcription and spelling. This study overcomes this obstacle by deploying a custom geocoding engine based on GeoNames, tailored transcription rules and fuzzy matching in order to achieve good accuracy even for noisy data while not depending on third-party services, resulting in lower costs than the comparable approaches. The results are presented on a subnational level for the immigrants coming to Czechia from the USA, Ukraine, Moldova and Vietnam, revealing important spatial patterns that are invisible on the national level.

Keywords: geocoding; transliteration; transcription; migration; Python; Czechia

check for updates

Citation: Šimbera, J.; Drbohlav, D.; Štych, P. Geocoding Freeform Placenames: An Example of Deciphering the Czech National Immigration Database. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 335. https://doi.org/10.3390/ijgi10050335

Academic Editors: Wolfgang Kainz and Giuseppe Borruso

Received: 14 February 2021
Accepted: 9 May 2021
Published: 15 May 2021

1. Introduction

International migration is a complex phenomenon which is highly relevant to modern society. On the one hand, technological changes and globalization make it easier than ever to relocate to a different country; on the other hand, immigration, in addition to its positive impacts, might also cause significant societal and political tensions. In order to better understand, measure, and manage migration movements, accurate and up-to-date data are necessary [1]. Spatial data about migrants, such as their places of birth, can reveal important insights about their material, social, and behavioral backgrounds (e.g., urban versus rural areas, large cities versus others, poor versus rich regions, etc.) allowing us to better assess their needs and intentions in new destinations [2].

Migration data, however, are often only available in a non-spatial form, bearing only freeform text descriptions about places, which need to be geocoded in order to obtain a geospatial database. In Czechia, the spatial identification of immigrants is stored in poorly arranged freeform text fields, which makes extracting any insights directly from them extremely challenging, and often impossible. Geocoding these fields—transforming textual descriptions into geospatial coordinates—would contribute to improving the situation.

Geocoding is the process of transforming the text description of a location into a geographical position specified as a point [3]. In this regard, querying for a location requires different approaches to common information retrieval [4]. Moreover, despite their abundance, relying on online geocoding services as external resources is impractical and sometimes impossible as a result of their terms of use [5], especially the imposed quotas. Given that our dataset is large compared to the allocated budget, the use of online services was not an option.

The trivial approach to geocoding is a simple lookup in a list of placenames that are matched to geographical locations—a gazetteer [6]. The precision of geocoding is then determined by two factors:

- The completeness of the gazetteer, both in names and other data that can resolve ambiguities [7].
- The accuracy of the input placenames—namely the presence of typographical errors, bad or non-conventional formatting, or spelling.

The insufficiency in the second factor is usually compensated for by using more complex geocoding algorithms. These usually come at the price of performance and geocoding speed, which can be a concern.

Geocoding is often considered a special subproblem of Spatial Named Entity Linking (NEL). NEL is the task of determining the unique identity of entities mentioned in a text [8,9], namely their subdomain of entity resolution, which uses other knowledge base resources besides the text itself (in our case, the gazetteer) [10]. NEL approaches usually consist of treating the knowledge base labels and/or the linked text in order to maximize the chances of finding the correct matches (see [11–14]). This is often, but not always, achieved by expanding the knowledge base labels [15]. In Spatial NEL specifically, using additional criteria such as population, area or popularity to enhance the matching accuracy is common, see [12] or [16,17].

The most commonly used gazetteer is the freely available GeoNames database [18]. It is widely used for geocoding data on settlement level [19]. Ahlers (2013) [20] reports a multitude of issues within the gazetteer, such as grid patterns, imprecise coordinates, overlaps and repetitions, and misclassifications, which are systematic in some cases but appear indistinguishable from correct data. Despite these detected errors, however, it appears to be of sufficient quality for general use. For the creation of a geospatial search engine in Latin America, it was supplemented with geotagged data from Wikipedia articles using data fusion based on entity merging and geographic conflation [21]. Valkanas and Gunopulos (2012) [5] merged GeoNames with the Flickr places dataset. OpenStreetMap, besides its open-source Nominatim geocoding service, can also serve as a valuable source of geospatial information, comparable to proprietary gazetteers [22].

A common approach to improve geocoding quality is to allow the geocoder to make inexact matches in the placenames, overcoming issues such as a single typographical error that would otherwise ruin the search. Liao and Wang [23] performed this for departmental data in China using the BPM-BM filter algorithm, which is especially well-formed for Chinese, and achieved an accuracy of 94.2%, higher than a simple SQL-based lookup. The Intiendo algorithm uses matching based on edit (Levenshtein) distance [24], as did Lan and Longley when geocoding historical British census addresses [25]. The edit-distance based fuzzy matching package was used for the geocoding of historical addresses in Scotland from 1855 to 1974 [26].

A simpler approach than fuzzy matching is to employ string substitution on selected, repeated error cases [27]. More involved methods are required when a change of script is needed (the case studied here) because there are many possible transliterations. A transcription system from Arabic to English was created using a neural network coupled with a knowledge-based system to vowelize the consonantal Arabic script and then filter out improbable transcription variants [28].

Some of the studies examining on-line geocoding services go beyond evaluating the services and try to propose pre- or post-processing strategies to enhance the accuracy while keeping the number of queries to a minimum [29]. Ahlers and Boll [30] analysed the Google and Yahoo APIs, which are based on different proprietary gazetteers, and presented a correction methodology to assemble their result. Karimi, Sharker and Roongpiboonsopit [31] presented a geocoding recommender algorithm that can recommend optimal online geocoding services according to the type of place name being searched, and their accuracy. Other approaches aim to utilize the spatial dimension of the problem. Ping and Yong [32] used a place name ontology to store the gazetteer data, which also allows the geocoder to consider more complex data such as distances, topologies, and geometric footprints. Coetzee and Rademeyer [24] presented a method to match the raw geocoding result by its spatial adjacency to a more suitable result.

The objective of this article is to present and test a new geocoding method invented for the purpose of geocoding the database of immigrants based on their birthplaces. The method enables us to geocode the noisy Foreigners Information System immigration dataset with completeness and accuracy exceeding the conventional services. This article aims to address the issue using rule-based transliteration and local batch geocoding. The method was tested and evaluated based on two criteria: completeness and accuracy. Selected results of the geocoded database of migrants are presented in this study. The usefulness of the method is demonstrated on several selected examples from the given data, and this enables us to see a development of the migration flows over time. The geocoding method is available for free for a wide range of the potential users.

The source dataset and the developed method are described in part 2, followed by the calculation results and the evaluation of their accuracy in part 3.1. The geocoded results are presented in maps and briefly commented upon in part 3.2. Part 4 contains a discussion of the potential usage and shortcomings of the obtained dataset. Conclusions are drawn in part 5.

2. Materials and Methods

2.1. Data

We had at our disposal, and processed, a unique database derived from the Foreigners Information System (Cizinecký informační systém—CIS), which is managed by the General Directorate of Alien Police of the Czech Republic. The Directorate provided us with data on registered foreigners in the territory of Czechia, coming from 10 selected countries, sorted into yearly batches covering the period from 2008 to 2017. The data contain anonymized data on individual applications, with no identifiers linking back to the individuals concerned. They contain the following fields:

- Type of residence permit requested—temporary or permanent. Whereas the former is represented mainly by stays in Czechia on the basis of a long-term visa, long-term residence permit, or (concerning EU citizens) temporary protection status, the latter takes into account chiefly permanent residence permits and international protection statuses (primarily asylum and subsidiary protection). Due to the marginal nature of the latter category in Czechia, the data mainly brings about an overall picture of documented labour and family migration into the country, of both third-country nationals or EU citizens.
- Citizenship, year of birth, and gender.
- Place and country of birth, both in freeform text. This contains the essential information to geocode the record.
- Country of previous residence. This often did not align with the country of birth if the immigrant had moved internationally before coming to Czechia.
- Application decision status: granted, denied, pending, or other. For our analyses, we only filtered out granted applications.

Apart from the place and country of birth, the attributes are not specific enough to enable back-linking to individuals through attribute combinations, as required by the EU's General Data Protection Regulation. The dataset, except for a fully anonymous random sample of birthplace strings passed to the external geocoding service as described in Section 3.1, was not shared with third parties, in order to further enhance the data privacy.

The data come from application forms filled out directly by immigration applicants, either in electronic form into templates provided at the Ministry of the Interior website, or in paper form at the respective offices of the Department of Asylum and Migration Policy of the Ministry of the Interior of the Czech Republic. The forms differ for EU citizens [33] and non-EU citizens [34]. The applications must be submitted and all data provided with the request for a residence permit or its prolongation. In the case of non-EU citizens, data from the immigrant's visa application, as submitted to the Czech embassy or consular office, are often reused. The officials are supposed to check the forms and data before accepting them and prompt the applicant to improve them, but in reality, they rarely do so, except when Latin script is not used; they especially tend to respect the data from the visa applications.

The instructions on filling out the form clearly stipulate "the answers in this application form must be typed or written in block letters in Czech" [34]. No instructions on placename spelling or transliteration (especially from Cyrillic, which is very common) are provided, nor is the specification of the administrative level that should be stated as the place of birth (the locality, municipality, region or their combination). This, along with the rather lax approach (and sometimes even limited geographical literacy) of immigrants and officials, contributes to the low quality of the CIS data. This creates a vicious circle in which the Alien Police do not use the contained information, and without use for the data, there is no incentive to improve its quality.

The aforementioned issues mainly affect place of birth data. They suffer from multiple issues that make their geocoding difficult.

- When the source language does not use the Latin alphabet, the transcription is often flawed:

○ Transcriptions into Czech and English are mixed both across words and within individual words.

○ Sometimes, a write-as-you-hear transcription is used, with a non-negligible amount of errors.

○ When the Czech transcription is used, diacritical marks are often omitted.

○ This issue is amplified by the fact that the gazetteer used for geocoding may also not contain all of the correct transcriptions.

- Exonyms in Czech, English, and from the source language are frequent. Exonyms are hard to geocode because there is usually no systematic way to derive them from endonyms; they must also be present in the gazetteer. A similar issue is presented by the usage of historical placenames (e.g., Soviet-era city names in Ukraine).
- Typographical errors are present in a significant number of cases.
- The specified country and placename of birth do not match (e.g., the country is given as Czechia but the placename is Moscow).
- Placenames of different hierarchical levels are mixed. This mostly happens in Ukrainian and U.S. placenames, where sometimes a specific placename is used, sometimes just the name of the administrative division (U.S. state or Ukrainian oblast), and sometimes they are used together, in no particular order.

This means that a single place is usually specified using multiple placenames. For example, for the Ukrainian city of Vinnycja, we counted 30 variants: Vinica, Vinicia, Vinicja, Vinitsa, Vinncja, Vinnica, Vinnice, Vinnicia, Vinnicja, Vinnicya, Vinnitca, Vinnitsa, Vinnitsja, Vinnitsya, Vinnyca, Vinnycia, Vinnycija, Vinnycja, Vinnycka, Vinnycya, Vinnytsa, Vinnytsia, Vinnytsja, Vinnytsya, Vinycya, Vynnycja, Vinyca, Vinycja, *Vinycla* and *Vinita*. (The last two variants—emphasized—do not match phonetically and are therefore most likely a typographical error.)

Nevertheless, despite their limits, the potential value of the provided data is significant. However, when we tried to geocode the data using standard tools such as on-line geocoding services, we failed to achieve acceptable levels of completeness and accuracy. Therefore, we attempted to devise a custom geocoding method that would address the issues of faulty data such as ours.

We geocoded and analysed the data on all of the concluded procedures concerning granted permanent or temporary residence between 2008 and 2017 for the citizens of eight source countries of immigration into Czechia: Ukraine, Vietnam, Russia, Poland, the USA, Moldova, Belarus and Georgia. Except for Georgia, all of these countries represent important sources of immigration for Czechia [35]; Georgia was included in order to expand the diversity of languages and writing systems.

The dataset had approximately 500,000 records. Some of the records were incomplete, and were therefore dropped when sorting out our data by the Alien Police; because of this, our data set does not precisely correspond to publicly available data sources, e.g., the total figures would not exactly match [36].

2.2. Methods

For the purposes of this study, as we were mostly dealing with a small amount of source languages and many issues specific to the source, we opted to use a knowledge-based system consisting of custom devised transcription rules tailored to their combination.

First, we apply a set of transcription rules to the lowercase raw birthplace string in order to produce one or more transcription variants; contrary to some approaches [11], we expand the linked string, not the gazetteer labels. Then, we successively query each of the transcribed variants against the gazetteer (using fuzzy matching) until one of them produces any matches; if there are multiple matches, we select among them using an objective function combining string similarity and place importance. Each of these steps is explained in more detail below.

The code to perform the geocoding was developed in Python as a set of scripts using a PostgreSQL database backend (see Supplementary Materials). It is available as open-source software from a repository at http://github.com/simberaj/migration-geocode/ (accessed on 1 May 2021) [37].

2.2.1. Transcription

Each transcription rule was specified as a regular expression to be evaluated against the source string; each match is then replaced with one or more variants, all of which are checked by a lookup in the GeoNames gazetteer.

The rules were derived from the following sources:

- The correspondence between the Czech phonetic orthography (which is often used in the input) and the standard Latin orthography of the language of the given country was examined; in the case of countries using non-Latin writing systems, the standard English transcription, which dominates in the gazetteer, was used. In many cases, this produces rather simple non-variant rules transforming strings from one orthography to the other.
- In the process of developing the geocoding engine on the input data, we devised further empirical rules that improved its accuracy on frequent non-standard transcriptions or typographical errors. This was achieved using a hold-out sample, which was distinct from the validation sample. The sample was repeatedly geocoded using the engine, and its mistakes and unmatched records were manually examined with the help of online web searches to produce new rules. The performance of these rules was tested in the following iterations. An example of a rule devised in this manner is the variant rule transforming between a single and a double 'N' in Ukrainian, as shown in Figure 1.

The rules are hierarchically grouped into rulesets applying to a particular language, in order to account for different transcriptions of, e.g., Ukrainian and Romanian. The ruleset is assigned to a placename based on the specified country. Table 1 shows the complete ruleset for Hungarian; other rulesets can be found in the configuration of the geocoding engine in its code repository.

The rules are applied to the input placename in a specified order. An example of such a progressive application of the rules from the Ukrainian ruleset for one of the forms of Vinnycja is shown in Figure 1. As noted above, the high data volume, furthered by a count of the variants produced, means that we cannot easily use standard web geocoding services due to the increased cost or exhausted quotas; instead, we used a custom local geocoding backend.

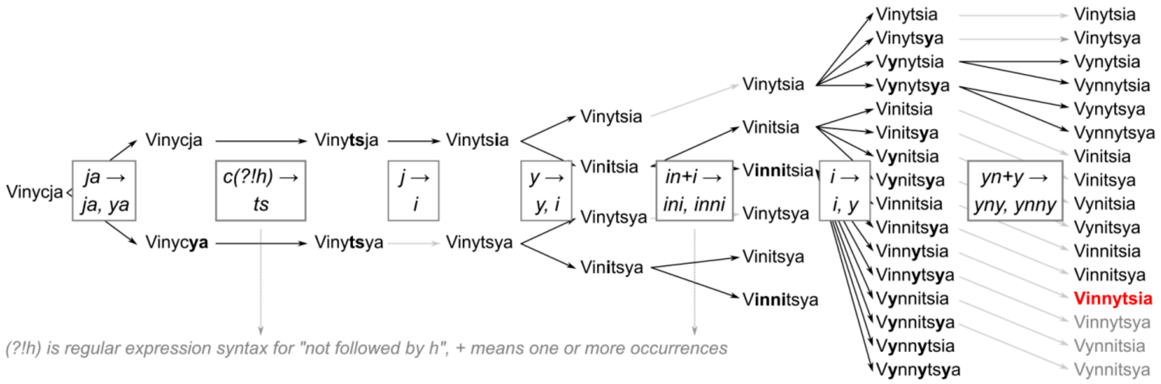

Figure 1. Transcription of an example placename using the developed system. The transcription rules are applied in the specified order from left to right; solid black arrows indicate where the specific rule matched; if that results in variant expansions, the arrows bifurcate. The output variants at the right side (after applying all of the rules from the ruleset) are tried top to bottom until one of them—in red—produces a match to the gazetteer. Duplicate variants from the bottom of the variant list are eliminated. Source: our own research.

Table 1. The transcription rules for Hungarian placenames. Source: own research.

Match in Input	Output Variant(s)	
i	i	yí
ö	ö	ő
č	cs	
s not before z	s	sz
ž	zs	
š	s	
d′	gy	
t′	ty	
j	j	ly

2.2.2. Gazetteer Matching

The gazetteer backend was built in the form of a PostgreSQL database with the pg_trgm extension that enables fuzzy matching; this means that even after transcription, the placename strings need not match exactly, but need to be sufficiently similar. Specifically, the pg_trgm module performs trigram matching; it computes the fraction of three letter sequences shared among the two strings [38]. This approach was chosen against the common alternative, edit distance matching (used e.g., in [add1]); while edit distance matching produced comparably accurate results, it was approximately one order of magnitude slower.

For a single place, the GeoNames database usually contains multiple alternative names, such as different national transcriptions, exonyms and historical names; we used all of these for the lookup. The exonyms and historical names are particularly useful, as they are often contained in our source data and there is no general way to arrive at them using fuzzy matching.

2.2.3. Result Selection

The fuzzy matching procedure usually returns multiple results for a single placename, ordered by similarity (the fraction of trigrams shared between the search term and the result). In order to decide which result to choose, we employ the following objective function:

$$s + \frac{\log P}{100}$$

where s is the trigram similarity and P is the population of the matched place. This gives a very slight preference (with mostly tie-breaking effects only) to places with a larger population, which naturally tend to occur more often, in order to prevent the assignment of large amounts of records to smaller places that happen to share the name of their larger counterpart.

Other types of features than populated places are also present in the GeoNames database; we distinguish among them using their one-letter feature code. Administrative units are given a fictitious population of 100 in order to give preference to them over very small settlements that would happen to share the name with the unit. Other types are less preferred and are given a fictitious population figure close to zero in the following order: buildings, roads, localities, natural features, and others.

The point coordinates of the result with the highest value of the objective function are then returned to geocode the placename.

2.2.4. Comparison

We compared our results with the results of the Nominatim [39] and Geoapify [40] on-line geocoding services on a sample of the input database in order to measure the performance of our approach approximately. Nominatim is a de-facto industry standard open-source geocoder based on OpenStreetMap, while Geoapify was chosen as an example of a commercial geocoding service with a freely accessible pricing tier.

3. Results

3.1. Algorithm

The performance of the created geocoding engine was evaluated using a sample of 1000 rows that were manually labeled with the help of web search where possible; 692 records were unique. The sample was created by the random sampling of the entire input dataset, such that the proportions of the country samples approximately matched the ones from the full dataset. The sampling produced some duplicate rows; these were kept in the sample in order to preserve the differences in the importance of the individual places. Out of the 1000 rows, nine strings were not geocodable, carrying no meaningful location information.

We also compared our engine with the Nominatimand Geoapify on-line geocoding servicesby running the same sample through it.

Because our engine produces point results, we chose a distance threshold of 10 kilometers; a result within this distance of the labeled location was considered a match. This threshold was checked manually in order to minimize false positives and false negatives. The Nominatim and Geoapify services return bounding boxes along with the points; therefore, it was considered a match when the labeled location fell within the bounding box.

The following accuracy metrics were measured, inspired by named entity linking metrics from [41]:

- geocoding precision: the fraction of correct matches out of all of the locations retrieved,
- geocoding recall: the fraction of correct matches out of all of the geocodable locations,
- geocoding F-score: a harmonic mean of geocoding precision and recall, regarded as the primary quality metric,
- nil precision: the fraction of ungeocoded records that truly did not carry location information,

- nil recall: the fraction of records that did not carry location information that were not geocoded (the lower the result, the more this set was "polluted" by false positives),
- completeness: the fraction of records for which a location was retrieved (although this is not in a true sense an accuracy metric, it nevertheless is an important measure of the usefulness of the result).

Variants of our geocoding engine with transcription rules or fuzzy matching turned off were also evaluated in order to show the effect of these components. From Table 2, it is clear that both components significantly improve the solution in all of the measured metrics. The contribution of the transcription rules seems to be higher than that of fuzzy matching, which, when used alone, tends to decrease the geocoding precision compared to the variant with both components turned off (where the geocoding is essentially reduced to the equality querying of the GeoNames database).

Table 2. Accuracy metrics for the developed geocoding engine and its variants with individual components turned off, compared to the third-party Nominatim and Geoapify engines. Source: our own research.

	F-Score	Precision	Recall	Nil Precision	Nil Recall	Completeness
Full engine	90.4%	91.7%	89.2%	22.2%	88.9%	96.4%
Transcription only	83.8%	90.9%	77.7%	5.2%	88.9%	84.7%
Fuzzy matching only	78.4%	87.2%	71.2%	4.2%	88.9%	81.0%
Both components off	72.3%	91.8%	59.6%	2.2%	88.9%	64.4%
Nominatim	73.4%	86.3%	63.8%	2.6%	77.8%	73.2%
Geoapify	64.5%	74.5%	56.8%	2.9%	77.8%	75.6%

Compared to the de-facto industry standard Nominatim geocoder, only the raw equality querying of the GeoNames database fared worse on the testing sample, while other alternatives presented an improvement in both accuracy and completeness. The Geoapify geocoder was even worse in comparison, hinting at the superior quality of the GeoNames gazetteer for the studied countries.

In the main method (using both fuzzy matching and transcription rules), geocoding errors are caused in about half of the cases by imperfect transcription (the placename being transcribed to a name of a different place) and in the other half by intrusion of the name of another entity (usually a higher administrative unit). These are also among the most common causes for the engine failing to geocode a location, along with mismatches between the placename and country, and heavily abbreviated placenames. Only one placename from the sample was missing from the GeoNames gazetteer but present in other sources.

Table 3 breaks the figures down by the given countries. Whereas for the Nominatim and Geoapify geocoders, countries using Latin script fare better than their Cyrillic counterparts, the differences are insignificant with our engine. The bad performance of our engine on USA placenames is caused by its tendency to prefer even less populated places at the expense of frequently occurring administrative division (state) names. The difference between Russian and Ukrainian data mainly stems from the higher share of smaller settlements among the records for Ukraine, for which the gazetteer usually does not contain many alternative names. Some records are also present with Czechia as the source country; this concerns some cases where the country does not match the birthplace, as mentioned above, as well as some cases of children who were already born in Czechia but due to Czech *ius sanguinis* nationality law nevertheless had to apply for a residence permit.

Table 3. Accuracy metrics for the main variant of the geocoding engine by input country, compared with third-party services. Countries with very few records are omitted. Source: our own research.

Country	Records	Full Engine F-Score	Nominatim F-Score	Geoapify F-Score
Belarus	23	90.5%	72.2%	75.7%
Czechia	54	91.6%	70.3%	72.5%
Moldova	42	98.8%	85.7%	86.1%
Russia	164	97.2%	77.2%	71.5%
Ukraine	466	88.6%	64.2%	42.0%
USA	75	68.9%	83.2%	85.3%
Vietnam	145	96.2%	85.6%	91.6%

3.2. Analysis of the Immigrants Coming to Czechia

We focused on the place of birth of migrants who have already migrated to Czechia. The results of the geocoding are presented in several maps, differentiated by the country of origin. For the purposes of presentation, the data were aggregated to areas—either to regular grids (in the case of Ukraine and Moldova, where the sufficient density of the data permits this) or to administrative divisions (in other countries where there are not enough input records to show the results in a grid with a meaningful resolution; this has an additional advantage of masking out the problematic cases where an administrative unit name was undetectably included in the placename, as discussed in Section 2.2). As the main purpose of this paper is to shed new light on methodological and methodical aspects, when interpreting the results of the analysis, we will limit ourselves to a simple description.

The map of the USA (Figure 2) tells us that the migrants coming to Czechia in the given years were mainly born on the east coast or in the central-eastern parts of the USA, rather than in the West. The importance of highly urbanized agglomerations like New York, Philadelphia, Washington, Chicago in the East, Miami in the South, and the Los Angeles and San Francisco areas in the West is prominent. Nevertheless, some other sources of immigration to Czechia are represented by more rural regions in Louisiana, Texas, or the Midwest. Altogether, there seems to be no particularly significant clustering of the given immigrants´ birthplaces, especially taking into account the fact that the data contain a high amount of error, as stated in Table 3.

As for Vietnam (see Figure 3), a clear spatial pattern is easily recognized. Significantly, the pattern shows a larger number of migrants from North Vietnam (the former communist part of the now-unified country) than South Vietnam. This has to do with the first immigration wave of Vietnamese students and trainees who came to former Czechoslovakia with the program of international aid among communist countries of the Soviet bloc during the 1970s and 1980s. The main birthplaces of Vietnamese immigrants to Czechia are the urban areas of Hanoi and Haiphong, along with their agglomerations, and the surrounding northern provinces of Nghe An, Ha Tinh, and Quang Binh.

Moldova, on the other hand, represents an opposing case, where a homogenized spatial pattern is clearly visible (see Figure 4). In fact, the whole country—both urban (including Chisinau, Balti, Cahul, Dunasari, and also Tiraspol, in the Dniester Moldovan Republic) and rural areas throughout the country—generate migrants who head for Czechia. Compared to Ukraine, migrants from Moldova to Czechia more often tend to request temporary residence (Figure 5), and the fraction of females is significantly lower (Figure 6). This shows the dominance of typical male professions in the Czech labor market (predominantly in construction and industry) for those migrants, though there has recently been an increase in the migration of Moldovan females (getting jobs, e.g., in services, as care-givers or cleaners, etc.).

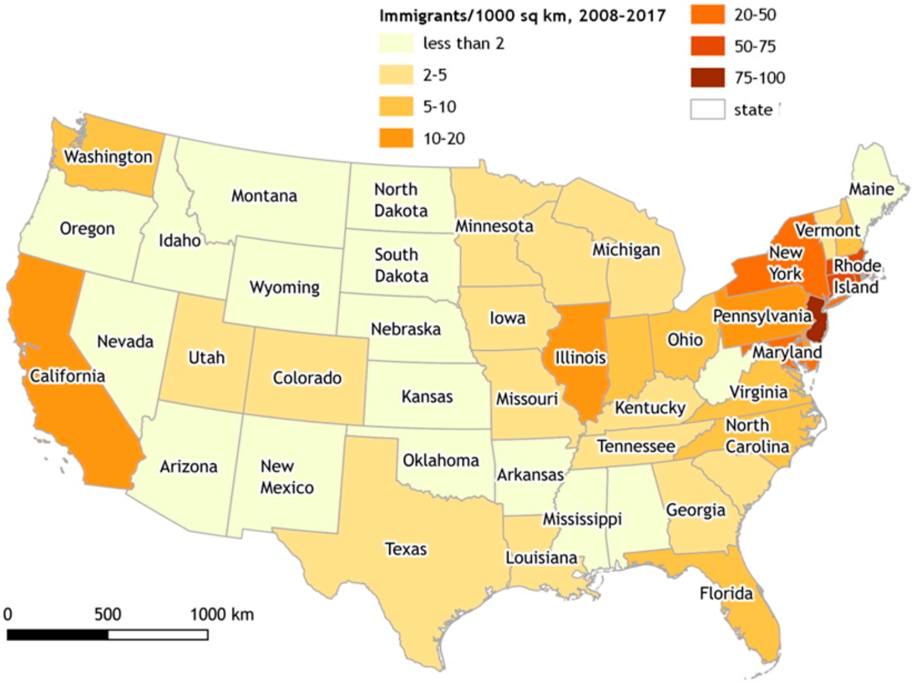

Figure 2. Immigrants' places of birth by states: those who migrated from the USA to Czechia between 2008 and 2017, including both permanent and temporary residence permits. Source: our own research.

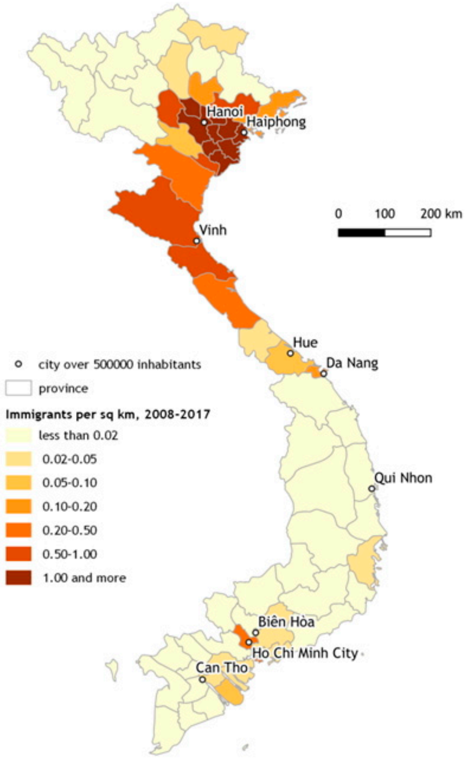

Figure 3. Immigrants' places of birth by provinces – those who migrated from Vietnam to Czechia between 2008 and 2017, including both permanent and temporary residence permits. Source: our own research.

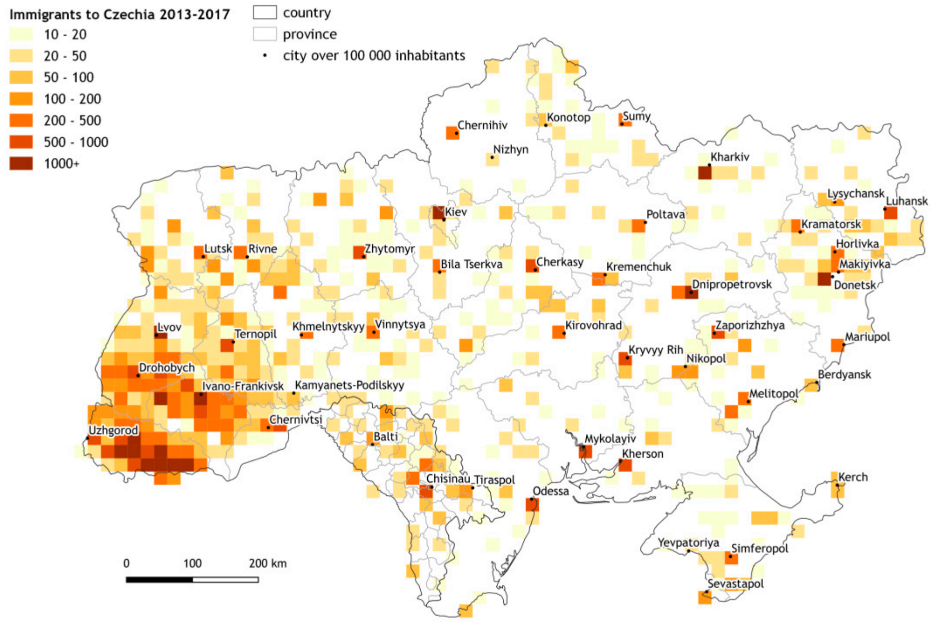

Figure 4. Spatial distribution of migrants into Czechia from Ukraine and Moldova between 2013 and 2017, including both permanent and temporary residence permits. Source: our own research.

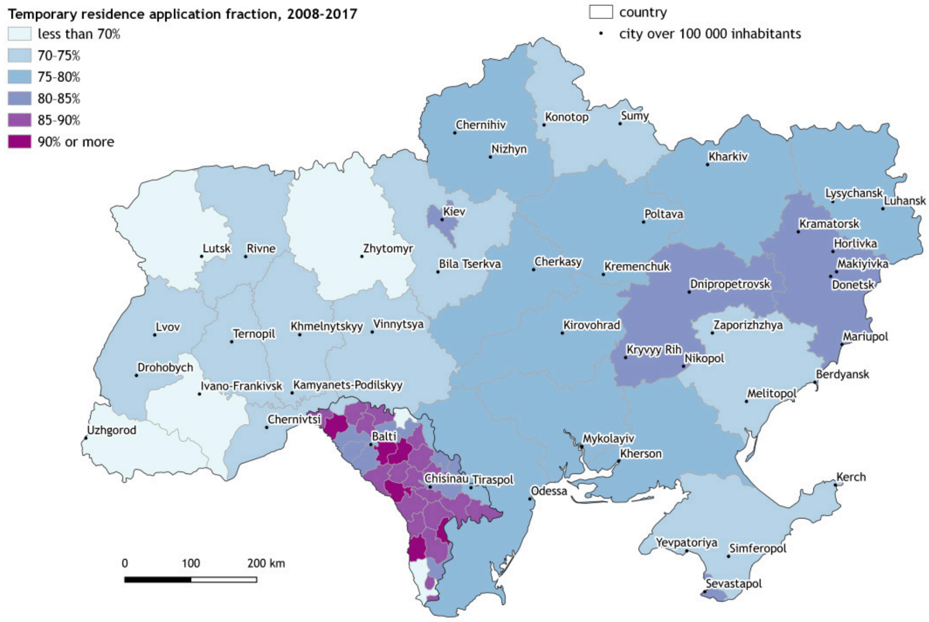

Figure 5. Fraction of temporary residence applications in the immigrant applications for Czechia from Ukraine (by oblast) and Moldova (by raion). Source: our own research.

People born in western Ukraine dominate among the Ukrainian migrants into Czechia (see Figure 4). Apart from the obvious advantage of spatial proximity, there are non-negligible cultural ties, namely in the southwestern oblast, e.g., for Transcarpathian Ukraine, which was a part of former Czechoslovakia during the 1920s and 1930s [42]. As a corollary, strong historical, cultural and psychological ties accompany the current outflow to destinations that were once situated in one common state. Although they have less migratory importance than the western region, some other individual sources, mostly administrative and industrial nodes like Kiev, Kharkiv, Dnipro, Donetsk, and Odessa are also worth mentioning.

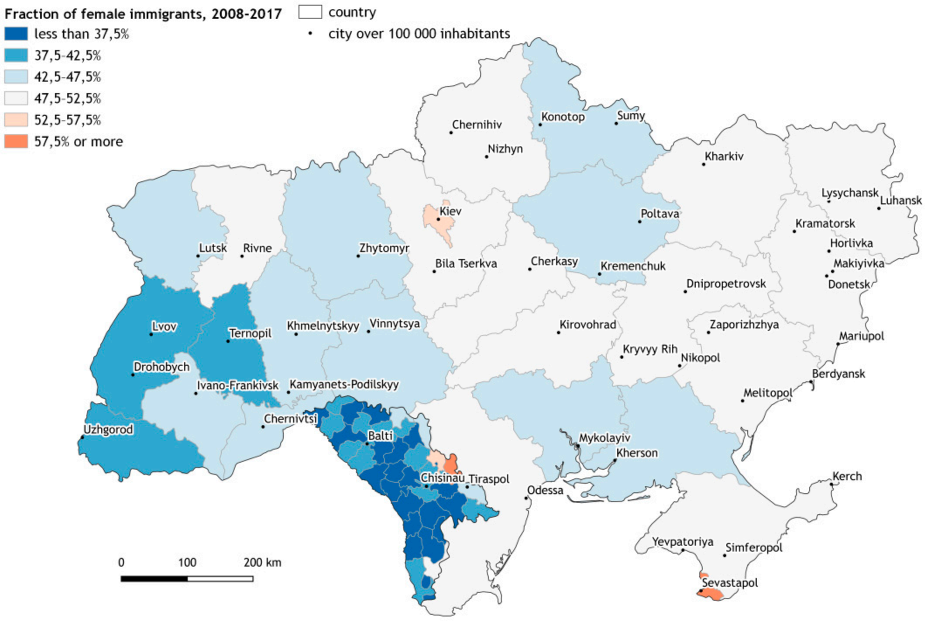

Figure 6. Fraction of female immigrants into Czechia from Ukraine (by oblast) and Moldova (by raion), including both permanent and temporary residence permits. Source: our own research.

The decline in the number of Ukrainians and Moldovans in Czechia came in the aftermath of the global economic crisis, which also hit the Czech economy (mainly in 2011–2013), simultaneously providing fewer economic opportunities for foreigners and causing the application of a more restrictive migration policy (see Figure 7). The current increase in demand for a foreign labour force has not yet compensated (until 2017) for the former losses.

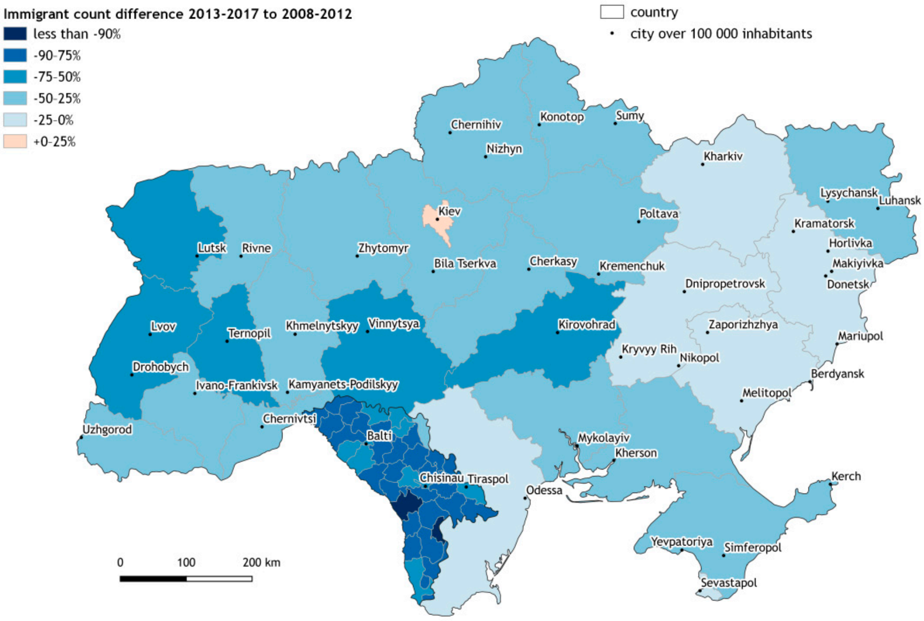

Figure 7. Difference in the count of migrants into Czechia from Ukraine (by oblast) and Moldova (by raion) in 2013–2017 compared to 2008–2012, including both permanent and temporary residence permits. Source: our own research.

See more information on various immigration and integration aspects of the respective immigration groups in Czechia, for example in [43–46].

4. Discussion

While our method is able to cope with most of the drawbacks outlined in part 2.2 that arise due to migrants coming from countries with different languages, spelling, and writing systems, resulting in a comprehensive Czech immigration dataset, it still suffers from a number of issues that need to be considered during its interpretation.

In a multitude of cases, the name of the administrative area is included in the original placename without any separator (e.g., "*UzinBelocerkovsky*", where "*Belocerkovsky*" designates a Ukrainian oblast). It is almost impossible to filter this out, as it may very well be an adjective of the placename itself. This often leads to the placename being geocoded to the capital of the administrative area as if only the administrative area part was given. Matching against a list of keywords for the most common cases could mitigate the issue.

The completeness of the GeoNames dataset was generally satisfactory for the purpose; however, more solid and complete results, especially in rural areas, could be obtained by data fusion with other open sources, such as Wikipedia [21] or OpenStreetMap [22].

Compared to edit distance, which is also used for fuzzy matching [47], the trigram similarity matching system that was used fares worse with shorter names. This proved detrimental with Vietnamese placenames, for which spaces are often inserted incorrectly. On the other hand, it is able to match words in a reversed order, which is also common for our Vietnamese data.

The performance of trigram matching degrades with very large tables, as is consistent with [46]; therefore, we had to limit the fuzzy search by the country of birth. This limitation sped the process up substantially but produced some misses where the given country and place of birth did not match, which was probably due to a misunderstanding at data input. An alternative would be to use a more powerful query engine such as Apache Lucene [5,19]. Given enough resources, the PostgreSQL backend could be replaced by any standard geocoding service; however, the high number of variants produced by the transcription engine means the process will probably be highly computationally demanding.

In a significant number of cases, only the name of the administrative area was recorded in the placename field. The aim was to geocode such cases to the location of the regional capital in order to provide at least some sense of association, and not to create fake settlements that would arise from the use of administrative area centroids. Unfortunately, because these rows often represent settlements so small that they cannot be named precisely; this skews the perceived urban/rural split of the immigrant population in favor of urban immigrants. This issue affects most gravely the USA where migration turnover and association with the state level are strong, but also Ukraine and to a lesser extent the other countries studied.

The main drawback of the method concept is the need to supply data-specific transcription rules. While some of them might work, for input that is not classified by language and/or country, more general and non-parametric methods such as neural networks might be more appropriate [28]. A neural network engine specifically focused on named entity linking, such as DeezyMatch [15], would be a strong contender. On the other hand, the transcription rules allow us to compensate for geocoding errors in a fine-grained fashion, and also make the process wholly interpretable, which would not be the case for neural network engines. The rules could also be at least partly generated automatically using phonological registers, achieving setup times comparable to the training times of the neural network engines.

Although Tables 1 and 2 show some accuracy metrics, comparing them straightforwardly to the figures reported by other researchers would be misleading, as the accuracy is not only determined by the geocoding method but also by the data and gazetteer used. Therefore, a numerical comparison to other studies was not performed. Potentially, a better comparison could be achieved by deploying a custom instance of the OpenStreetMap Nominatim search engine locally and intertwining it with the transcription rules; however, that is a very laborious task that was not undertaken due to technical difficulties. Furthermore, more online geocoding services could be included in the comparison.

Without applying our method, spatial patterns showing the places of birth of persons (in selected countries of origin) who have migrated to Czechia would still be visible. Nevertheless, the use of our method contributed to a more accurate picture. Crucially, without our attempt to improve the accuracy of the information, this data might never have been used. Regarding the migration issues, the importance of our method may have a great potential residing especially in solving more nuanced tasks, like studying family and/or community/neighborhood social networks within migration processes (via connections between the origin and destination).

Using only the birthplaces of migrants for their spatial identification has some limitations. It signals the type of environment in which the migrant was born, but not how long he or she spent in that particular place, nor does it reveal anything about his or her migratory history before coming to Czechia. However, the result is still a clear improvement over presenting the results only on the national level.

5. Conclusions

Using the geocoding method, we were able to geocode the noisy CIS immigration dataset with completeness and accuracy exceeding conventional services which are attuned to less noisy data [19]. The non-geocoded results can be safely discarded, as they do not exhibit any significant spatial association. The comparison with the online Geoapify geocoder shows that the engine has an overall advantage, especially in countries which do not use the Latin script.

The concern was that GeoNames may not contain the names of some very small settlements; this was not an issue. We can thus conclude that GeoNames is a fruitful source of placenames for noisy geocoding, confirming the previous study [20].

We selected several examples from the given data set, through which we demonstrated usefulness of our geocoding method. More specifically, we contributed to a more accurate picture informing us about places where immigrants to Czechia (in the given time period) were born. In the USA, the eastern and central-eastern parts together with large urban areas are the main sources. As for Vietnam, the northern part with several northern provinces—including Hanoi and Haiphong, along with their neighborhoods—dominated. In Moldova the spatial pattern is uniform in its high intensity, with males migrating more often than females. Ukrainian migration to Czechia created a specific spatial pattern in which mainly migrants born (and living) in the west of the country, especially Transcarpathian Ukraine, and also those from big administrative and industrial centres throughout the whole country head for Czechia.

Our analytical perspective also enabled us to see a development of the respective migration flows over time (between 2008 and 2017), reflecting important changes in the rate of migration due to the global economic crisis and its aftermath in Czechia (fewer working opportunities for foreigners and more restrictive policies). For more about migration and immigration patterns in Czechia, see [34,48].

We used the resulting dataset on several examples in order to demonstrate that the geocoding method has many potential uses for further research, and several questions may be asked on the spot:

- Why, in terms of migrants' birthplaces, have the encountered spatial patterns (or the absence thereof) arisen?
- What is the role of migration networks in their formation?
- What differences are there for different subnational cultures, ages, and genders?

On the other hand, the dataset can also be used to quantitatively verify similar questions and hypotheses arising from qualitative migration research.

The method itself, although an ad-hoc development for the problem at hand, might be applied to the geocoding of data from similarly noisy sources, especially where manual input from people with different spelling and transliteration conventions is encountered. Hence, one can imagine using our method not only to more precisely determine migrants' birthplaces but also, for example, when ascertaining migrants' latest, usual, or permanent residences before immigration to a particular destination country, or, by contrast, their preferred potential destinations abroad, etc. For example, during the ongoing Covid-19 pandemic, the method could be leveraged in order to automatically extract traveler routes where more sophisticated recording methods have not been put in place yet, especially if sub-national travel restriction conditions are enacted. A further example could be the automated correction and geocoding of placenames extracted by the automated speech-to-text transcription of oral testimonies, where placenames in a different language are recorded by a system tuned for the language used for the rest of the testimony. Further uses might be for databases in areas of state security, economy, social affairs, health, transportation, or science, as such may also contain similarly noisy data, e.g., names of individuals, organizations, or localities that could be geocoded using this method. An application of the transcription engine on nonspatial data could even be imagined, wherever the data are manually recorded by a person using a different orthography than the language they normally use.

Because the solution does not rely on third party services or commercial software and is easily deployable on commodity hardware, the solution has lower costs than comparable approaches. Thanks to its open-source implementation, the method can be easily adapted to different contexts, which is especially relevant given the sharp rise of user-generated content [49]. The transcription engine has a clearly defined interface and could thus also be coupled with a different geocoding backend, such as OpenStreetMap's Nominatim or commercial engines, if they provide sufficient capacity.

Finally, an improvement of the data lies not only in applying a new geocoding method but also in putting the current process of data creation into sharp relief. It is clear from our work that if migration into Czechia is to be effectively documented, the officials must check the applications more thoroughly and explain more. On the other hand, the applicants must be more willing to cooperate, and the process rules (especially with regard to form filling) have to be clearer and more specific. The digitalization of the process would allow for the usage of gazetteer-based autosuggestion, which can be expected to lower the error rate considerably.

After the current pandemic situation, a new post-pandemic era will come. Nevertheless, states throughout the whole world will strive, via their policies and practices, to reduce the risk of being hit by other possible pandemics. It will also be carried out by partially limiting arrivals to their territories, and through the more intensive monitoring and controlling population movements, be it those tied to business trips abroad and from abroad; short-term, long-term, permanent international migration; or tourism. For this purpose, new databases (mostly at the national or other regional–hierarchical levels) will be established (or the existing ones will be newly designed) with regard to how to better register and record movements, and especially related to the places visited by the domestic population as well as foreigners. Paradoxically, the partial de-globalization of the already highly diversified world will be accompanied by its more intensive spatial monitoring and recording. This is an opportunity in which our tool might be successfully applied.

Supplementary Materials: The code is available online at http://github.com/simberaj/migration-geocode.

Author Contributions: Conceptualization, Jan Šimbera, Dušan Drbohlav, Přemysl Štych; Methodology, Jan Šimbera, Dušan Drbohlav, Přemysl Štych; Software, Jan Šimbera; Validation, Jan Šimbera, Dušan Drbohlav, Přemysl Štych; Formal Analysis, Jan Šimbera; Investigation, Jan Šimbera, Dušan Drbohlav; Data Curation, Jan Šimbera, Dušan Drbohlav; Writing—Original Draft Jan Šimbera, Dušan Drbohlav; Writing—Review and Editing, Dušan Drbohlav, Přemysl Štych; Supervision, Dušan Drbohlav, Přemysl Štych; Funding Acquisition, Jan Šimbera, Dušan Drbohlav, Přemysl Štych. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Ministry of Culture of the Czech Republic, programme NAKI II, Project no. DG18P02OVV065: "Living map: Topography of the History of Natural Sciences in the Czech Lands (Živá mapa: Topografie dějin přírodních věd v Českých zemích)".

Acknowledgments: We would like to thank to the Ministry of Culture of the Czech Republic, programme NAKI II, Project no. DG18P02OVV065: "Living map: Topography of the History of Natural Sciences in the Czech Lands (Živá mapa: Topografie dějin přírodních věd v Českých zemích)". We would like to thank the anonymous reviewers for their substantial help with improving our manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fassmann, H. European migration: Historical overview and statistical problems. In *Statistics and Reality; Concepts and Measurements of Migration in Europe*; Fassmann, H., Reeger, U., Sievers, W., Eds.; University Press: Amsterdam, The Netherlands, 2008; pp. 21–43.
2. McHugh, K.E. Explaining migration intentions and destination selection. *Prof. Geogr.* **1984**, *36*, 315–325. [CrossRef]
3. The Open Geospatial Consortium (OGC). Reference Model. Version 2.1. 2011. Available online: http://www.opengis.net/doc/orm/2.1 (accessed on 24 January 2021).
4. Sanderson, M.; Kohler, J. Analyzing Geographic Queries. In Proceedings of the 27th Annual International ACM SIGIR Con-ference, Sheffield, UK, 25–29 July 2004; pp. 37–39.
5. Valkanas, G.; Gunopulos, D. Location Extraction from Social Networks with Commodity Software and Online Data. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium, 10–13 December 2012; pp. 827–834.
6. Densham, I.; Reid, J. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*; Association for Computational Linguistics (ACL): Edmonton, AB, Canada, 2003; pp. 79–80.
7. Huck, J.; Whyatt, D.; Coulton, P. Challenges in geocoding socially-generated data. In *Proceedings of the GIS Research UK 20th Annual Conference: Volume 1—Presentations*; Whyatt, D., Rowlingson, B., Eds.; Lancaster University: Lancaster, UK, 2012; pp. 39–45.
8. Derczynski, L.; Maynard, D.; Rizzo, G.; van Erp, M.; Gorrell, G.; Troncy, R.; Petrak, J.; Bontcheva, K. Analysis of named entity recognition and linking for tweets. *Inf. Process. Manag.* **2015**, *51*, 32–49. [CrossRef]
9. Hirschmann, L.; Chinchor, N. MUC-7 coreference task definition. In Proceedings of the MUC-7 Conference, Fairfax, VA, USA, 19 April–1 May 1997.
10. Rao, D.; McNamee, P.; Dredze, M. Entity linking: Finding extracted entities in a knowledge base. In *Multi-Source, Multilingual Information Extraction and Summarization*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 93–115.
11. Amitay, E.; Har'El, N.; Sivan, R.; Soffer, A. Web a where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*; ACM: New York, NY, USA, 2004; pp. 273–280.
12. Ardanuy, M.C.; Sporleder, C. Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*; Association for Computing Machinery (ACM): New York, NY, USA, 2017; pp. 175–180.
13. Brando, C.; Frontini, F.; Ganascia, J.-G. Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets. In *Advances in Service-Oriented and Cloud Computing*; Metzler, J.B., Ed.; Springer: Cham, Switzerland, 2015; pp. 505–514.
14. Kim, J.; Vasardani, M.; Winter, S. Similarity matching for integrating spatial information extracted from place descriptions. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 56–80. [CrossRef]
15. Ardanuy, M.C.; Hosseini, K.; McDonough, K.; Krause, A.; van Strien, D.; Nanni, F. A Deep Learning Approach to Geographical Candidate Selection through Toponym Matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*; Association for Computing Machinery (ACM): New York, NY, USA, 2020; pp. 385–388.
16. Li, H.; Srihari, R.; Niu, C.; Li, W. InfoXtract location normalization: A hybrid approach to geographic references in in-formation extraction. In Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References, Stroudsburg, PA, USA, 10–13 July 2003; pp. 39–44.
17. Overell, S.; Rüger, S. Using co-occurrence models for placename disambiguation. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 265–287. [CrossRef]
18. GeoNames. GeoNames. Available online: http://geonames.org/ (accessed on 28 February 2019).
19. Mattmann, C.A.; Sharan, M. An Automatic Approach for Discovering and Geocoding Locations in Domain-Specific Web Data (Application Paper). In Proceedings of the 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA, USA, 28–30 July 2016; pp. 87–93.
20. Ahlers, D. Assessment of the accuracy of GeoNames gazetteer data. *Proc. Python High-Perform. Sci. Comput.* **2013**, 74–81. [CrossRef]
21. Ahlers, D. Applying Geographic Information Retrieval. *Datenbank-Spektrum* **2014**, *14*, 39–46. [CrossRef]
22. Chow, T.E.; Dede-Bamfo, N.; Dahal, K.R. Geographic disparity of positional errors and matching rate of residential addresses among geocoding solutions. *Ann. GIS* **2015**, *22*, 29–42. [CrossRef]
23. Liao, Y.; Wang, J. A method for matching Chinese place-name data. In *Proceedings of the Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Advanced Spatial Data Models and Analyses*; International Society for Optics and Photonics: Washington, DC, USA, 2009; Volume 7146, p. 71461.

24. Coetzee, S.; Rademeyer, M. Testing the spatial adjacency match of the Intiendo address matching tool for geocoding of ad-dresses with misleading suburb or place names. In Proceedings of the 24th International Cartography Conference, Santiago, Chile, 15–21 November 2009; pp. 10–18.
25. Lan, T.; Longley, P. Lan Geo-Referencing and Mapping 1901 Census Addresses for England and Wales. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 320. [CrossRef]
26. Daras, K.; Feng, Z.; Dibben, C. HAG-GIS: A spatial framework for geocoding historical addresses. In Proceedings of the 23rd GIS Research UK Conference, Leeds, UK, 15–17 April 2015; pp. 3–6.
27. Singh, S.K. Evaluating two freely available geocoding tools for geographical inconsistencies and geocoding errors. *Open Geospat. Data Softw. Stand.* **2017**, *2*, 11. [CrossRef]
28. Arbabi, M.; Fischthal, S.M.; Cheng, V.C.; Bart, E. Algorithms for Arabic name transliteration. *IBM J. Res. Dev.* **1994**, *38*, 183–194. [CrossRef]
29. Cui, Y. A systematic approach to evaluate and validate the spatial accuracy of farmers market locations using multi-geocoding services. *Appl. Geogr.* **2013**, *41*, 87–95. [CrossRef]
30. Ahlers, D.; Boll, S. Adaptive geospatially focused crawling. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management–CIKM '09*; ACM: New York, NY, USA, 2009; pp. 445–454.
31. Karimi, H.A.; Sharker, M.H.; Roongpiboonsopit, D. Geocoding Recommender: An Algorithm to Recommend Optimal Online Geocoding Services for Applications. *Trans. GIS* **2011**, *15*, 869–886. [CrossRef]
32. Ping, D.; Yong, L. Building placename ontology to assist in geographic information retrieval. In Proceedings of the Computer Science Technology and Applications, IFCSTA'09, International Forum, IEEE Computer Society, Washington, DC, USA, 25–27 December 2009; Volume 3, pp. 306–309.
33. MVČR. Formuláře a Žádosti. Praha: Odbor Azylové a Migrační Politiky, Ministerstvo Vnitra ČR. Available online: http://www.mvcr.cz/clanek/formulare-zadosti.aspx (accessed on 28 December 2019).
34. MVČR. Některé Náležitosti Žádosti. Praha: Odbor Azylové a Migrační Politiky, Ministerstvo Vnitra ČR. Available online: https://www.mvcr.cz/clanek/obcane-tretich-zemi-nektere-nalezitosti-zadosti.aspx (accessed on 28 December 2019).
35. Foreigners, Total by Citizenship as at 31 December 2017. Directorate of the Alien Police Service, Czech Republic. Available online: https://www.czso.cz/documents/11292/27914491/1712_c01t01.pdf/ff9e9fee-08d3-4bdc-a11b-d0cc1e3ac184?version=1.0 (accessed on 29 January 2021).
36. CZSO. Foreigners: Number of Foreigners. Available online: https://www.czso.cz/csu/cizinci/1-ciz_pocet_cizincu (accessed on 28 December 2019).
37. Šimbera, J. Python-Based Placename Geocoder for Noisy, Badly Transcripted and Erroneous Data from Migration Geodata-bases. 2018. Available online: http://github.com/simberaj/migration-geocode/ (accessed on 28 December 2019).
38. Korotkov, A.; Zakirov, A. Fuzzy Substring Searching with the pg_trgm Extension. Available online: https://dl.acm.org/citation.cfm?id=1463460 (accessed on 15 November 2019).
39. Nominatim Geocoding Service: About & Help. Available online: https://nominatim.openstreetmap.org/ui/about.html (accessed on 2 April 2021).
40. Geoapify.com geocoding. Available online: https://www.geoapify.com/ (accessed on 1 February 2021).
41. Brando, C.; Frontini, F.; Ganascia, J.-G. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Syst. Inf. Model. Q.* **2016**, *7*, 60–80. [CrossRef]
42. Halemba, A. Not looking through a national lens? Rusyn–Transcarpathians as an anational self-identification in contemporary Ukraine. In *Debatten um Polen und Polentum in Geschichte und Gegenwart, Polen: Kultur–Geschichte–Gesellschaft 1*; Brückner, A., Ed.; Wallstein Verlag: Göttingen, Germany, 2015; pp. 123–146.
43. Šimon, M.; Křížková, I.; Klsák, A. Immigrants in large Czech cities 2008–2015: The analysis of changing residential patterns using population grid data. *Geografie* **2020**, *125*, 343–374. [CrossRef]
44. Ignatyeva, E.; Sýkora, L. Strangers among their own: Local interaction, integration, and segregation of Russian immigrants in Prague. *Geografie* **2019**, *124*, 341–364. [CrossRef]
45. Klvaňová, R. *The Brother of the Other. Immigration from Belarus, Russia, and Ukraine to the Czech Republic*; EDIS Publication Series; Masaryk University, MUNI Press: Brno, Czech Republic, 2017; 167p, Volume 16.
46. Freidengerová, T. Vietnamci v Česku a ve světě. Migrační a Adaptační Tendence. Praha, Sociologické Nakladatelství (SLON), Prague. 2014, p. 232. Available online: https://sreview.soc.cas.cz/artkey/csr-201604-0001_living-together-in-an-urban-neighbourhood-the-majority-and-vietnamese-immigrants-in-prague-libus.php (accessed on 1 May 2021).
47. Angel, A.; Lontou, C.; Pfoser, D.; Efentakis, A. Qualitative geocoding of persistent web pages. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems–GIS '08*; ACM: New York, NY, USA, 2008; p. 10.
48. Drbohlav, D.; Medová, L.; Čermák, Z.; Janská, E.; Čermáková, D.; Dzúrová, D. *Migrace a Migranti v Česku. Kdojsme, Odkud-Přicházíme, kamJdeme?* Sociologické nakladatelství (SLON): Prague, Czech Republic, 2010; p. 184.
49. McKenzie, G.; Slind, R.T. A user-generated data based approach to enhancing location prediction of financial services in sub-Saharan Africa. *Appl. Geogr.* **2019**, *105*, 25–36. [CrossRef] [PubMed]