Surveying GeoNames Gazetteer Data for the Nordic Countries

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway, dirk.ahlers@idi.ntnu.no

Abstract

This paper takes a look at freely available gazetteer data for the Nordic countries. We examine locations in this region to understand their characteristics and the quality of the available data. Several indicators are developed and discussed to estimate the expected data quality. The distribution and coverage of the data is mapped and the accuracy and quality indicators are visualized. The used method focuses on populated places as locations of interest but can be extended to arbitrary types of locations. The results give insights into the distribution of issues based on multiple indicators and give an estimate of per-country data quality.

1 Introduction

Location information plays a vital role in many applications. For example, place names can be geocoded to provide coordinates for a named place or vice versa, to provide place names to coordinates; information systems need a basic knowledge about the world's countries and cities; sizes and population numbers are important for many statistics. One source of such information are gazetteers – geographical thesauri – that provide detailed information about places, including names, coordinates, types, and more [9]. They can contain data about physical and geographical features, most importantly cities, villages, and other populated places, structures such as airports or harbours, but also natural features such as lakes, mountains, forests, etc. As with other data sources, quality issues play a role in the use of spatial data [6, 15], such as coverage, consistency, and accuracy [1] and many location-based services, data mining or statistics approaches rely strongly on accurate gazetteers [8, 10, 14].

GeoNames.org is the most widely used freely available gazetteer. It has a worldwide coverage and also good coverage in the Nordic countries. For example, it contains about 120000 places for Norway alone, of which almost 10000 are populated places. There is little previous work on the analysis of gazetteer data quality. There has been previous work on OpenStreetMap data quality [4], relying strongly on the user contribution to the dataset. To the best of our knowledge, our approach is the only one working on quality indicators for the GeoNames dataset. We started on this

This paper was presented at the NIK-2015 conference; see http://www.nik.no/.

before, but with a rather limited list of countries and indicators [2] based on a local search project [3]. It mostly concerned the investigation of truncated coordinates as well as manual assessments of overlapping, overspill, and feature classes. We extend this to more fully detect and classify potential errors automatically. For this, we are working on more global and robust indicators. In this paper, we take an in-depth look at a selection of the data, namely for the example of the Nordic countries of Norway, Denmark, Finland, Iceland, and Sweden, as well as associated territories. We take this as an opportunity to understand the indicators on a limited variety of countries and examine them in-depth (and already test them on a world scale) and additionally estimate the influence and characteristics of associated or autonomous territories and how they compare to the respective state they are related to. The Nordic countries are an interesting subject of this type of analysis due to the high number of dependencies that need to be considered for a full picture. This format gives us the opportunity to describe the relationship and idiosyncrasies of the countries and the data in-depth and in detail to consider bias and quality of the data on a country by country level.

2 Data source and country characteristics

For this study, we focus on the Nordic countries of Norway (NO), Denmark (DK), Finland (FI), Iceland (IS), and Sweden (SE). Many of these are not limited to their mainland, but have additional dependent territories in other regions of the world. This allows us to explore how well these are handled in the gazetteer data. The addition includes the territories of Faroe Islands (FO) and Greenland (GL) for Denmark, Svalbard and Jan Mayen (both use SJ) for Norway, and Åland Islands (AX) for Finland (cf. Fig. 2 for a map). The Danish ones are autonomous countries, AX is an autonomous region of Finland, Svalbard is an unincorporated area under sovereignty of Norway, but subject to a special status, and Jan Mayen is a dependency of Norway, integrated into the county-level administration.

Gazetteer comparison

GeoNames is a freely available gazetteer and currently contains data for 250 territories[1], identified by their ISO country code[2]. Its data comes from a wide variety of mostly official public data sources[3] of varying density, coverage, and quality that is merged internally to provide its gazetteer data on a world-scale.

Other gazetteers follow a very similar structure and collect basically the same data, so the methodology and the indicators developed here can be adapted to analyse them as well. Such alternative commercial, thematic, or national gazetteers may be preferred in certain cases because they provide better coverage, more current data, or better capture places in the local language, are a defined authoritative source, or provide better integration and interchange. For example, many museums use the Getty Vocabularies as an authoritative source so that, e.g., pieces of art, architecture, artists, cultural or archaeological artefacts are classified consistently. Their well-known gazetteer is the Getty Thesaurus of Geographic Names (TGN)[4].

[1] http://download.geonames.org/export/dump/

[2] ISO Online Browsing Platform: Country Codes https://www.iso.org/obp/ui/#search

[3] http://www.geonames.org/data-sources.html

[4] Getty Thesaurus of Geographic Names (Getty Research Institute) http://www.getty.edu/research/tools/vocabularies/tgn/index.html

Table 1: Quantitative Comparison of TGN and GeoNames (GN). Grouped by dependencies, sorted alphabetically. TGN adapted shows the adapted count to the GeoNames hierarchy, GN.all and GN.ppl show all and only populated places, %GN.ppl is their percentage, the last two columns show the percentage of TGN place counts compared to GN.

	TGN	TGN adapted	GN.all	GN.ppl	%GN.ppl	TGN per GN.all	TGN per GN.ppl
Denmark	878	878	14520	7299	50,3%	6,0%	12,0%
Faeroe Islands	19	19	1746	224	12,8%	1,1%	8,5%
Greenland	184	184	7391	278	3,8%	2,5%	66,2%
Finland	554	553	42379	14013	33,1%	1,3%	3,9%
Åland Islands		1	3234	385	11,9%	0,0%	0,3%
Iceland	136	136	15781	414	2,6%	0,9%	32,9%
Norway	679	679	119039	9886	8,3%	0,6%	6,9%
Svalbard	9	11	5525	17	0,3%	0,2%	64,7%
Jan Mayen	2						
Sweden	1781	1781	94212	32572	34,6%	1,9%	5,5%

We provide a brief quantitative comparison of GeoNames to TGN to highlight some issues and show the size of the data in Fig. 1. We have adapted the TGN structure to GeoNames: In TGN, Åland is not a separate entity, but is maintained under Finland as a 'former administrative division' with additional historical types of 'autonomous province' and 'first level subdivision'. Yet, its country code AX is listed only as an alternative name. Then again, TGN separates Svalbard from Jan Mayen, whereas in GeoNames, they are joined as they share the same country code, but they could be separated by administrative divisions. AX seems highly underrepresented in TGN, but overrepresented in GeoNames, but at manual inspection, the data seems to be valid. We see this in other areas as well, where certain countries or regions have a much higher place density than would be expected. We speculate that this is due to the integration of data sources that also contain very small villages or hamlets that may miss from more general data sets.

It is obvious that GeoNames has a much broader coverage with much more places available than TGN, which mostly contains major places. Only for Greenland and Svalbard does it reach over 60% of GeoNames' places, for the rest it covers on average only 10% (and 22% for the whole list). However, comparing both qualitatively, TGN has a comprehensive system of place types and also allows multiple place types (feature type in GeoNames) and alternative geographical hierarchies, but we mostly see inhabited places and administrative areas. TGN also has rich provenance data, listing contributing sources, which is not available in GeoNames.

We continue to use GeoNames in the remainder of the paper for three reasons. First, it is freely available under a CreativeCommons license, while TGN only allows free queries, but no full data access. Second, it is substantially larger and provides better coverage, and it is the most widely used source. It would be future work to examine in more detail and possibly merge features [7, 11] to improve the quality.

Country characteristics

The Nordic countries have a decreasing population density towards the North and on average lower than the rest of Europe, which partly shows in large amounts of lakes, mountains, forests, icecaps, or glaciers as uninhabited areas [13], which will be seen in more detail later in the mapping.

There is no direct way to get dependency relations from the GeoNames dataset, these have to be uncovered by other means[5]. For example, for every country in GeoNames, there is an entry that contains the feature type of the country as a political entity, and also for further subdivisions[6]. In most cases, there is an entry found which is noted as PCLI (independent political entity). If this does not exist, there can be entries with other codes. We find PCLD (dependent political entity) for FO, GL, and AX and TERR (territory) for SJ. Additionally, for SJ, we find an ADM1 code for Jan Mayen and three ADM2 codes for administrative subdivisions of Svalbard. Note, however, that this does not reliably show the relations, as Jan Mayen is an ADM1 of Norway, not of Svalbard. Also, the language field does not provide conclusive evidence. For example, Faroe Islands are listed as "fo,da-FO", Svalbard as "no,ru", and AX as a Swedish-speaking island is listed as "sv-ax". Furthermore, this would fail for English-speaking countries. TGN maintains multiple hierarchies, for example for Greenland, one geographical as 'World > North and Central America > Greenland', and an alternative political one as 'World > Europe > Denmark > Greenland' but as shown before does not separate Åland. This illustrates the problems with the semantic extraction of such relations and also how their treatment can differ in separate sources. As there is a limited number of dependencies, for this study, the relation between the territories was drawn manually.

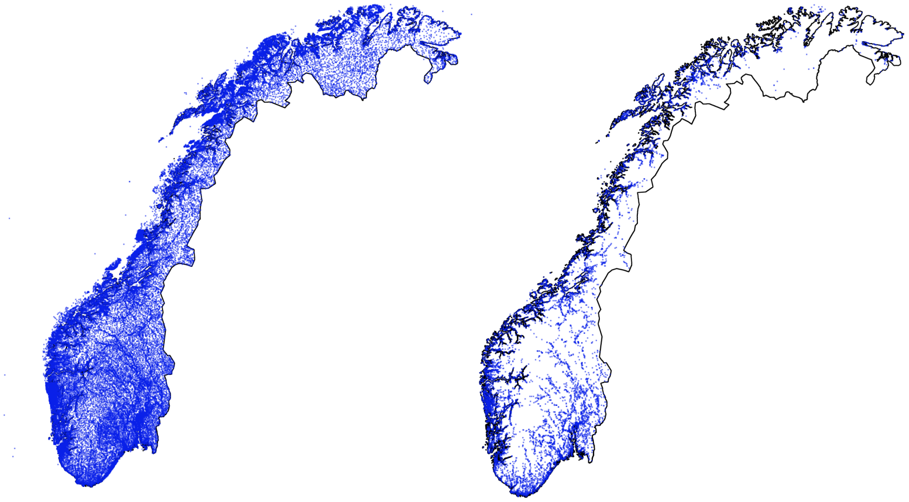
Of the countries under study here, Norway additionally has three uninhabited dependent territories (biland). These are Bouvet Island, Queen Maud Land, and Peter I Island. The first is a sub-antarctic island (BV), the other two are part of Antarctica (AQ). Since for Antarctica, no subdivisions are defined in the ISO codes and thus in the GeoNames export, we cannot directly retrieve places in those territories, but due to the pie-shaped distributions of territories, these could be easily filtered by longitude. As they are uninhabited, no populated places should be available. However, we see research stations varyingly mapped as PPL (populated place) or STNB (scientific research base). In the Norwegian part, we see the Troll research station and airfield and the Tor research station. There is a possible conflict in the use of PPL as places in Antarctica are technically not populated by permanent residents, which is also why the dataset shows zero population. On the other hand, Jan Mayen also only houses staff that operates the stations there, but is listed with a population of 18. Furthermore, the settlement Olonkinbyen is listed as PPLA (seat of a first-order administrative division, similar to a capital) despite it being administered from the mainland.

These remarks have already shown that the use of the data is not without challenges. Most complications are related to different definitions or notations of countries and the ISO codes. While for most generally accepted states it works as expected, for certain tricky regions, there can be a mismatch between different data sources and even conflicts in definition. For example, the (technical) ISO codes are not fully consistent with the country definition of the UN[7]. The ISO codes correspond mostly to top-level domains, but with some exceptions. For example, Svalbard uses the Norwegian .no ccTLD. In other cases, relations between

[5] They are available in the premium subscription: http://www.geonames.org/products/premium-data.html

[6] http://www.geonames.org/export/codes.html

[7] http://unstats.un.org/unsd/methods/m49/m49alpha.htm



(a) All Norwegian places

(b) Norwegian populated places

Figure 1: Norwegian entries in GeoNames

countries and dependencies are not stated. For example, SJ actually consists of two separately administered territories, which share the same ISO code. Vice versa, many sources list the territories of Norway and Svalbard as one joined entity with the name of Norway and subsequently cannot map Svalbard. These irregularities and idiosynchrasies have to be considered in using this or any other spatial dataset. It helps to understand countries in the GeoNames definition as territories or regions based on geographical cohesion or hierarchical relations. This is an issue that has to be considered when trying to get a full view of a state as additional countries might need to be included. For the sake of legibility, we will continue to use the term country as used by GeoNames for the different types of territories.

3 Methodology

Our goal is to develop a measure of quality for gazetteer data. At this stage, we focus on intrinsic indicators or such that can be intrinsically calculated as quality proxies. The reason is that we have no other fully reliable source on a world or regional level to compare the dataset to because GeoNames itself is already considered a ground truth. Examining data available from individual countries is discussed as future work. Even if we did a comparison, as noted above, TGN is quite small in comparison and would not match for most places. Additionally, in case there is a mismatch of coordinates or other features, there would be a tie when only using two sources. This concerns both false negatives and positives. If both gazetteers are based on the same erroneous third source, they would agree. If there would be a mismatch, it would be unclear which one is correct. The same would be true for other gazetteers. So a prerequisite to gazetteer merging [7] would also be a quality analysis of individual sources. This issue is part of our future work.

The approach of intrinsic indicators [4] instead can work on just a single data source. With this approach, we may not pinpoint individual places that exhibit an issue, but that we can show patterns and estimate anomalies on a country scale. The indicators then are only proxy for quality, but are helpful in determining fitness for use of the data. We select all places available in GeoNames for a country. Features we use are name, coordinate (latitude & longitude), feature class (the rough type of place, e.g. city or building), feature code (the detailed type of place, e.g., farm village or amphitheater), alternative names including links, and country. Additionally, aggregate data and metadata per country such as population and area is available. From this we develop intrinsic indicators based on either directly available features and metadata or on derived or computed features.

The application scenarios for fitness for use are based on the following: locating place names and disambiguating placenames for geoparsing and geocoding; linking to external sources; inverse geocoding to find the place for a coordinate or nearness calculations, such as nearest town, for location-based services; population numbers as they are used for estimates of place sizes or importance ranking; availability of POIs (points of interest); and the use of the data as basis for a knowledge base.

4 Analysis

As an initial overview, we start with the example of Norway to showcase the data. We plot all Norwegian places available in GeoNames in Fig. 1a. In a second step, we reduce this to only populated places shown in Fig. 1b. For this, we remove purely geographical features and only keep all places under the PPL (populated place, anything from city or capital to villages, towns, hamlets, neighborhoods) and the ADM (administrative subdivisions) feature types. This is more in line with major uses of gazetteers of geoparsing, geocoding, or similar tasks.

Norway is very thoroughly covered with place names for numerous geographical features, while populated places are more rare. There is a discernible pattern for the populated places which are more numerous along the coast and along rivers and their number declines towards the North. Their clustering follows well the country's population density. The pattern is less visible but still existing for all places. The populated places in Fig. 1b make up around 8% of all places in Fig. 1a.

We proceed to map the data for all Nordic countries, using an azimuthal equal-area projection to maintain the relative sizes, which would otherwise be heavily distorted. The Nordic countries are shown in Fig. 2 with all places, and only the populated places in Fig. 3.

Our ongoing work is to understand GeoNames quality and understand the available data [2]. Quality indicators mainly concern geospatial cohesion and accuracy, but also topical information. The main indicator developed is based on the raw coordinates, which in some cases are truncated to the minute, with removed seconds. This can mask a positional error up to 1.8km. The truncation is directly visible in the coordinates and thus is a direct accuracy indicator. Especially for smaller settlements, this can mean that the coordinate lies outside of the actual settlement. One possible source of this error can lie in manual digitisation of low-resolution paper maps or errors or omissions in processing steps. Truncated coordinates make up between 6 and 76% of places for our selected countries, details are found in Table 2. We map the results also in Fig. 4 for the Nordic countries. In the figures, exact coordinates are given in blue, truncated coordinates in red. The visualisation shows very well the huge differences in place density that are also seen in Table 2. Since the countries with higher density are difficult to make out, we zoom into Norway, Sweden, and Finland with Åland in Fig. 5. We leave out detailed figures for Greenland and Svalbard because the point density is too low to show much more than already visible in Fig. 4.

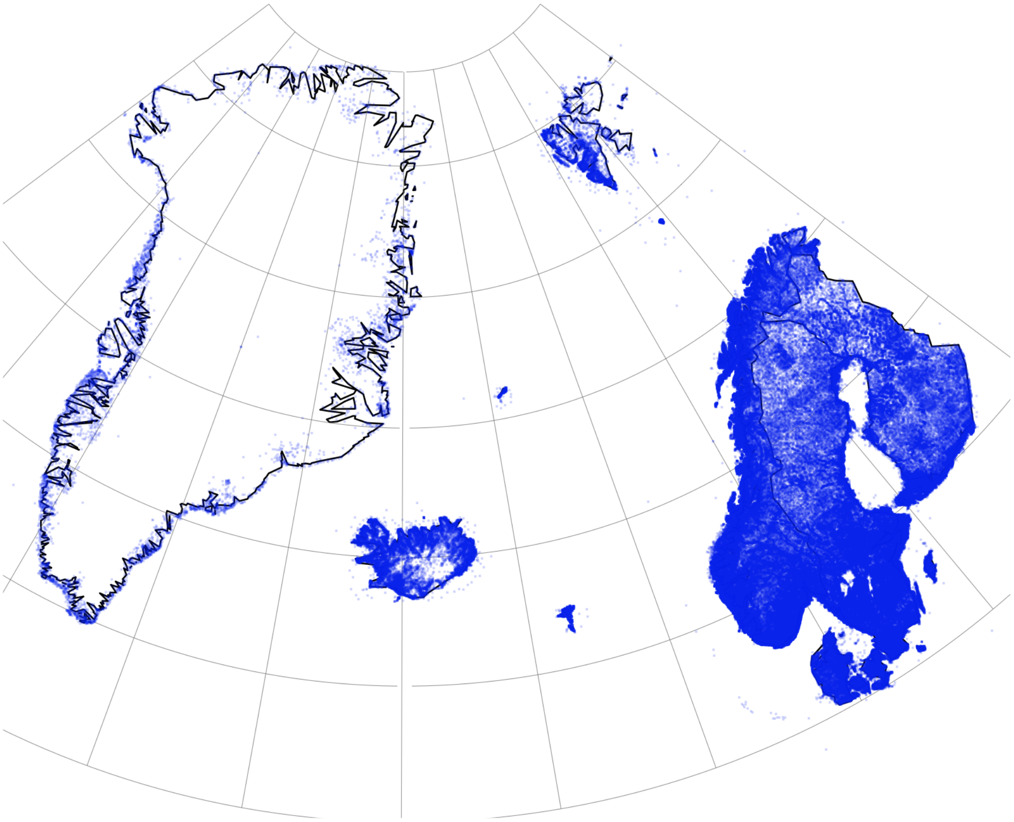


Figure 2: All places in the Nordic countries

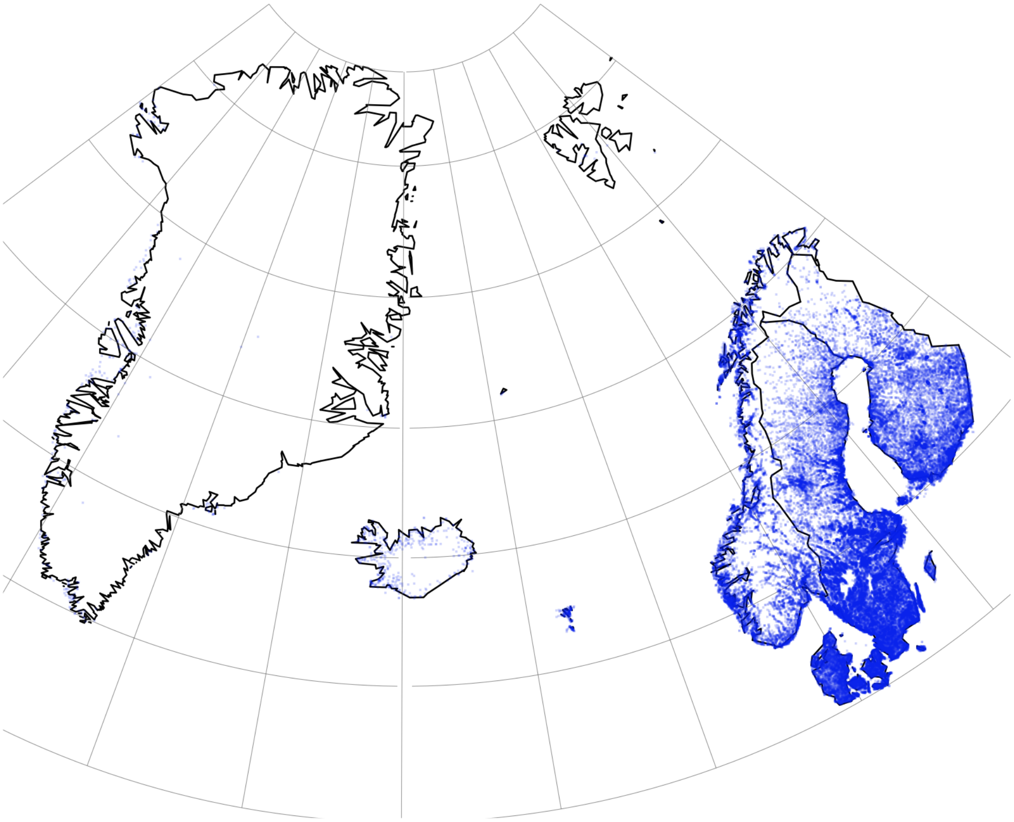


Figure 3: Populated places in the Nordic countries

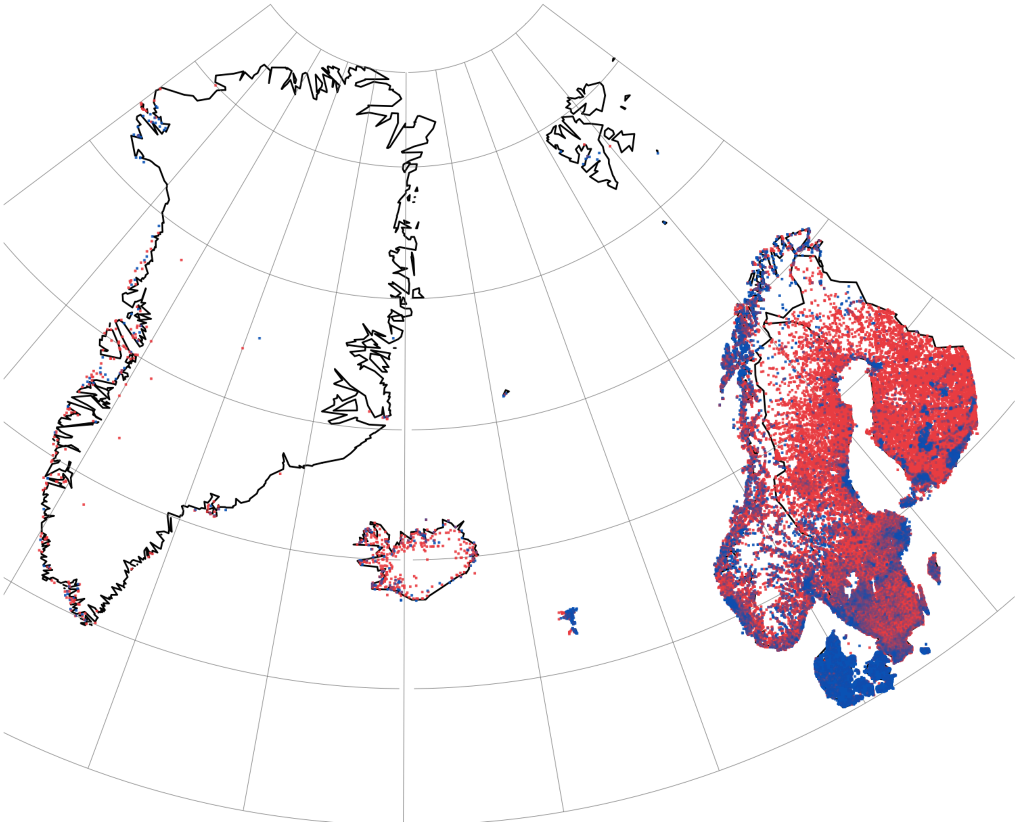


Figure 4: Populated places in the Nordic countries partitioned by accuracy, exact coordinates in blue, truncated in red

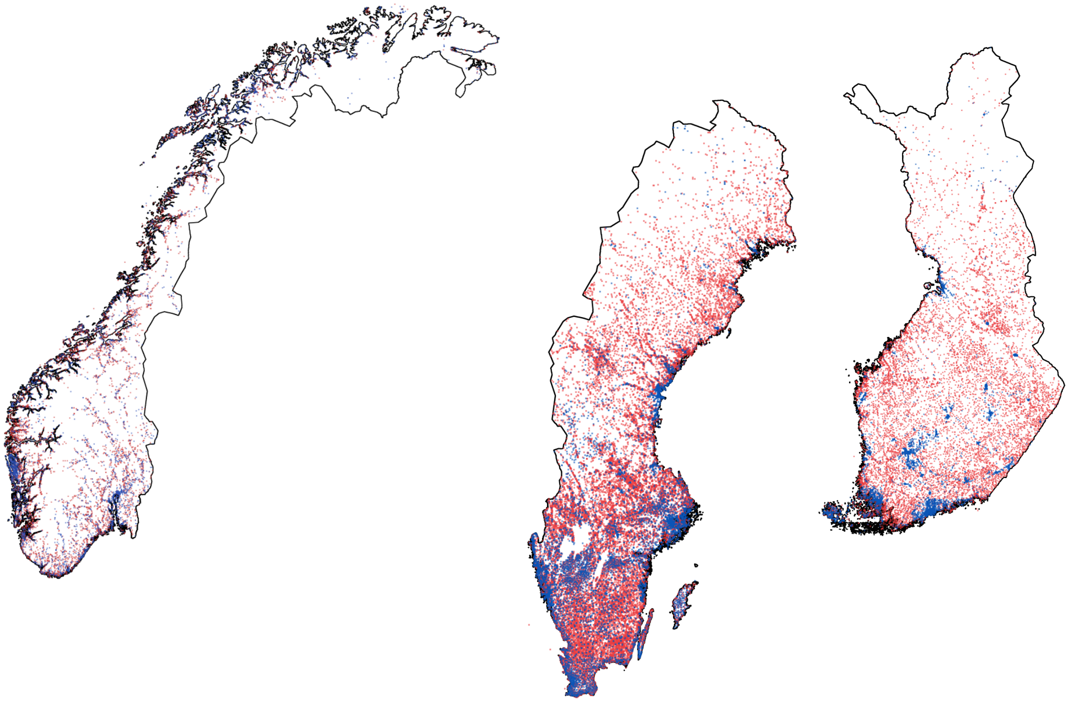


Figure 5: Places partitioned by accuracy, Norway, Sweden, Åland and Finland

Table 2: Selected indicators. %ppl states the percentage of populated places in all entries. Other values are calculated based on ppl. popno states the amount of populated places for which population numbers are available, links are links to the respective Wikipedia page for a place, %pois is the amount of POIs compared to the amount of populated places, and %poi.trunc shows the percentage of truncated POIs. Countries are sorted alphabetically and grouped by dependencies.

										%poi.
Country	all	%ppl	%trunc.	%overl.	%popno	pl/km^2	pl/1000	%links	%pois	trunc
Denmark	14520	50	6	0	6	0.169	1.331	8	44	19
Faroe Islands	1746	13	20	0	50	0.160	4.645	54	10	22
Greenland	7391	4	69	0	39	0	4.931	33	77	64
Finland	42379	33	74	4	6	0.042	2.672	6	73	95
Åland Islands	3234	12	32	0	9	–	14.414	7	22	18
Iceland	15781	3	74	1	24	0.004	1.34	25	1324	14
Norway	119039	8	56	3	11	0.030	1.974	18	312	12
Svalbard & JM	5525	0	24	0	53	0	6.667	76	376	50
Sweden	94212	35	76	9	12	0.072	3.409	8	49	68

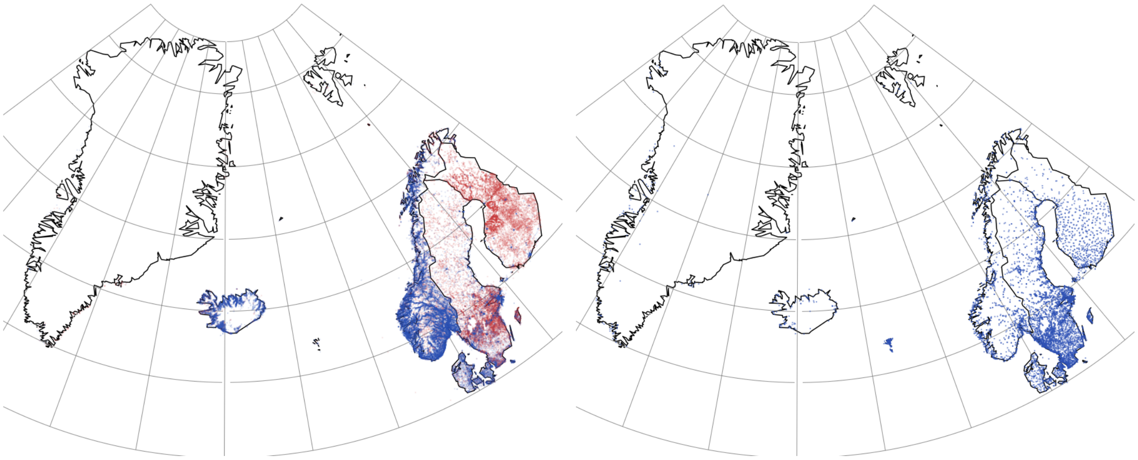
Finland, Iceland, and Sweden have very high numbers of truncation reaching levels of over 70%. Norway is still at 56%, but Denmark has very good quality at only 6%. For the dependencies, the data is often better than the mainland, except in Denmark due to its high quality. Greenland is close to 70% as well.

Repeated places, i.e., places that are identical in the features and close to each other their coordinates, can show errors in the internal merging and can provide hurdles for place disambiguation. They were only found in less than 0.5% of cases. Overlap is defined as places with the same feature code at the exact same coordinates, but with a differing name, which again complicates place disambiguation for inverse geocoding or nearness calculations. This was found in 0–9% of places. We did not find any exact duplicates.

We have developed additional indicators as seen in Table 2. One is the number of populated places that have population numbers associated with them. This indicates the level of detail in the metadata and is used in size estimations for geoparsing or coverage calculations. The linkage of GeoNames places to Wikipedia is determined by the amount of places that have a link to a Wikipedia article in any language. This is useful from a Linked Data perspective and also to gain additional knowledge about places, including descriptions and metadata.

Of the many potential errors described in our previous work, many were much less prominent in this study. A remaining issue are places outside the country's border. For maritime and undersea features, this is expected, but populated places need to lie inside the landmass. For some ppl of the Faroe Islands, we find wrongly mapped places up to 25km offshore.

There are certain patterns in the data regarding the places per area and the places per 1000 population. Normally these numbers are rather stable, but we get some extreme cases, due to small sizes, large countries with few available places, or very small populations. For example, Greenland has a very low density per area, but since it also has a low population density, that value is slightly higher as also many small places are available. This is even more extreme for the Åland Islands, which have a very high data density per population. The area for Åland was not given in GeoNames and thus not calculated. It should be noted that on a global scale, the average of data per 1000 inhabitants is around 3 and values above 4 are



(a) POIs partitioned by accuracy (b) Places with population numbers

Figure 6: Indicator mapping

mostly small states or islands, which is demonstrated well in this dataset.

We want to note that Denmark has a surprisingly high percentage of populated places of all it places at 50%. Usually, this number is much lower because a multitude of other features are available. It still has comparatively good coverage regarding area and population. This leads to the hypothesis that there is an underrepresentation of natural geographical places in the dataset. Still high values at around 35% can be observed for Sweden and Finland.

We also look at available points of interest. For this, we have manually built a list of feature codes that qualify. This is often used to bootstrap location-based services or to populate maps. There are mostly usual values, of an amount of POIs making up around 40–70% of the amount of PPL for the larger countries, with variations for the smaller dependencies. But there are surprising massive outliers for Iceland and Norway (including Svalbard). These have over 13 respectively over 3 times the amount of POIs than populated places as shown in Table 2 and in Fig. 6a when compared to Fig. 3. Usually the POI numbers are much lower, so that they can only be a very first starting point, but in these cases, there is an unexpected richness. As we do not have provenance data available, we can only speculate that some of the constituent sources contains a huge national POI list.

We also see an interesting pattern in the number of population data annotations and the Wikipedia links. The prevailing trend is that countries with lower population density tend to have more annotations in these fields. This holds for the dependencies and also for Iceland. The outlier is Åland, which might be due to its less isolated geographical location. There is a strong correlation between both fields which might indicate that values are reimported from Wikipedia articles. Yet this correlation also hints at an issue for the population numbers. They are often used as a weighting factor in nearness calculations or rankings. Fig. 6b maps only those PPL that have the respective annotation, compared to all PPL in Fig. 3. This is a rather low number, around 6–24% for mainlands with an average of 12% and 9–53% with an average of 38% for dependencies.

5 Conclusion

The presented analysis furthers the understanding of the GeoNames dataset and its characteristics in the Nordic countries. We have taken this limited country list as an opportunity to dive deeper into the data, but this is easily scaled to the full world dataset. Together with the country discussion in Section 2 on manually identified issues, the developed indicators allow for a detailed analysis of gazetteer data.

There is a generally good availability of data, but strong variations between countries. Dependent territories, or small countries on a world scale, can be treated differently than expected. Yet there is no clear overall picture, in some cases the mainland, in other cases the dependencies come out ahead. Checking back to Fig. 2, the general availability of places follows population densities, and general geographical features are usually well covered. Previous work has argued that the real world is inadequately mirrored in geospatial datasets [12], which is certainly true for user-generated media content. The strongest variations are observed in the small dependencies. The strong linkage with Wikipedia for less populated places needs to be better examined to understand the reasons behind it and possible transfer processes to improve other places as well. The distribution of POIs shows strong variations, which can make GeoNames an unexpected source for this type of data.

In our future work, we will develop and improve additional indicators and examine their correlation beyond those discussed here and also examine correlation of individual places [5]. Additional work will concern the applicability of robust indicators in widely varying regions on a world-scale with widely varying quality of data. While the gazetteer data is supposed to be used on its own, we might have to examine certain external data to test some hypotheses about aspects of the data, for example external counts of cities or other feature types to assess completeness. We further have to consider the robustness of indicators for small or less populated places, which otherwise easily show extreme unexplained values. We then aim to make our results available online as a service. Another possible angle of approaching the issue would be to go down to a country scale and compare available data there. For example, for Norway, the Norwegian Mapping Authority makes place names available and this could act as a starting point for a comparison to ground truth that is not yet part of GeoNames.

In general, care needs to be taken when using geospatial data. Gazetteer use includes the choice of source as briefly discussed in the comparison of TGN and Geonames and assessing whether it has the right data structure and richness of data. This work cannot deliver clear advice on the data use. It will not pick a 'winner', but instead shows the range of quality issues to consider, how to identify them, and how to use these intrinsic indicators. The indicators have to be assessed individually based on the intended use case for the data. With these results, researchers and practitioners are able to easier gain insights and judge the data quality to their requirements when using this geospatial data.

References

[1] A. I. Abdelmoty and C. B. Jones. Towards maintaining consistency of spatial databases. In *CIKM'97*. ACM, 1997.

[2] D. Ahlers. Assessment of the Accuracy of GeoNames Gazetteer Data. GIR '13. ACM, 2013.

[3] D. Ahlers. Applying geographic information retrieval – an experience report on developing local search for a developing country. *Datenbank-Spektrum*, 14(1):39–46, 2014.

[4] C. Barron, P. Neis, and A. Zipf. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 2014.

[5] T. J. Brunner and R. S. Purves. Spatial Autocorrelation and Toponym Ambiguity. GIR '08, 2008.

[6] R. Devillers, A. Stein, Y. Bédard, N. Chrisman, P. Fisher, and W. Shi. Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS*, 14(4), 2010.

[7] J. T. Hastings. Automated conflation of digital gazetteer data. *Int. J. Geogr. Inf. Sci.*, 22(10), 2008.

[8] A. Henrich and V. Lüdecke. Measuring Similarity of Geographic Regions for Geographic Information Retrieval. In *ECIR '09*. Springer, 2009.

[9] L. L. Hill. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *ECDL '00*, 2000.

[10] L. Isaksen. Lines, damned lines and statistics: unearthing structure in Ptolemy's Geographia. *e-Perimetron*, 6(4), 2011.

[11] H. Manguinhas, B. Martins, and J. L. Borbinha. A Geo-Temporal Web Gazetteer Integrating Data From Multiple Sources. In *ICDIM*. IEEE, 2008.

[12] V. Murdock. Your mileage may vary: on the limits of social media. *SIGSPATIAL Special*, 3(2), 2011.

[13] Nordic Council of Ministers. *Nordic Statistical Yearbook 2014*. Copenhagen, 2014.

[14] L. A. Souza, C. A. Davis, Jr., K. A. V. Borges, T. M. Delboni, and A. H. F. Laender. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In *LA-WEB '05*. IEEE, 2005.

[15] J. Zhang and M. Goodchild. *Uncertainty in Geographical Information*. Taylor and Francis, 2002.