urban science

MDPI

Article

Street Name Data as a Reflection of Migration and Settlement History

Sarah J. Berkemer [1,2,3,*] and Peter F. Stadler [1,2,4,5,6,7,8,*]

1 Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstr 16-18, 04107 Leipzig, Germany
2 Competence Center for Scalable Data Services and Solutions Dresden/Leipzig, 04109 Leipzig, Germany
3 Image and Signal Processing Group, Department of Computer Science, Universität Leipzig, 04109 Leipzig, Germany
4 Centre for Integrative Biodiversity Research (iDiv), and Leipzig Research Center for Civilization Diseases, Universität Leipzig, 04109 Leipzig, Germany
5 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
6 Facultad de Ciencias, Universidad National de Colombia, Sede Bogotá, Ciudad Universitaria, Bogotá D.C. 111321, Colombia
7 Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria
8 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
* Correspondence: bsarah@bioinf.uni-leipzig.de (S.J.B.); studla@bioinf.uni-leipzig.de (P.F.S.)

Received: 23 November 2020; Accepted: 8 December 2020; Published: 11 December 2020

Abstract: Street names (odonyms) play an important role not only as descriptors of geographic locations but also due to their sociological and political connotations and commemorative character. Here we analyse street names in Europe and North America extracted from OpenStreetMap, asking in particular to what extent odonyms reflect early European settlements in the New World, i.e., the immigration of German, Austrian and Scandinavian minorities. We observe that old street names of European origin can predominantly be found in rural areas. North American street names indeed recapitulate local and regional settlement histories. The aim of this study is to demonstrate that easily accessible data sets from freely available map data such as street names convey usable information concerning migration patterns and the history of settlements in the case of European immigrants in North America as well as colonial history. We provide a freely available pipeline to analyse this kind of data.

Keywords: street names; OpenStreetMap; geographic comparison

1. Introduction

The history of North America is in large parts driven by huge migration waves. After the arrival of the first Europeans, settlements, cities, and states were created and developed based on cultures, religions and national identities of their founders. Woodard [1] describes in great detail how groups of early settlers formed several societies centered around various principles and beliefs. These groups developed the "dominant cultures" and shaped the basis for the "eleven American nations" that can still be found today and are responsible for the difficulties of making the United States a unified nation based on common concepts and values [1,2].

Three large immigration waves from mostly Europe to the North American continent can be identified between 1830 and 1924 [1]. The first wave between 1830 and 1860 bringing mainly Irish, German and British people was followed by the second and even larger wave until 1890 where people from the afore-mentioned countries together with Scandinavians and Chinese arrived at the east coast. The third and largest wave brought people from mainly southern and eastern Europe such as Italy,

Greece and Poland consisting to a large part of Catholics and Jews between 1890 and 1924. European immigrants arrived at the east coast and from there spread across North America leading to early settlements in the eastern states and only later to a migration to the western and southern parts. At the same time Spanish settlers arrived in the Mexican area and moved up north reaching the southern part of today's US area, a region also described by the name "El Norte" in Woodard [1].

European migrants included their own cultures, religions and languages when founding new settlements. Hence, place, street and city names reflected the countries of origin as well as the names, religion or culture of their founders [3,4].

Place names have been shown to clearly reflect migration patterns and historical population profiles. Naming and renaming of place names has various purposes [5,6]. As listed by Nash [7], changing place names can be part of capitalist modernization [8,9], colonial settlement [10], state formation [11], national independence [12] or official commemoration [13]. This practices have mainly been used in colonial times by Europeans to claim the new territories and erase former cultures and collective identities of the indigenous population [7] but also for identification purposes and as a means of forming and shaping settlements in the "new world" as their home [3,4].

Today, street names (odonyms) primarily serve as a means to uniquely determine locations of private houses, businesses and public spaces as well as to ensure basic supplies such as water and electricity. Street naming practices go back to medieval Europe where at first, street names solely fulfilled functional roles. Since then, most streets underwent numerous renaming events, reconstructions and changes during history. As mentioned in Badariotti [14], five different epochs in Europe are mirrored within street naming practices:

I. During medieval times street names mainly fulfilled functional roles pointing out the usage of a certain space such as social, institutional or industrial/business zones (see also Algeo [15]).
II. Commemorative street names became popular during the 17th and 18th century serving as glorifications of popular and mighty personalities such as the king.
III. During the time of the French revolution, the first large street renaming events occured where mainly names referring to religious terms were changed to terms popular in revolutionary thoughts, such as philosophers and non-religious values.
IV. Napoleon's era then reestablished old names and introduced street names referring to the Empire, victories, battles, officer's names and other military terms.
V. Street names in the 19th and 20th centuries are mostly based on commemorative street names of well-known people, geographic terms for certain locations or local areas with specific themes such as plants' or birds' names.

However, until today, changes of regimes also lead to renaming of mainly commemorative street names [13,16–19]. In recent history, streets were renamed extensively after World War I. In the US mainly German street names were removed [20]. Also in Europe, naming of streets is considered to have political and historical impact, i.e., as shown in the case of West and East Germany [21]. Renaming also occured within communities of mixed descent such as English and Hispanic cultures in New Mexico. Here, many street names changed from a Spanish to an English name [15].

Based on the street naming epochs and today's usage of themes among street names, phases of growth of a city and the construction of new suburbs can be traced back regarding street names [14]. This also includes street names reminding of buildings or landmarks that have been located in a certain area, such as former schools, churches, rivers, or communities of people.

Exceptions to such street names can be seen in North America, where many streets are given a number depending on their location to a reference street. In Managua, the capital of Nicaragua, no street names are used at all. Instead, directions and reference points inside the city are used to describe locations [14].

OpenStreetMap (OSM, osm.org) is an online database providing geographic data and maps of cities, for routing, public transport, hiking or cycling. The data is freely available online and users

can add and edit information, similar to Wikipedia (wikipedia.org). In this contributiion, we analyse street names retrieved from OSM to investigate to what extent this type of easily accessible data can be harnessed to gain information into settlement and migration histories. We focus here on European settlements in North America and street name patterns in Europe, mainly because the corresponding data are most abundant and most easily interpretable to us. We then briefly investigate reflections of Africa's colonial history.

Despite a large number of newly created streets due to fast growing cities and many renaming events, we observe that there are still easily detectable traces of very old and conserved street names in Europe and North America. Our results of street name comparison and resulting clusters of European origin in North America agree with results obtained by Han et al. [22] and Woodard [1]. We observe that the oldest and most conserved European clusters can be found in rural areas which are less prone to reconstruction and renaming of streets. The work by Woodard [1] states the historical developments of the "11 American nations" that were formed by the first waves of immigration to North America. The results by Woodard [1] are confirmed in the study by Han et al. [22], which analyses reveil genetic diversity and ancestry of 770,000 genomes in the US. Similarly, French influences in northern and central Africa as well as Dutch influences in South Africa are consistent with the regions' colonial history, even though the OSM street name data for the African continent are comparably sparse. Clearly, the analysis of street name data cannot replace a detailed historical analysis. Nevertheless, it serves to provide an easily accessible overview and is potentially useful as a means of generating hypotheses.

2. Materials and Methods

We consistently used street names downloaded and extracted in February 2020 (Europe and North America) and September 2020 (Africa) in the language that OSM defines as mainly spoken language in that area. No translations were applied. For further details on Materials and Methods see the Supplementary Material.

2.1. Data Extraction and Curation

Geographic data has been downloaded from geofabrik (download.geofabrik.de, February 2020, September 2020). The data were extracted and converted using osmium tool (osmcode.org). Street names have been extracted filtering the data by the tag *highway*, which defines the category of all types of streets in OSM. Streets are equipped with geographic coordinates which have been used for the analyses. Street entries in OSM are assigned a street type. In this analysis, we only include streets with types primary, secondary, tertiary, residential, pedestrian and unclassified. In this way, highways and motorways, which are typically named just by numbers, are excluded.

2.2. Extraction of Street Terms from Street Names

The extraction of street terms is done for each country separately and aims to split street names into specific and generic terms. Numbered streets are excluded from the analysis in order to be able to extract street names of European origin.

In many countries, street names are composed of two parts, a *specific* term that makes the street name unique in a certain area and a *generic* term, referring to the kind of street, such as *road*, way or *boulevard* [14]. In numerous cases, a third term is added which denotes the direction or orientation of the street such as east and west or upper and lower.

Splitting street names into single terms is done in a two-step process as depicted in the workflow in Figure 1. Most of the street names in the data set are composed out of two or three terms concatenated by a space character (' '). Even in countries where street terms are usually concatenated into one single word, we find street names separated by a space character or by a dash ('-'). In order to extract and split street terms in street generics and specifics, street names are first split into terms by space and dash characters. This results in a mixed multi-set $M = T \cup N$ of street terms T and street names N. Note that the same street name usually has many occurences in a country and thus, also in M, however,

each street is uniquely defined by its ID, i.e., we treat M as multiset. This process is also described by an example in Figure 1. Here, the analysis starts with two English street names (*hill street, pine road*) and two German street names (*Bergstraße, Hans-Meier-Straße*, English translation: hill street and Hans Meier street). In a first step, the street names are decomposed into terms by only splitting at space and dash symbols. This results in the mixed set of single terms such as *street* and complete street names for entries such as *Bergstraße*, which consists of just a single word.

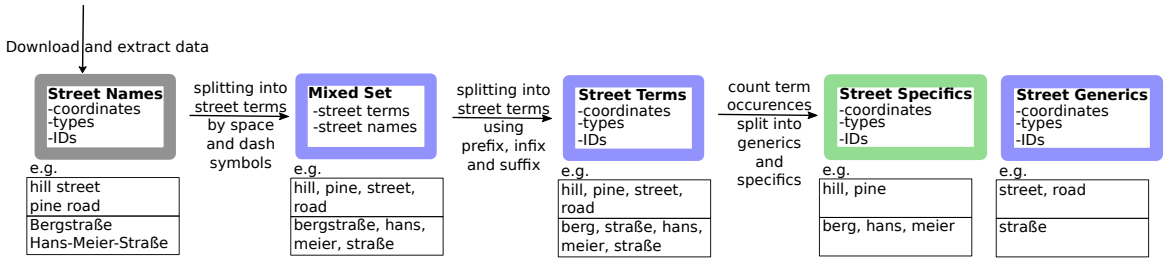

Figure 1. Workflow describing the data extraction and preparation in order to create a set of street specifics and street generics for European and North American countries and states. Street names (grey) are split into specific and generic street terms in a two-step process (middle blue). The lower part of the figure shows examples for English and German street names and how they are split into specific and generic terms. In order to find exact matches, all street names and terms are converted to lower-case symbols. Resulting street specifics (green) will further be used to map street terms to their countries of origin, see also Figure 2. For more details, see description in the text and supplemental data.

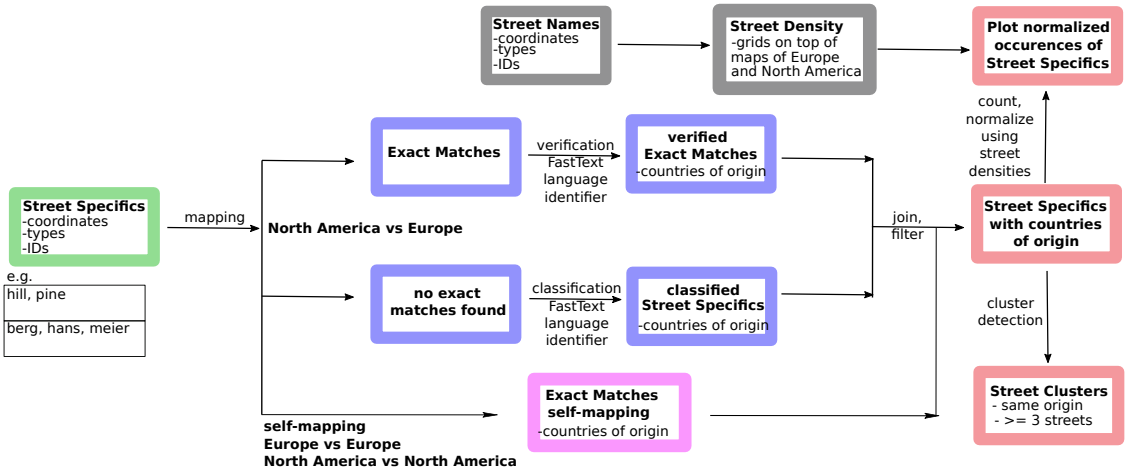

Figure 2. Workflow describing the analysis of specific street terms in Europe and North America (green). In order to classify American street specifics, exact matches to European street terms are used. Resulting exact matches are verified with a language identifier whereas the language identifier is used to classify countries of origin for street terms where no exact matches could be found (blue). In case of self-mappings of Euopean street terms to European street names and analogously for North America, only exact matches are used (pink). Plotting occurences of streets of common origin (upper red) is done by normalizing counts using street density data (grey). Resulting street specifics are also used for the detection of street clusters (lower red). See description in the text for more details.

In a second step, we now compare each entry $m \in M$ to all other entries $m' \in M$, $m' \neq m$ and check if m occurs as a proper prefix, infix or suffix of m' or if m and m' are equal. If they are equal, the term is counted occuring as 'complete term', thus it is part of a street name that could be split by a space or dash sign. A proper prefix, infix or suffix is here defined such that at least 2 letters of the original word m' remain on both sides (infix) or as suffix or prefix, respectively. In the example in Figure 1, the German street name *Bergstraße* will be split into the prefix *Berg* (hill) and the suffix *Straße* (street) by comparing the complete term *Straße* to the street name *Bergstraße* to identify the proper

prefix and suffix. For each $m \in M$, the number of matchings as prefix, infix, suffix or complete term is counted as well as the total number and percentage of appearances in the data set in comparison to the total number of street terms (not street names!) in the data set. In this way, the composition of street names in a certain country is analysed and terms appearing in a much higher frequency than others are identified as street generics. More detailed information and examples can be found in the supplemental data.

2.3. Calculating the Street Density in North America and Europe

As depicted in Figure 2, data for street densities are used in order to normalize counts of streets classified to having the same country of origin. We use a grid based on geographic coordinates and count street entries within each cell of the grid. Each cell c_{ij} covers an area of 0.5° in longitude and latitude The grid is set in the region from 130° W to 60° W and 15° N to 55° N for North America, 30° W to 30° E and 35° N to 70° N for Europe and 25° W to 60° E and 36° N to 37° S for Africa, respectively. For the Africa data a coarser grid with cells covering an area of 1° in longitude and latitude is used due to the sparsity of named streets in the OSM data set. For each cell in the grid we count the number of streets n_{ij} intersecting the cell. Thus, for each street, we check in which grid of the cell it is located by looking at the street coordinates. As each street is described by several coordinates, we calculate a midpoint of the coordinates which is used when counting the streets. We compute $d_{ij} = \ln(n_{ij}/N)$, where $N = \sum_i \sum_j n_{ij}$ is the total number of streets in the area of interest, i.e., the density values d_{ij} for North America are shown in Figure 3. Density plots for streets in Europe and Africa can be found in the Supplementary Materials. Plots have been created using `Python matplotlib/basemap` [23].

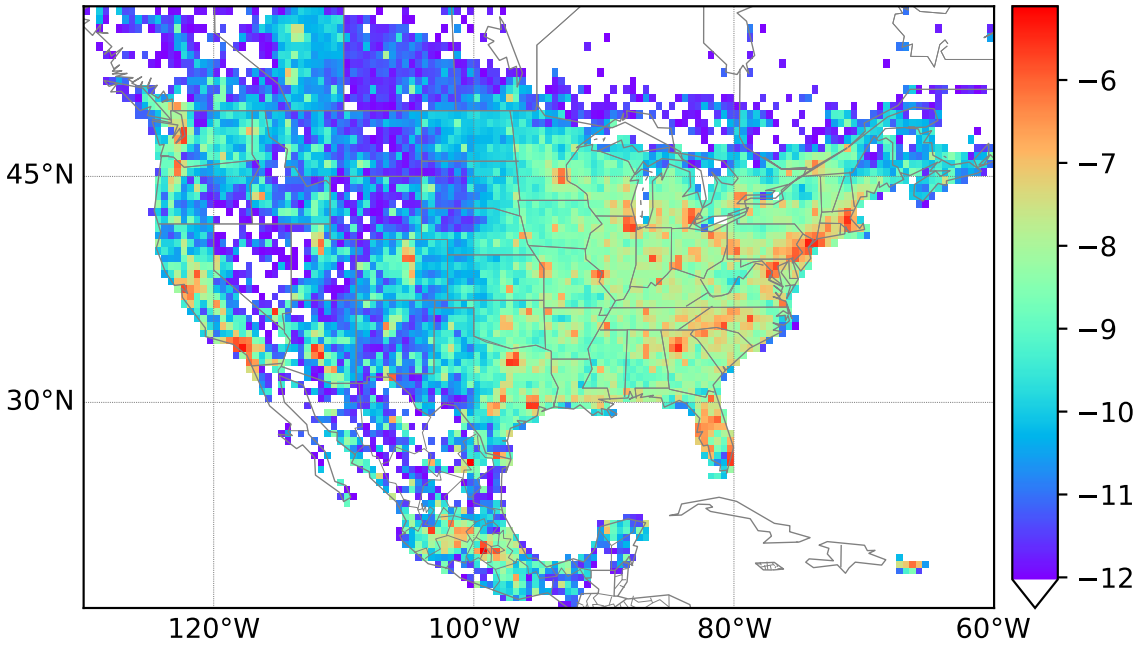

Figure 3. Heatmap of North America showing the density of streets. Colors indicate the density of streets in this area, thus the logarithmic normalized number of streets per cell of the grid. Higher values and red colors show a higher density of streets. Each cell c_{ij} covers an area of 0.5° in longitude and latitude, for which we count the number of streets n_{ij} intersecting the cell.

2.4. Classification of Street Terms and Names

After the extraction of street terms, we now have a data set of street terms with total number and percentage of appearance for each European country and North American state. As depicted in Figure 2, we only use the set of street specifics obtained after data extraction and preparation

as described above and in Figure 1. To classify North American street terms by their country of origin, we count exact matches of North American street specifics with European street specifics. Thus, for each North American street term, we obtain a list of putative European countries of origin. Approximately 2/3 of the street terms in North America can be assigned to a European correspondent by an exact match. In order to verify assigned countries of origin and classify terms that have not yet been classified, we apply the `FastText` language identifier [24,25] that has been trained on Wikipedia data. For each street term, we extract the three most probable languages and corresponding countries, i.e., if the language is German, possible countries are Germany, Austria and Switzerland.

For most of the street terms, it was possible to correctly verify the country of origin using the language identification tool. The results agreed to more than 80% with the results obtained by matching street terms of American countries to street terms in Europe. Difficulties occured for very similar languages and countries with the same official language. Of course, also classifications based on street term matching might not always reflect the term's language of origin. However, our aim is to detect clusters of North American street names based on European street names. The highest amount of misclassified street terms can be seen when the street term (specific) is a person's or a city's name, i.e., *Lindbergh, Lionel, Christie, Frazer, Mozart, Dessau, Marionville, Amalfi*. Here, matching of the street terms results in different origins than using the language identifier tool. For only a few terms, no language identification is possible. These terms are excluded from further analyses. For further details we refer to the supplemental data.

Final assignments of European countries to North American street terms are based on the intersection of street term matchings and results of the language identifier. If the intersection is empty, we remain with the street term matching. If the intersection contains more than one country, we keep all assignments in the intersection. North American street terms that could not be matched to European street terms are only classified by the language classifier. We see that in total, street generics are less accurately identified by the language identifier as the sets contain more abbreviations, e.g., 'dr' for drive. However, only street specifics are included in the cluster detections.

Taken together, the final assignments of North American street specifics to their European country of origin reveils that many street names can possibly be mapped to several countries of origin. In order to avoid too generic street names, we only take into account street names that have been assigned to at most four European countries of origin. This number especially accounts for European countries sharing the same language such as Germany, Austria and Switzerland being German-speaking countries and Belgium, Luxembourg and Netherlands having German minorities.

We not only calculate mappings of European street specifics to North American street names but also create self-mappings, thus mappings of European street specifics to European street names and analogously for North America. These mappings are only based on exact matches of street specifics but omit the verification and classification step using the language classifier (see also Figure 2).

In an additional step, we downloaded street name data for the African continent (September 2020). Their analysis is conducted analogously to the analysis of North American street names. However, much less data is available and thus, we obtain a smaller amount of information. Still, the former colonial powers can in most cases be identified based on street names with European origins.

2.5. Construction of Street Clusters Based on European Origin

Dense regions of streets named by the same European origin are used to detect clusters of streets. Each cluster has to contain at least three streets with the same European origin. Only regions with a log density value of larger than -4 are taken into account. In this way, no distance threshold between the streets has to be defined as distances might be much larger in rural areas than in cities.

2.6. Code Availability

All scripts used to analyse the data sets and lists of data sources are freely available and can be found in the corresponding github repository, see www.github.com/bsarah/osm-streetnames.

3. Results

3.1. Regions with Higher Density of European Street Names

We identify regions in North America by plotting the density of street names with common European origin. This has been done for the main language groups English, Spanish, French and German. The results agree with the expected and known distribution of currently spoken languages and historical data about European settlements in North America [1,5,22].

The mainly spoken language in North America is English and thus, the distribution of English street names covers most of North America. Exceptions are francophone regions in Canada, Mexico, and the parts of the US that once were part of Spain or Mexico, i.e., the southwestern states (in particular NM, AZ, CA, CO, and NV), Florida, and Puerto Rico, see Figure 4.

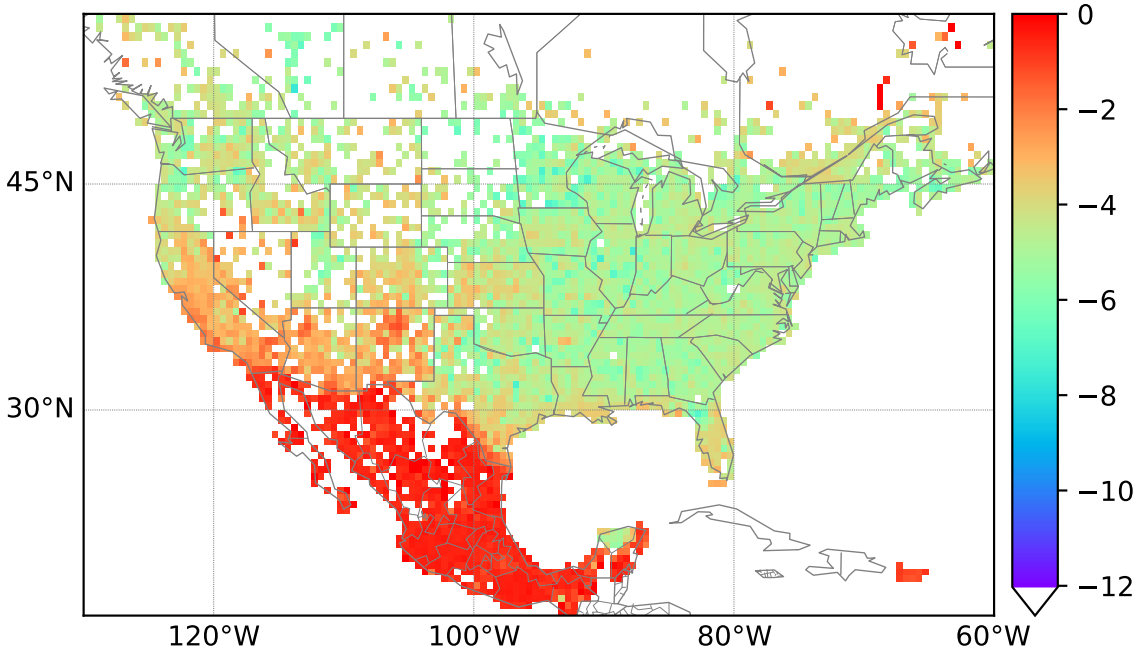

Figure 4. Heatmap of North America showing the density of streets classified as Spanish. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

The distribution of street names classified as French agrees with the historical distribution of French speakers based on Woodard [1] and the genetic distribution by Han et al. [22], with the main regions being Quebec and southern Louisiana as shown in Figure 5. The density plot additionally shows the Francophone regions in Saskatchewan, also called Fransaskois, which go back to migration of French Canadians from Quebec to the West [26]. German settlers were mainly distributed throughout the former regions located south of the Great Lakes, thus the states of Minnesota, Wisconsin, Illinois, Indiana, Michigan, Ohia and Pennsylvania as described by Woodard [1], Fuchs [6] and depicted in Figure 6. Table 1 shows numbers of streets that are used in the plots.

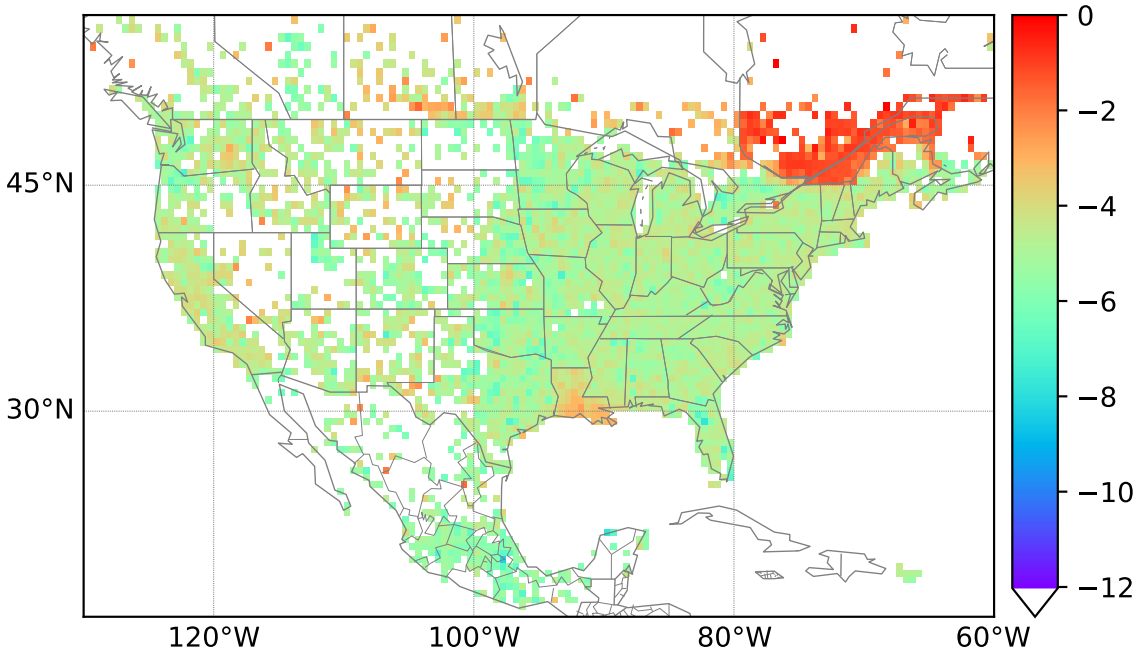

Figure 5. Heatmap of North America showing the density of streets classified as French. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

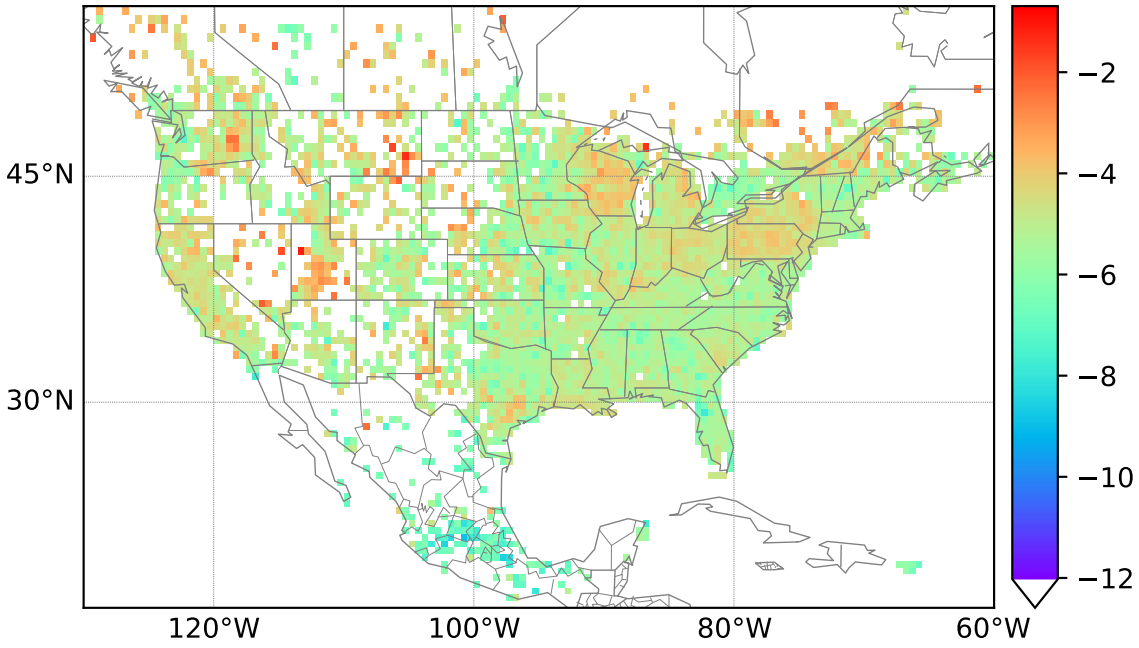

Figure 6. Heatmap of North America showing the density of streets classified as German. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

Table 1. Total numbers of streets in North America that have been mapped to Spanish, French and German streets. After a first classification, the resulting set is filtered again in order to reduce noise and remove streets that are classified by many different countries. In this way, street specifics that are used in many countries and languages are removed (i.e., popular first names). The numbers confirm the signals we see in the plots as Spanish signals are much stronger than French or German ones.

Country	Originally Classified	Filtered/Noise-Reduced	Cells w/Signals	Cells w/Clusters
Spain	1,439,928	544,213	4035	1384
France	1,721,161	107,584	3524	410
Germany	1,156,203	52,002	3198	219

The statistical distribution of street names thus broadly agrees with historical distributions. We also see that our approach has limited resolution, at least in part because street names from similar languages cannot be safely distinguished. There is no clear signals for Dutch and Irish street names. As being part of the first settlers arriving at the North American east coast, early settlements and many street names likely have undegone reconstructions and renamings, in particular in highly populated areas. As mentioned by Woodard [1] and Han et al. [22], many Scandinavian immigrants joined German settlements in the regions close to the Great Lakes and the Canadian border. Again, language similarity and possibly renaming resulted in only weak signals for Scandinavian street names in North America.

Interestingly, the density of Spanish odonyms is expectionally low in the Mexican state of Yucatán, presumably reflecting the prevalent usage of Mayan languages in this region. For more details, see the Supplementary Figure S7 depicting a self-mapping from Mexican street specifics onto street names in Mexico. The region of Yucatán can be clearly distinguished in the plot. These data confirm that the relative frequency of common Mexican street terms is also much reduced in Yucatán.

3.2. Clusters of European Street Names in North America

Figure 7 shows geographic locations of clusters of streets with names originating from the same European country for Belgium (representing French clusters, as France would result in a large amount of data points), for Austria and Germany (German), and Scandinavian including the countries Sweden, Denmark, Norway and Iceland. A street cluster contains at least three streets that are located close to each other. See also Materials and Methods (Section 2) and the supplemental data for further details.

Belgian clusters cover the dense areas of French street names as shown in Figure 5, thus Quebec, southern Louisiana and Saskatchewan. Clusters of German street names match the areas around the Great Lakes in the east of the US as shown in Figure 6. Few Scandinavian clusters have been detected, which agrees with the Scandinavian immigrants joining the German settlers [1]. The most western and central regions of the US consist of a mixture of clusters of European street names with German, Austrian and Scandinavian origins. This reflects a large number of smaller European settlements in the western and less densely populated parts of the US. The observation agrees with Woodard [1] and is explained by the fact that railway companies advertised regions in the Far West in order to attract railway workers and farmers to settle down. Many people from Europe but also from the eastern parts of the US followed the advertisement. However, harsh climate conditions forced many settlers to give up. The southeastern regions of the US seem to mainly include English street names with relatively low numbers of closely located and clearly distinguishable European street names that could form clusters. As described by Woodard [1], these regions were less welcoming for immigrants and mainly populated by people identifying with the new American lifestyle, clearly separating from European roots. Many catholic European immigrants moved further towards Mexico, as catholicism was practiced by most of its inhabitants but was not well accepted in many other parts of North America at that time. US census data show that there are around 9.6% of inhabitants in Texas with German,

6.9% with Irish, 2.0% with French and 1.9% with Italian ancestry (US census data: https://data.census.gov/cedsci/table?g=0400000US48_1600000US4865000&tid=ACSDP1Y2014.DP02&q=DP02).

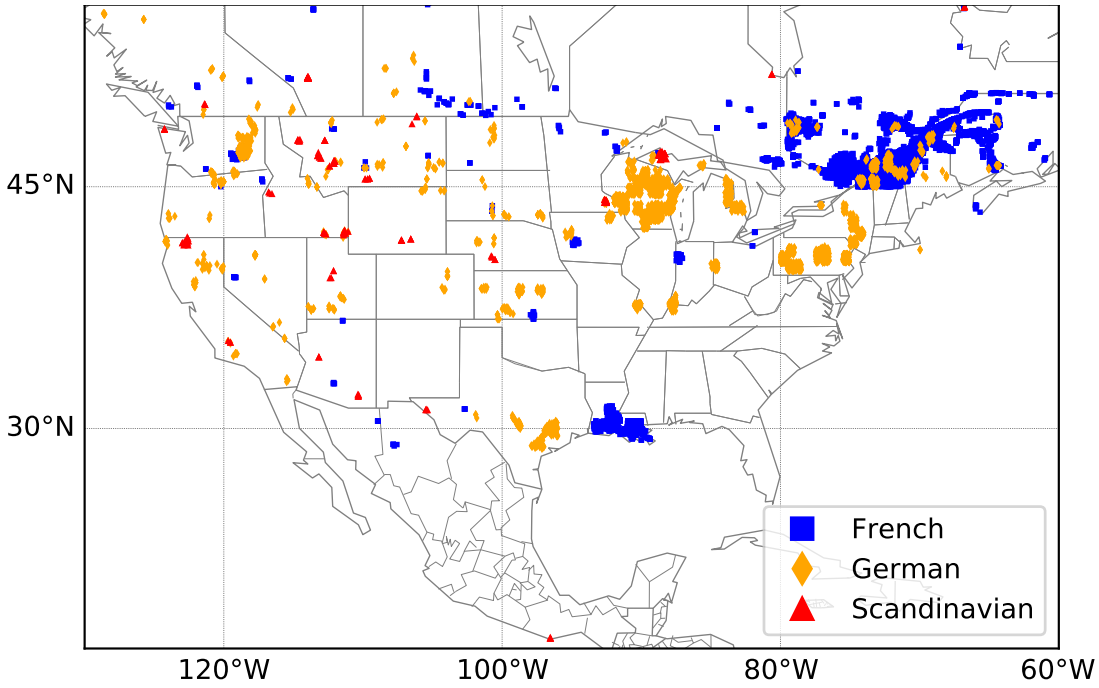

Figure 7. Clusters of European streets based on closely located streets whose names are classified as originating from the same European country.

3.3. Oldest Street Names Can Be Found in Rural Areas

The central states of the US have a much lower population density than the coastal areas as shown in Figure 3. Streets and other infrastructure only change slowly in rural areas and a lot of renaming has only been done in cities [20]. This contributes to the high number of clusters regarding European street names in the rural areas, not only in the central states of the US. Early European settlers named streets based on terms referring to places in Europe or names of settlers creating their farms in that street (and possibly the street itsself, too).

Figure 7 shows locations of street clusters of European street names for German, Austrian and Scandinavian origins. Taking a closer look at the specific locations of the clusters, it can be seen that they are mainly contained in rural areas. Figure 8 shows an example of a cluster containing street names with mainly German and Austrian origin. The cluster is located in the rural area in the state of Washington (WA), more precisely Lincoln County with the city of Davenport as county seat. The streets correspond to the German and Austrian clusters in Washington (most northwestern state of the US), close to the Canadian border as shown in Figure 7. The example cluster contains streets with specific terms named Kramer, Mohler, Teschner, Oestreich or Platte and also the area around the cluster contains further street names of German origin such as Schirr, Wollweber and Reith. Lincoln county has been reported that around 30% of inhabitants are of German origin by the census data of 2014 (US census data: https://data.census.gov/cedsci/table?q=DP02&tid=ACSDP5Y2014.DP02&g=0400000US53_0500000US53043_0600000US5304390848&vintage=2014&layer=VT_2014_050_00_PY_D1) and is listed in Table 2.

A slightly different picture can be seen in southern Louisiana which has been highly influenced by French immigrants. Here, French street terms also appear in larger cities whereas some street names are completely French, i.e., using the French street generic *rue* instead of *street* or *road*. In a few cases, the French street generic *rue* is combined with *road* or *street*, e.g., *rue des chenes road* or *rue de commerce street*.

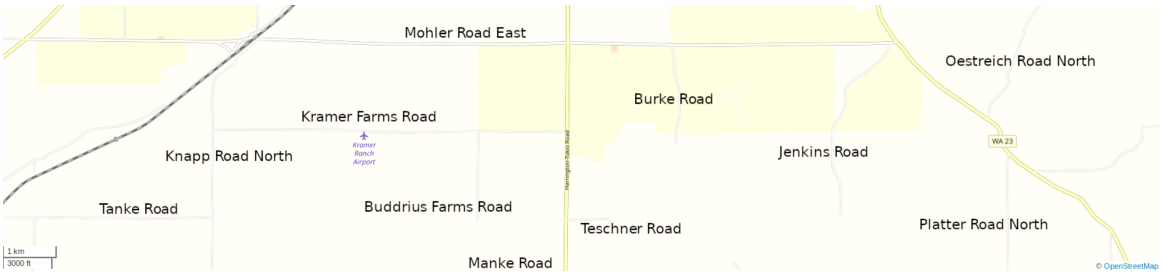

Figure 8. Cluster of streets with names mainly from German speaking origin. The cluster is located in Lincoln County in the state of Washington. The map has been created using OpenStreetMap (www.osm.org).

Table 2. Numbers of the US american census data of 2014 for the state of Washington (WA), Lincoln County and its county seat Davenport for the total population and inhabitants of German origin thereof.

	Washington State (WA)	**Lincoln County (LC), WA**	**Davenport CCD, LC, WA**
Total Population	6,899,123	10,409	4950
German Origin	1,270,685 (+/−10,392)	3559 (+/−356)	1621 (+/−334)
German Origin in %	18.4	34.2	32.7

3.4. Europe

Self-mapping of European street specifics to European street names results in clearly distinguishable regions with high densities of streets of the same origin. Mapping street names of a single European country to the complete region of Europe clearly shows countries with minorities or several official languages such as Belgium (see Figures 9 and 10) and Switzerland (see Figures 9 and 11). Also the border regions show mixed occurences of street names originating in one of the countries incident to the border, i.e., Germany and France, Austria and Italy, Germany and Poland, France and Italy (see Figures 9–11). European countries share a set of street specifics including common first or last names or names of popular historical personages. This introduces noise in the data set by showing signals for common street names for mainly unrelated countries. Therefore, we omit street names that at are classified at the same time as British, Spanish and German or Dutch for the self mappings in Figures 10 and 11. For the mapping of French street specifics to Europe (Figure 9), only streets which are at the same time classified as British are omitted. Table 3 shows numbers underlying the plots.

We observe several plausible patterns. German street names not only cover Germany, Austria, and the German-speaking parts of Switzerland, but also appear, at lower relative densities in Denmark and the Netherlands (where language similarity may serve as explanation) but also in the the parts of Poland that were part of Germany until WW2, as well as in the Czech Republic and a region in southern Poland that were part of the Habsburg empire, Figure 11. French street specifics appear in high density in all areas where a Romance language is spoken, and in Hungary, Figure 9. With the exception of Hungary, language similarity may provide an explanation. This in particular pertains to the Catalan-speaking areas in north-eastern Spain including the Balearic Islands. More puzzling is the high density in the Basque region of Spain. The signal in Hungary is consistent with immigration from France in the 18th and early 19th century. As expected, Dutch steet specifics appear in the Netherlands and the northern half of Belgium, but also, at much reduced density in northern Germany, which roughly coincides with the region in which Low Saxon dialects, which is also prevalent in parts of the Netherlands [27].

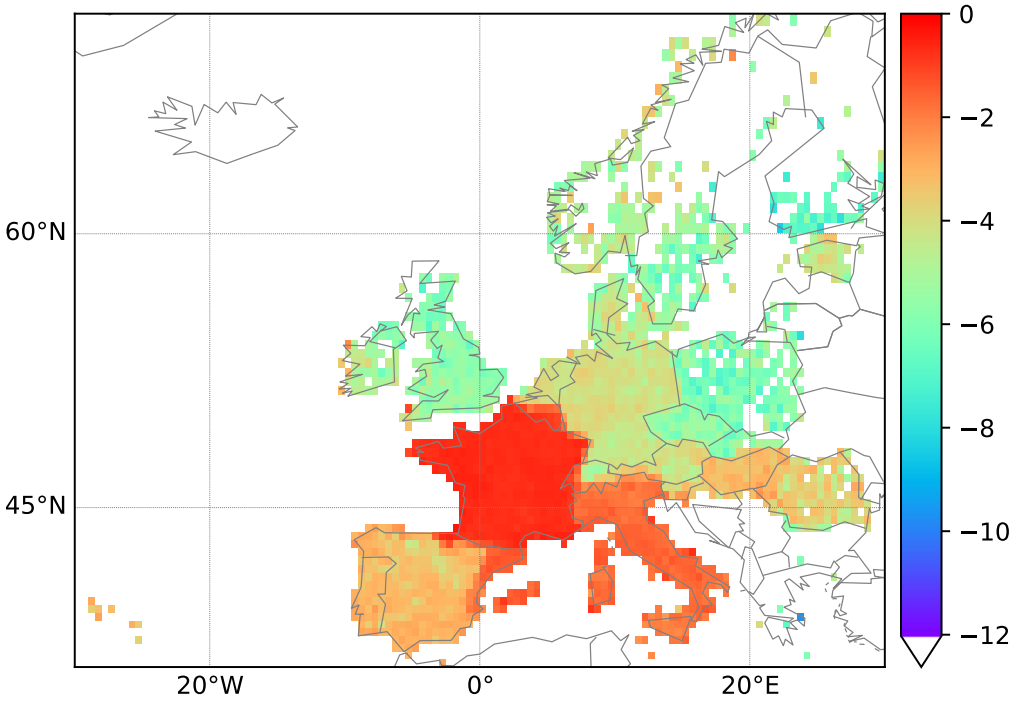

Figure 9. Heatmap of Europe showing the density of streets mapped to French street specifics. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

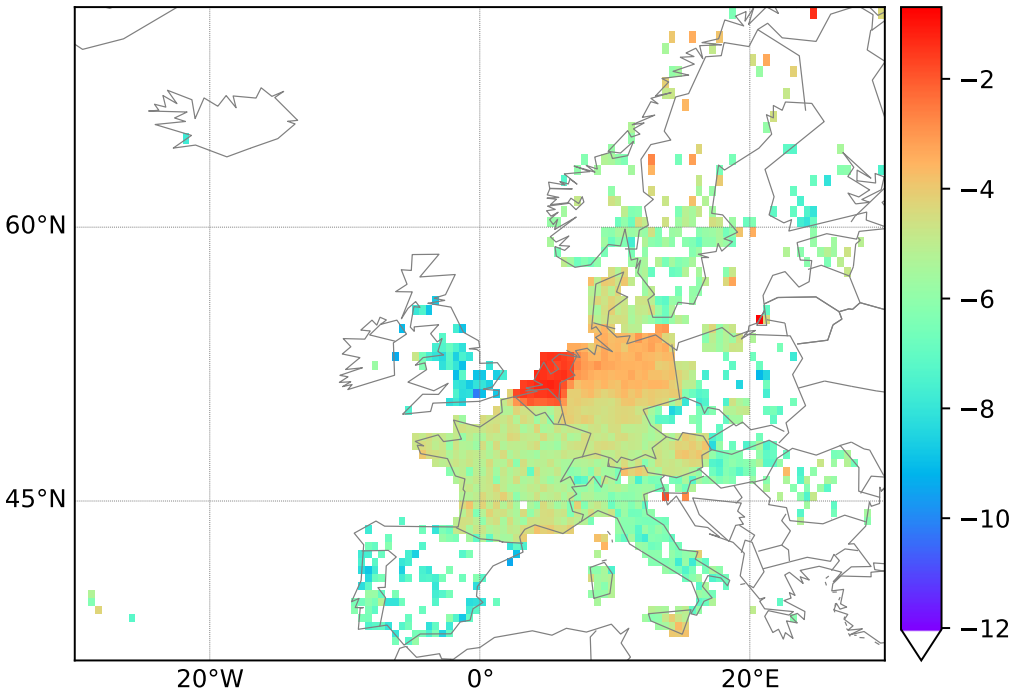

Figure 10. Heatmap of Europe showing the density of streets mapped to Dutch street specifics. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

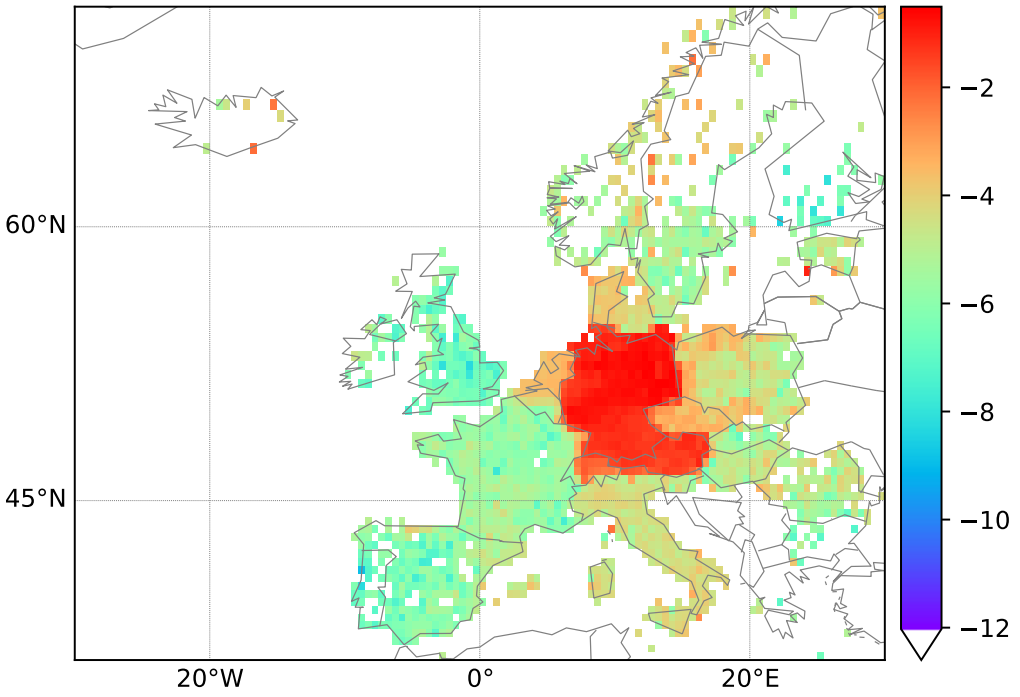

Figure 11. Heatmap of Europe showing the density of streets mapped to German street specifics. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

Table 3. Number of streets in Europe that have been mapped to Dutch, French and German streets. After a first classification, the resulting set is filtered again in order to reduce noise and remove streets that are classified by many different countries. In this way, street specifics that are used in many countries and languages are removed (i.e., popular first names).

Country	Originally Classified	Filtered/Noise-Reduced	Cells w/Signals	Cells w/Clusters
Netherlands	232,211	112,028	1432	151
France	1,518,943	963,479	2025	1083
Germany	1,008,306	522,256	1925	467

3.5. African Street Names and Mapping to European Street Specifics

Street name data for 53 African countries has been downloaded in September 2020. In total, much fewer data are available in OSM than in the case of Europe or North America which can also be seen when plotting street densities for all Africa. Please see the Supplementary Materials for more information. The results for mapping Dutch and French street specifics to African countries are depicted in Figures 12 and 13. We observe an increased density of French street specifics in particular in Algeria and Marocco, the former French colonies in western and central Africa and Madagaskar, and in the Democratic Republic of Congo (presumably relecting its history as a Belgian colony). Conspiciously, some countries that would have been expected based on their colonial history are missing. In Cameroon, for instance, almost all street specifics are numeric and thus omitted from our analysis. Dutch names, Figure 12, are largely confined to South Africa and Namibia, again in line with expectation. Signals in the data for Africa are less clear and much more noisy than in North America or Europe. This is caused in part by the small amount of available data and the large fraction of numeric street names. Another important factor is that the language classifier `fasttext` used in this study only includes a small sample of African languages.

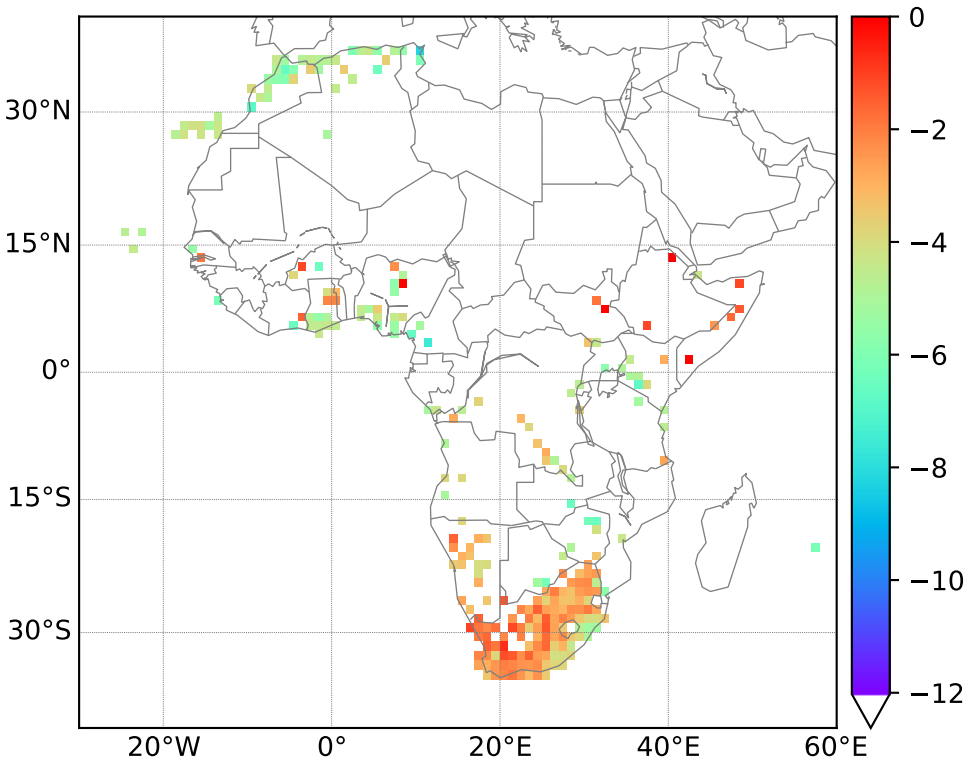

Figure 12. Heatmap of Africa showing the density of streets mapped to Dutch street specifics. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

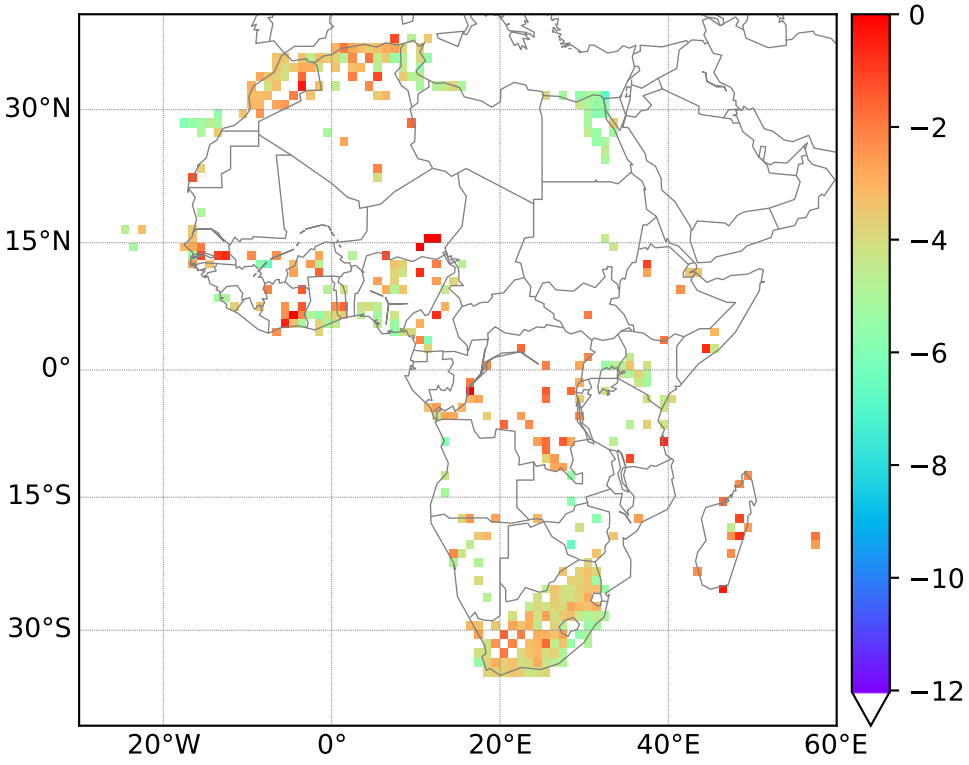

Figure 13. Heatmap of Africa showing the density of streets mapped to French street specifics. Colors indicate the density of streets in this area, thus the logarithmic value of the number of streets normalized by all streets in the cell of the grid.

4. Discussion

We have shown in this contribution that it is possible to detect former European settlements in North America based solely on the comparison of street names. Here, mainly the street specifics are encountered for the comparison. The results confirm Spanish and French speaking areas in North America such as Mexico and Quebec. However, Spanish street names can also be found in the southwestern states of the US and Florida which are known to have been settled by Spanish-speaking immigrants. A similar case exists for regions with a large portion of French street names. In addition to Quebec, Saskatchewan in Canada and Louisiana in the US show high amounts of French street names which confirms the immigration of settlers of French origin. As described by Woodard [1], German-speaking immigrants as well as Scandinavians moved into the areas south of the Great Lakes which is confirmed by the high density of German street names in these regions.

The analysis of dense regions also reveils smaller clusters of streets of European origin. The central and western states of the US and Canada show a high amount of small European street clusters which indicate European settlements. Most of these clusters are located within rural areas as they are less prone to undergo reconstructions and renaming of streets. We again see German, Austrian and Scandinavian clusters not only in the vicinity of the Great Lakes but also in the central and western US states.

As a first step, our aim was primarily to show that the available data indeed contain a detectable signal. Our approach therefore was purely data driven, i.e., we did not use any knowledge based curation of street names and instead relied on statistical (co)occurences along. The simple statistical approach taken here can certainly be improved in future work, in particular regarding the assignment of the country or region of origin of a street name. While Spanish, French or German origins are different enough to yield clearly distinguishable signals, the resolution is insufficient to clearly distinguish e.g., Spanish and Portugese, or Austrian and German influence since each pair of countries shares too many street specifics. This observation suggests, in particular, in a next step to develop statistics to detect street specifics that are characteristic for geographically limited regions of origin.

Several methodological issues arise in this kind of quantitative analysis. Most importantly, extensive preprocessing is required to prepare good origin-specific collections of street name specifics. Starting from European data, street names in countries with a single dominating language provide a good starting point but are not specific enough. In particular language similarities as well as very frequently used geographic terms or surnames tend to generate spurious matches. Using the street specifics that appear in many different regions as a negative filter seems to work well, at least on the Europe-centric datasets used here. The development of training methods that are also applicable when the initial sets cannot not defined in an obvious manner by the border of countries or lower-level adminstrative entities within them remains as an interesting topic for future research.

In our initial analysis we have focussed on the most obvious signals reflecting settlement and migration history related to the ethnicity of migrant and minority populations. In particular in Africa, the patterns reflect colonial history. However, street name data can also reflect subsequent political changes such tendencies to eradicate colonial names after independence. Active renaming is in particular a likely explanation in former colonies that show no signal for street specifics matching the former colonial power. Such effects are difficult to detect with our current methodology. Since they involve a change in street naming conventions, temporally stratified data could be used. The most convenient data would be street name data for different points in time, which would allow a direct measurement of renaming rates and language/concept transitions. Since OSM exists only since 2004 and data were comparably sparse initially, a direct analysis of renaming is at least at present limited to very recent history. As a proxy, one could also associate streets in rapidly growing communities with the time at which they appeared on maps, even if past names are difficult to access.

5. Conclusions

In summary, we have shown that a statistical analysis of odonyms allows conclusions about regional settlement histories, despite renamings of streets, reconstructions and expansions of settlements, and a change in predominant language. The results agree with the historical context of European settlement as described in Woodard [1]. Also the results of previous studies in North America [5,6,22] are relected at least in part in our results.

We demonstrated that freely accessible and easily evaluated sources such as street name data can be used as additional information to review and confirm historical events such as settlements and migration profiles. However, this is only possible for countries or regions that homogenously have used one official language for a significant amount of time. Countries with a large number of different languages make comparison and mapping more complicated and languages cannot be distinguished readily based on geographic data alone. Despite the results obtained here, there are limitations to odonyms as the sole data source. Most naturally, they could be combined with additional socio-economic data such as census data.

For th technical point of view, the approach presented here is not limited to OSM data. However, as we use `osmium` tool to parse OSM files, the software accompanying this contribution will need to be adapted to use specific parsers for other data sources such as the Google Earth project. The use of proprietory data sources in addition may encounter legal constraints.

Supplementary Materials: The following are available at http://www.mdpi.com/2413-8851/4/4/74/s1 .

Author Contributions: Conceptualization, P.F.S.; Data curation, S.J.B.; Methodology, S.J.B. and P.F.S.; Writing—original draft, S.J.B. and P.F.S.; Writing—review editing, S.J.B. and P.F.S. All authors have read and agreed to the published version of the manuscript.

Funding: DAAD FITweltweit travel grant to SJB. Publication cost are covered by the Open Access Fund of Leipzig University. This work was supported by the German Federal Ministry of Education and Research (BMBF, 01/S18026A-F) by funding the competence center for Big Data and AI "ScaDS.AI Dresden/Leipzig".

Acknowledgments: S.J.B. thanks Simon DeDeo for helpful discussions and useful hints.

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

References

1. Woodard, C. *American Nations: A History of the Eleven Rival Regional Cultures of North America*; Penguin: London, UK, 2011.
2. Weber, J. The Three American Wests. *Prof. Geogr.* **2019**, *71*, 239–252. [CrossRef]
3. Stuckey, M.E.; Murphy, J.M. By any other name: Rhetorical colonialism in North America. *Am. Indian Cult. Res. J.* **2001**, *25*, 73–98. [CrossRef]
4. Alderman, D.H. Place, naming and the interpretation of cultural landscapes. In *The Ashgate Research Companion to Heritage and Identity*; Brian, G., Peter, H., Eds.; Routledge: London, UK, 2016; pp. 195–213.
5. Berger, T.; Engzell, P. American geography of opportunity reveals European origins. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6045–6050. [CrossRef] [PubMed]
6. Fuchs, S. An integrated approach to Germanic place names in the American Midwest. *Prof. Geogr.* **2015**, *67*, 330–341. [CrossRef]
7. Nash, C. Irish placenames: Post-colonial locations. *Trans. Inst. Br. Geogr.* **1999**, *24*, 457–480. [CrossRef]
8. Pred, A. *Lost Words and Lost Worlds: Modernity and the Language of Everyday Life in Late Nineteenth-Century Stockholm*; Cambridge University Press: Cambridge, UK, 1990.
9. Pred, A. Languages of everyday practice and resistance: Stockholm at the end of the nineteenth century. In *Reworking Modernity: Capitalisms and Symbolic Discontent*; Rutgers University Press: New Brunswick, NJ, USA, 1992; pp. 118–154.
10. Carter, P.; McKenzie, L. *The Road to Botany Bay: An Essay in Spatial History*; Faber & Faber London: London, UK, 1987.
11. Cohen, S.B.; Kliot, N. Place-names in Israel's ideological struggle over the administered territories. *Ann. Assoc. Am. Geogr.* **1992**, *82*, 653–680. [CrossRef]

12. Yeoh, B.S. Street-naming and nation-building: Toponymic inscriptions of nationhood in Singapore. *Area* **1996**, *28*, 298–307.
13. Azaryahu, M. The power of commemorative street names. *Environ. Plan. D Soc. Space* **1996**, *14*, 311–330. [CrossRef]
14. Badariotti, D. Les noms de rue en géographie. Plaidoyer pour une recherche sur les odonymes/Street names, an argument for a geographic research. In *Annales de Géographie. JSTOR*; Armand Colin: Paris, France, 2002; pp. 285–302.
15. Algeo, J. From classic to classy: Changing fashions in street names. *Names* **1978**, *26*, 80–95. [CrossRef]
16. Palonen, E. The city-text in post-communist Budapest: Street names, memorials, and the politics of commemoration. *GeoJournal* **2008**, *73*, 219–230. [CrossRef]
17. Gill, G. Changing symbols: The renovation of Moscow place names. *Russ. Rev.* **2005**, *64*, 480–503. [CrossRef]
18. Azaryahu, M. German reunification and the politics of street names: The case of East Berlin. *Political Geogr.* **1997**, *16*, 479–493. [CrossRef]
19. Faraco, J.C.G.; Murphy, M.D. Street names and political regimes in an Andalusian town. *Ethnology* **1997**, *36*, 123–148. [CrossRef]
20. Mencken, H.L. American street names. *Am. Speech* **1948**, *23*, 81–88. [CrossRef]
21. Thiel, S.; Pippig, K.; Burghardt, D. Analysis of street names regarding the designation of cities. In Proceedings of the 26th International Cartographic Conference. International Cartographic Association, Dresden, Germany, 25–30 August 2013.
22. Han, E.; Carbonetto, P.; Curtis, R.E.; Wang, Y.; Granka, J.M.; Byrnes, J.; Noto, K.; Kermany, A.R.; Myres, N.M.; Barber, M.J.; et al. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat. Commun.* **2017**, *8*, 14238. [CrossRef] [PubMed]
23. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
24. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv* **2016**, arXiv:1607.01759.
25. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. FastText.zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
26. Dupuis, S. Francophones of Saskatchewan (Fransaskois). In *The Canadian Encyclopedia*; Historica Canada: Toronto, ON, Canada, 2019.
27. Gooskens, C.S.; Kürschner, S. Low Saxon Dialects across borders. In *Zeitschrift für Dialektologie und Linguistik—Beihefte*; Franz Steiner Verlag: Stuttgart, Germany, 2009; Volume 138.