Computers, Environment and Urban Systems 64 (2017) 309–320

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A quantitative analysis of global gazetteers: Patterns of coverage for common feature types

Elise Acheson [a], Stefano De Sabbata [b], Ross S. Purves [a]

[a] University of Zurich, Department of Geography, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[b] Department of Geography, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom

ARTICLE INFO

Article history:
Received 21 September 2016
Received in revised form 28 January 2017
Accepted 13 March 2017
Available online 13 April 2017

Keywords:
Gazetteers
Data quality
GeoNames
Placenames
Geocoding

ABSTRACT

Gazetteers are important tools used in a wide variety of workflows that depend on linking natural language text to geographical space. The spatial properties of these data sources, such as coverage, balance, and completeness, affect the performance of common tasks such as geoparsing and geocoding. However, little attention has focused on how these properties vary in global gazetteers, particularly across country boundaries and according to feature types. In this paper, we present a detailed investigation of the spatial properties of two open gazetteers with worldwide coverage: GeoNames, and the Getty Thesaurus of Geographic Names (TGN). Using point density maps, correlations, and linear regressions, we analyze the global spatial coverage of each data source for the full set of features and for top feature types: populated places, streams, mountains, and hills. Results show wide discrepancies in coverage between the two datasets, sharp changes in feature type coverage across country borders, and idiosyncratic patterns dominated by a few countries for the more sparsely covered natural features. As more and more systems rely on recognizing and grounding named places, these patterns can influence the analysis of growing amounts of online text content and reinforce or amplify existing inequalities.

1. Introduction

Gazetteers play a central role in linking text to space, influencing a multitude of application outcomes through their use in tasks such as identifying placenames[1] in text, disambiguating placename references, and associating placenames with a geographical footprint and type information. Until recently, gazetteers were primarily produced top-down, typically as curated resources for placenames in a prescribed area such as a country. Today, with data easily stored and shared online, and vast quantities of data released as open data, the ways in which gazetteers are being produced and distributed is evolving. At one end of the spectrum remains a top-down, strongly regulated process, where organizations such as national mapping agencies produce gazetteers according to explicitly defined data quality standards and local laws. At the other end are crowdsourcing efforts collecting information about places from anyone who wishes to contribute, often largely relying on the notion of the 'wisdom of the crowd' principles for data quality (Goodchild & Li, 2012). Somewhere on this spectrum are two gazetteers with some level of data curation, nominally global coverage, but limited explicit information with respect to data quality: GeoNames (GeoNames, 2016), and the Getty Thesaurus of Geographic Names (TGN, 2016).

These gazetteers form the focus of the present paper. Perhaps because of their worldwide coverage and their ready availability, both are popular in many research projects and applications, with GeoNames arguably the most commonly used gazetteer today. Despite this popularity, there has been limited scrutiny of its contents, with attention typically limited to a particular region or country, and focused largely on populated place features rather than a broader set of feature types. Smart et al. (2010) mapped the overall coverage of GeoNames in Great Britain, contrasting it with national mapping agency data and crowdsourced datasets. Ahlers (2013) conducted a broader examination of data quality in GeoNames, identifying anomalies and quality indicators for populated places in Central America, Germany, and Norway. Looking at both GeoNames and TGN, De Sabbata and Acheson (2016) quantitatively compared their coverage for all features and populated places in Great Britain, finding the datasets less detailed and less balanced than national mapping agency data. Although these studies have revealed that coverage in these products is unbalanced even within individual countries, the overall picture remains unclear since to date, an in-depth systematic global analysis, looking across country boundaries and at a range of feature types across gazetteers, has not been carried out.

An initial exploration of such properties examined global coverage of GeoNames alone and explored the distribution of a single feature type

E-mail addresses: elise.acheson@geo.uzh.ch (E. Acheson), s.desabbata@leicester.ac.uk (S. De Sabbata), ross.purves@geo.uzh.ch (R.S. Purves).

[1] We use the more vernacular term *placename* interchangeably with *toponym* in this paper.

http://dx.doi.org/10.1016/j.compenvurbsys.2017.03.007

(populated places) as a function of population (Graham & De Sabbata, 2015). Expanding on this work, we undertake a detailed comparative investigation of the global spatial properties of both GeoNames and TGN. We not only look at the full datasets, but also present a worldwide analysis of coverage for the four most frequent feature types in GeoNames, matched with corresponding types in TGN: populated places, streams, mountains, and hills. These four feature types account for a large portion of the full datasets in both gazetteers, particularly populated places which comprise over a third of all the data in both GeoNames and TGN. As for streams, mountains, and hills, they are among the most common natural features found in the data sets, and in the case of mountains, the most commonly referenced examples of a geographic feature in empirical experiments (Smith & Mark, 2001). Understanding the global coverage of these named natural features is particularly important in the context of any work analyzing the distribution of common toponym types (Campbell, 1991) and analysis of texts containing references to natural features (Moncla et al., 2014). For both gazetteers, we examine and compare feature distributions at fine, medium, and coarse granularities.

As discussed in the review that follows, *coverage* and *balance* are two pivotal quality indicators to assess the fitness for use of gazetteers for many common tasks. We therefore pose the following research questions:

1. How do GeoNames and TGN compare in terms of overall global coverage and balance?

2. How are important feature types in GeoNames and TGN distributed globally, and how do they compare in terms of coverage and balance?

We review previous work focusing on gazetteer properties, sources, and quality, as well as tasks in which gazetteers play a role. We then introduce in more detail the properties of the two gazetteers we analyzed, before setting out the analysis methods to characterize and compare GeoNames and TGN. Our results are presented as both graphical and numerical data, before we discuss their implications, particularly in terms of the suitability of these data sources for relevant tasks. We conclude with a list of key gazetteer shortcomings and propose future research focused on addressing these.

2. Gazetteers

> "*There is remarkable diversity in approaches to the description of geographic places (...)*".
>
> [Linda Hill, Georeferencing, p. 94]

Gazetteers are resources that store structured information about places, minimally providing name, type, and location (or footprint) information for each place or record (Hill, 2000; Mostern et al., 2016). Each record may also contain other attributes such as alternative names, population information for populated places, and containment relationships - for example which country or region the place is in. Records may contain links to matching records in other datasets. These 'linked data' records are ones deemed to be about the same place through a matching process that, for instance, compares text, positional, and type information across resources (Sehgal et al., 2006; Smart et al., 2010). Placenames have in fact become a central node in linked open data, with GeoNames lying at the center of the linked open data cloud diagram (Schmachtenberg et al., 2014), demonstrating the efficacy of placenames as a way of relating information in the developing semantic web.

2.1. Gazetteer sources and production

Gazetteers have traditionally been produced in a top-down process, most commonly by national mapping agencies to serve as official placename resources for a defined area of interest such as a country, sometimes under specific legal or regulatory conditions. For example, Ordnance Survey (OS) produces the OS 1:50k gazetteer (2016) (and more recently, OS Open Names) for the extent of Great Britain, and SwissTopo produces SwissNames 3D (2016) for the extent of Switzerland. In the case of the United States, examples include a national resource for domestic names, the Geographic Names Information System (GNIS, 2016), developed by the U.S. Geological Survey, and an international resource for foreign names, the GEOnet Names Server (GNS, 2016), developed by the National Geospatial-Intelligence Agency.

As well as general purpose gazetteers, typically created by national mapping agencies and other government authorities, purpose-built gazetteers are created for a wide range of purposes. Among these are the TGN, a structured gazetteer with the aim of improving access to art, architecture, and material culture by enabling indexing. Due to its focus on these topics, historical names are important elements of the TGN, allowing links of historical artifacts to be made between present day locations and texts describing them in a historical context.

More recently, gazetteers have also been produced by incorporating bottom-up methodologies, where data is collected from multiple sources and integrated. Two heavily used global spatial datasets, OpenStreetMap and GeoNames, are produced this way: their sources include authoritative data, such as those described above where licensing permits, but also original data contributed by individuals, also known as volunteered geographic information (VGI) (Goodchild, 2007). Further still along the spectrum from top-down to bottom-up production are approaches to creating structured gazetteers using only crowdsourced data, through the extraction, analysis, and merging of multiple sources. One such example, the Gazetiki project, mined Wikipedia and Panoramio data to automatically create a gazetteer, relying on linguistic cues, search hits, and the GeoNames feature type hierarchy for entity typing (Popescu et al., 2008).

A complementary body of research focuses on both augmenting and enriching existing gazetteers and the generation of so-called meta-gazetteers to build better resources, whether more complete (with more features, or with richer annotation for existing features), or deemed more suitable for a particular task (Kessler et al., 2009; Smart et al., 2010). In one example using VGI, Gao et al. (2017) present a framework for efficiently creating new gazetteer entries from large numbers of user-tagged photographs, many of which contain feature types like 'park', 'museum', or 'river' as tags. Finally, OpenStreetMap has also been used as a gazetteer source directly, or to augment existing placename resources (de Oliveira et al., 2016; Hess et al., 2014; Yin et al., 2014).

As feature types are one of the three basic requirements of a gazetteer entry (Hill, 2000), any work seeking to integrate or augment gazetteers faces the challenge of assigning appropriate types to features, and potentially having to align different feature type ontologies to each other. A common use case in gazetteer conflation is to consider feature type information as evidence of (dis)similarity when trying to detect whether records are about the same feature (Fu et al., 2005; Hastings, 2008; Smart et al., 2010). However, this is a challenging task since feature types may vary widely between gazetteers, and the process of feature type alignment is itself complex (Janowicz & Keßler, 2008; Zhu et al., 2016). These difficulties are illustrated by for example Fu et al. (2005) who established "equivalence links" between feature type hierarchies, but found that strong constraints on feature type alignment led to poor performance. The underlying problem is further illustrated by Smart et al. (2010) who noted that even in national mapping agency data, large proportions of features were simply classified as "other". Zhu et al. (2016) recognize this challenge and combine top-down ontology analysis with bottom-up data-driven methods using spatial signatures related to instances of feature types to explore alignment issues in GeoNames, TGN and DBPedia Places.

2.2. Gazetteer quality

As introduced above, a wide variety of gazetteer and gazetteer-like resources exist, all of which may vary with respect to their data quality. In exploring gazetteer quality, we take as a starting point the so-called famous five, as listed by the US Federal Geographic Data Committee: attribute accuracy, positional accuracy, logical consistency, completeness, and lineage (Guptill & Morrison, 1995). Van Oort (2005) added to this list semantic accuracy, fitness for use, and temporal quality. In the more specific context of gazetteers, Leidner (2004) proposed seven criteria for gazetteer quality, a list extended and refined by Hill (2006, p.107) (Table 1).

An implicit quality which is not explicitly listed in the criteria above, but often mentioned in discussions of the nature of gazetteers, is coverage. In this paper we define the *coverage* of a resource as the feature density across space ('spatial coverage', as in Hill, 2006, p. 144). We define *balance* as in Table 1 as the uniformity of a resource across its scope of coverage, including the uniformity of its currency, accuracy, granularity, and richness of annotation. Thus balance and coverage are clearly related, since balance depends on the feature density across space (coverage) across the resource. As an example, a gazetteer covering features down to street-level detail in London but only down to neighborhood detail in Paris would be less balanced, but have better coverage in London, than a resource covering only neighborhood-level features in both cities. In the analysis that follows, we primarily focus on balance as the uniformity of coverage, as commenting on the uniformity of currency, accuracy, and richness of annotation is beyond the scope of this work.

An important upstream factor impacting gazetteer quality is the way in which the datasets are produced, as previously described. In the case of top-down datasets from mapping agencies, the organizations producing these resources typically ensure adherence to, and document, data quality standards, for example by sending surveyors out into the field in a structured manner, such that errors or omissions may be assumed to be randomly distributed in the dataset. However, this is not the case for crowdsourced data, which tend to show bias - that is, data quality which varies non-randomly as a function of the properties of the underlying space. For example, in a seminal paper on VGI quality, Haklay (2010) conducted a systematic analysis of OpenStreetMap data quality in terms of positional accuracy and completeness of street network data, and found geographical biases towards both urban (and therefore more populated) and more affluent regions in England, which have important implications for the balance of datasets produced through crowdsourcing in general. For gazetteers, such quality comparisons with authoritative datasets are possible for regions that have national mapping agency counterparts. However, a quality evaluation of global gazetteers must proceed in other ways, since no authoritative global database of placenames exists.

Table 1
Gazetteer quality criteria from Hill (2006).

Criterion	Description
Availability	"Degree to which the gazetteer is freely available and not limited by restrictive conditions of use"
Scope	"Small communal database, regional/national coverage, or worldwide coverage"
Completeness	"Degree to which the scope of the gazetteer is covered completely"
Currency	"Degree to which the gazetteer has incorporated changes"
Accuracy	"Number of detectable errors in names, footprints, and types"
Granularity	"Includes large, well-known features only or features of all sizes and those that are less well known"
Balance	"Uniform degree of detail, currency, accuracy, and granularity across scope of coverage"
Richness of annotation	"Amount and detail of descriptive information, beyond the basics of name, footprint, and type"

2.3. Using gazetteers

In general, gazetteers play a key role in tasks linking text to space, with applications ranging from disaster response or disease tracking through social media geolocation (Dredze et al., 2013; Zhang & Gelernter, 2014), to historical and literary text analysis (Cooper & Gregory, 2011; Southall et al., 2011). By enabling the organization of data according to geographical space, textual datasets can be spatially analyzed, opening up a wide range of possibilities for descriptive and predictive modelling of previously aggregated data.

More specifically, a number of distinct tasks require gazetteers or benefit from their use. A first task is the detection of placenames (or more broadly, geographic references[2]), for which a common approach is to look for textual matches between placenames in the gazetteer and each word, N-gram, or candidate placename in the text (Leidner & Lieberman, 2011). Thus, gazetteers and the placenames they contain influence whether a word or sequence of words is classified as a location in the first place (Leveling, 2015; Purves et al., 2007). Second, gazetteers are important in disambiguating placenames, as many disambiguation strategies use gazetteer data such as geometry, type, and attribute data (e.g. population) to rank candidates (Buscaldi, 2011). In fact, gazetteers typically determine whether a placename can be considered 'geo/geo' ambiguous, as this type of ambiguity in practice is defined as when a placename matches more than one entry in a gazetteer (Amitay et al., 2004; Zhang & Gelernter, 2014). Thus, toponym ambiguity is heavily influenced by the resources used, where gazetteers covering larger areas and finer granularities raise the potential for multiple matches (Buscaldi, 2011) - in other words, potentially increasing recall but decreasing precision.

The link between text and space is completed in a further task, focused on selecting a relevant geometry (footprint) for an input placename (Hill, 2000). This is crucial in geographical information retrieval, both to obtain geographical representations of textual queries and to index documents according to the geographical space they refer to (Purves et al., 2007). Of particular importance, setting aside geoparsing performance considerations, are the types of geometries available to model placenames. The simplest and most common geometry used to represent a placename is a point, but other geometries may be more appropriate according to place granularity and the nature of the reasoning to be carried out with the data (Alani et al., 2001; Guo et al., 2008). Currently, however, both GeoNames and TGN provide only latitude, longitude tuples as points referring to features in their free, downloadable versions.

An increasingly frequent text-to-space task is the geolocation of social media users or content. On Twitter for example, users may indicate a home location on their profile in free-text form. Thus, matching this text field to a gazetteer entry is often a key first step to analyzing Twitter users' attitudes and beliefs according to their location, using for instance sentiment analysis. Furthermore, since only approximately 1% of Twitter posts are explicitly tagged with GPS coordinates (Hecht et al., 2011), geocoding the posts themselves (or using the geocoded profile location as a proxy) can make more content available for analysis in use cases such as disaster response (Zhang & Gelernter, 2014) or disease tracking (Dredze et al., 2013). Jurgens et al. (2015) emphasize that gazetteer choice impacts the performance of geolocating Twitter users from the textual profile location information, with GeoNames returning matches for 500 k users and DBPedia only 75 k. Clearly coverage in general, and in particular balance with respect to the underlying properties of interest, are key indicators in assessing the quality of the results of such processes.

[2] Geographic references are considered a superset of placenames, which includes not only placenames but also place codes, such as addresses or postal codes, and more complex expressions, such as composite expressions like "North of Lake Ontario".

3. Data & methods

In order to obtain high recall for placename detection in text, it is important to use a gazetteer that provides both good coverage and completeness. To increase precision and limit the impact of ambiguity, it is furthermore fundamental to select a balanced gazetteer with an appropriate level of detail. This section first presents GeoNames and TGN, then describes the methods used to assess the coverage and balance of the two datasets.

3.1. GeoNames

GeoNames is arguably the most-used placename data source today, widely cited in academic works. It is not hard to understand why considering its unique combination of desirable properties: it contains over 10 million entries worldwide (coverage), is freely available online (availability), and has daily data exports (currency). Where its properties become less clear is concerning balance, data precision, completeness, and lineage. With respect to lineage, GeoNames consists of data originating from a variety of sources, some official sources and some individual contributors, but the source(s) for each particular record is not provided in the free version. This lineage issue muddles the already unclear picture with respect to balance, precision, and completeness. For instance, the true precision of a record may vary depending on the source or simply be unknown. As for balance and completeness, it is unclear to what extent each country or region is captured in the dataset, and thus the coverage may rather vary as a function of data availability for a particular area rather than as a function of the true concentration of named geographical features at those locations.

In addition to the standard elements of name, type, and geometry, GeoNames also provides for many entries a rich set of structured information including alternate names and spellings, population information for populated places, and hierarchical information such as containing administrative areas including countries. Geometries provided in the free gazetteer data download are latitude-longitude points for each feature, including for features with very large extents like countries and lakes. All features are classified according to a two-level type hierarchy consisting of a 9-feature-class top level and a 645-feature-code sublevel, with a short description provided for most feature codes. However, the distribution of feature counts is dominated by just a few of these 645 feature codes, with for example 3.4 million features having the code 'PPL' for 'populated place'. Though the feature type hierarchy has two levels, in practice a third-level is arguably encoded in the feature codes themselves, with for example 'STM' standing for 'stream', 'STMI' for 'intermittent stream', and 'STMIX' for 'section of intermittent stream'. We make use of this information when considering feature type alignments in Section 3.4.

3.2. TGN

TGN is a gazetteer resource developed for cultural heritage applications, freely available as linked open data since 2015. Unlike GeoNames, it is curated, has a stated focus on places of historical significance, and an intended use for "cataloguing, research, and discovery of art historical, archaeological, and other scholarly information" (TGN, 2015). Its coverage is nominally global and, appropriately for its historical focus, also covers a temporal range from "prehistory to the present".

With over 1.4 million entries of named places around the world, including over half a million populated places, TGN is a useful resource for text analysis, particularly for texts of a historical nature (Overell & Rüger, 2008; Smith & Mann, 2003). Similarly to GeoNames, TGN records provide names, feature type information, latitude-longitude coordinates, as well as hierarchical information where appropriate. Records also generally include sources and contributors, and may also contain descriptive notes and dates for historical places, as well as linkages between places signifying relationships such as 'successor of' and 'distinguished from'. TGN uses feature types from the Art & Architecture Thesaurus (AAT, 2017) type hierarchy, also from the Getty Research Institute. This type vocabulary and hierarchy features many more explicit levels than GeoNames and provides information about type semantics not only through a definition, but also with information about overlaps with other types and with lineage information about the type itself and its position in the hierarchy. In practice, TGN's feature type distribution is however also heavily skewed to a small number of types and again features a widely used category for population centers of all sizes, known as 'inhabited places'. In Fig. 1, the 'hills' type is shown in the AAT hierarchy with four sub-types, but while in our dataset 25,756 TGN features have 'hills' as a primary type, no TGN features have as primary type 'foothills', 'hillocks', 'hummocks (hills)', or 'knolls'.

3.3. Gazetteer quality criteria

In Table 2, we present a comparison of our two gazetteers of study and two representative national mapping agency gazetteers, OS 50k for Great Britain and SwissNames 3D for Switzerland, along nine quality criteria. These quality criteria are as in Table 1, with the addition of lineage (from the famous five of spatial data quality) and precision (which is documented explicitly for our national mapping agency data), and the removal of accuracy, which in practice is not available because it requires testing against a reference dataset. Though coverage does not appear explicitly in the table, our definition relates it to scope, completeness, granularity, and balance as discussed above.

These four datasets all share the characteristic of being freely available (though licensing conditions vary), but in the other dimensions, TGN and GeoNames are more similar to each other than either is to the national mapping agency datasets. Importantly, completeness and balance are unknown for GeoNames and TGN, not being explicit aims for either data source. Precision of the feature coordinates is not documented, though TGN states that their coordinates are approximate only. On the other hand, GeoNames and TGN offer the advantage of nominally rapid update cycles, with GeoNames providing daily data downloads online. The authoritative datasets have slower release cycles, consistent with a process requiring extensive quality control to ensure completeness, precision, and balance of their contents. Our analysis of GeoNames and TGN aims to enable informed statements about balance and coverage.

3.4. Feature type selection and matching

Full data snapshots of both GeoNames and TGN were obtained on June 30th 2015. Though differences in feature density are to be expected, since GeoNames has almost 10 times the number of features as TGN,

Fig. 1. AAT type hierarchy containing the feature type 'hills', used in TGN. (Source: AAT, 2017).

Table 2
Gazetteer quality criteria for four gazetteers.

Criterion	GeoNames	TGN	OS 50 k	SwissNames 3D
Availability	Free	Free	Free	Free
Scope	Worldwide	Worldwide	Great Britain	Switzerland
Completeness	?	?	✓	✓
Currency	Daily	Two weeks	Annual	Annual
Precision	Varied	Approximate	1k grid cell	0.2 m–3 m
Granularity	Medium to fine	Medium	Medium to fine	Fine
Balance	?	?	Uniform	Uniform
Richness of annotation	Medium	Rich for portion	Medium	Medium
Lineage	Various sources	GNIS, experts	OS maps	SwissTopo maps

we expect datasets to become more similar as they tend to more accurately sample the real world, and in the future, as they become more integrated through the practice of linked data. Our snapshot of TGN was taken shortly after the product was launched in its linked open data form, and before it had had time to be propagated into GeoNames.

A first step towards our goal to compare GeoNames and TGN was to match countries from one dataset to the other, which was done manually and resulted in a set of 237 common countries which could be used to select features by country from the raw gazetteer data. A second step was to select feature types for analysis, a process primarily driven by looking at the most common features types in GeoNames and matching these to types in TGN.

Aligning feature types between gazetteers is a complex problem, as resources may have different feature type ontologies, the same words may take different meanings across resources, and the meaning of a particular word may also itself vary among geographical regions and individuals (Zhu et al., 2016). In our alignment process we considered type names, definitions, feature type hierarchies and, since our analysis of coverage and balance relies on feature counts, type frequencies in each dataset. Given the potential influence of feature alignment on any results comparing gazetteer content, we furthermore carried out a series of sensitivity tests to explore such effects, where we varied the alignment choices using one-to-one, one-to-many, many-to-one, and many-to-many links between GeoNames and TGN types.

Fig. 2 shows the most frequent feature types in each gazetteer (using unique feature codes for GeoNames), the results of our feature type selection and alignment, and the number of features of each type selected for analysis.

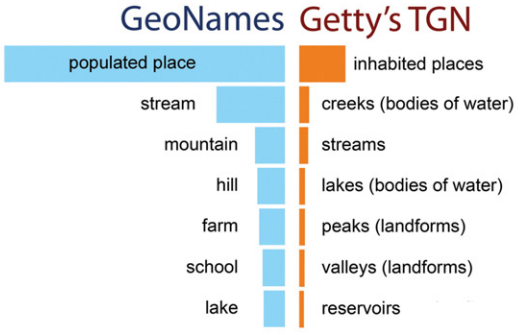
Based on our sensitivity tests, we aligned 'populated place' (PPL) and its implicit sub-types (PPL*) in GeoNames to the 'inhabited places' place type in TGN, and proceeded similarly for the other types, as shown in Fig. 2. Whereas GeoNames has a large number of features typed with implicit sub-types (for example, PPLL is used for 141,798 features and STMI for 177,531 features), TGN has very low counts for sub-types, perhaps because of its richer hierarchical type structure and its use of a preferred place type for each feature with optional secondary types.

3.5. Analysis methods

For a multi-scale, feature-type specific understanding and comparison of the global coverage of features in GeoNames and TGN, three analysis scales were chosen. At the finest level, 10x10km cells with 30 km neighborhoods were chosen as the point aggregation unit for global point density maps. For the full set of features ('all') and for the four selected feature types ('populated places', 'mountains', 'hills', 'streams'), maps were produced using the ArcGIS Point Density tool, in the Goode Homolosine Land equal area projection, resulting in 10 global maps. These maps allow for a visual overview of the global coverage of features in each of GeoNames and TGN, and a visual comparison of this coverage between the two gazetteers for each feature type. Furthermore, by exploring differences in coverage, it is possible to gain some insight into balance (for example where feature density varies greatly in a dataset across national boundaries).

For the second part of the analysis, features were aggregated at two coarser-grained spatial units: 100 × 100 km cells, and individual countries. Aggregated counts for both spatial units were calculated for each feature type (five in total, including 'all'). The country of each feature was available directly as attribute data in both datasets, thus country counts could be obtained by summing the number of features with each country attribute. For the 100 × 100 km cells, counts of features in each cell were obtained through a spatial join. For each aggregation unit (2) and each feature type (5), counts in the two datasets were plotted and correlation coefficients computed using ranks (Kendall's method), since values were not normally distributed.

Based on the outcomes of the second part of the analysis, linear models relating the two gazetteers were calculated to establish descriptive, quantitative relationships between their coverage. Through linear models, feature coverage can be compared not just on the basis of



Feature type name	GeoNames feature codes	TGN place types	GeoNames count	TGN count	GeoNames % of total	TGN % of total
all	(all)	(all)	10,203,772	1,439,138	100.00%	100.00%
populated places	PPL*	'inhabited places'	3,767,721	564,533	36.92%	39.23%
streams	STM*	'streams'	1,027,913	108,445	10.07%	7.54%
mountains	MT*	'mountains'	390,418	25,975	3.83%	1.80%
hills	HLL*	'hills'	351,926	25,680	3.45%	1.78%

Fig. 2. Most frequent feature types for GeoNames and TGN (top); Types selected for analysis and their respective counts and frequencies in GeoNames and TGN (bottom). (Only features considered for the country and graticule analyses are included in the table. Excluded features, not in any of the matched countries, accounted for 0.26% of all features GeoNames and 0.15% in TGN.) * Indicates that all GeoNames feature codes taking this base were included.

rank, but also magnitude, by describing a proportional relationship between feature counts in corresponding cells or countries.

4. Results and interpretation

4.1. Point density maps

We first present the global point density maps of all the features in each of GeoNames and TGN (Fig. 3). These maps show the density of placenames in each data source using quantiles calculated on the GeoNames data for both maps to ease comparison. Overall, for both data sources, we observe higher densities of placenames in regions such as Europe and in the eastern United States, and much lower densities in entire continents such as South America and Africa. These global differences in coverage are particularly marked in the TGN data, where there is widespread data scarcity in South America with the exception of Chile, and likewise for Africa aside from a concentration of placenames in Egypt, a place of great historical significance. Another observation common to both maps, but particularly pronounced in TGN, is that in many places coverage seems particularly uneven across country borders. For example, GeoNames placename density is markedly different between Norway and Sweden, and in TGN the eastern and southern borders of Germany are clearly distinguishable due to a sudden drop in coverage, and similarly for the border between Canada and the United States. India is described in relatively similar detail by GeoNames and TGN, whereas the amount of features in neighboring Nepal and Sri Lanka is strikingly different in the two datasets. These results also indicate that balance is an issue in both datasets, since these patterns are unlikely to reflect real toponym density.

Breaking down the datasets by feature type helps shed light on whether these observations are consistent across, or driven by, particular types. Fig. 4 shows small multiples of coverage in GeoNames and TGN for each of the four feature type data subsets, starting at the top with the most frequent feature type, populated places, down to the least frequent (in GeoNames), hills. This sequence illustrates that as a feature type decreases in numbers overall in a gazetteer, its global coverage also becomes sparser, concentrated in a smaller area of the globe. Whereas in GeoNames populated places show non-zero density across large swaths of the globe, most 10 km density cells are zero (white) for mountains and hills. This observation is more pronounced in TGN,

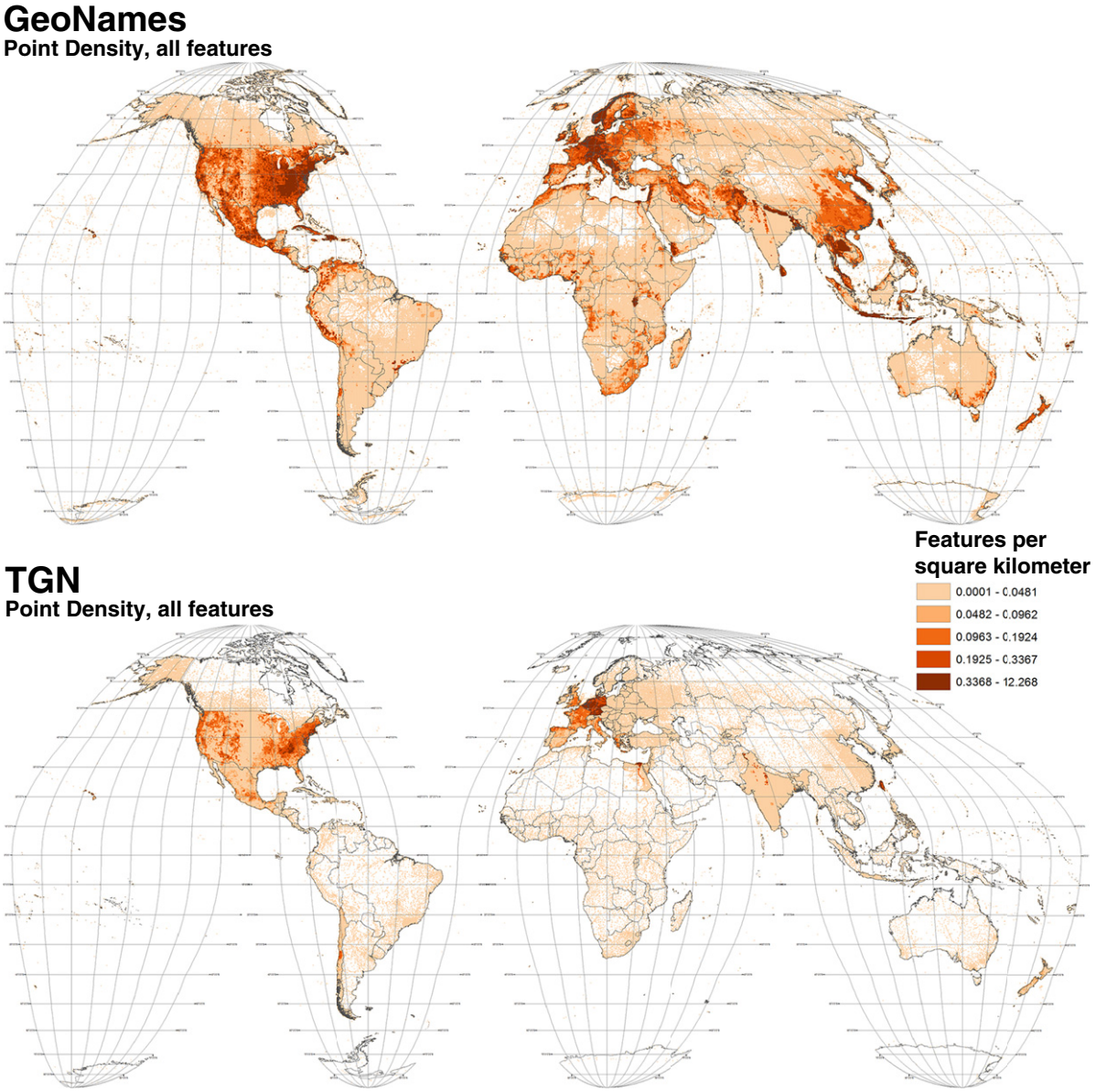


Fig. 3. Point density maps for all features in GeoNames (top) and TGN (bottom) rendered in terms of GeoNames quantiles, in the Goode Homolosine Land projection.

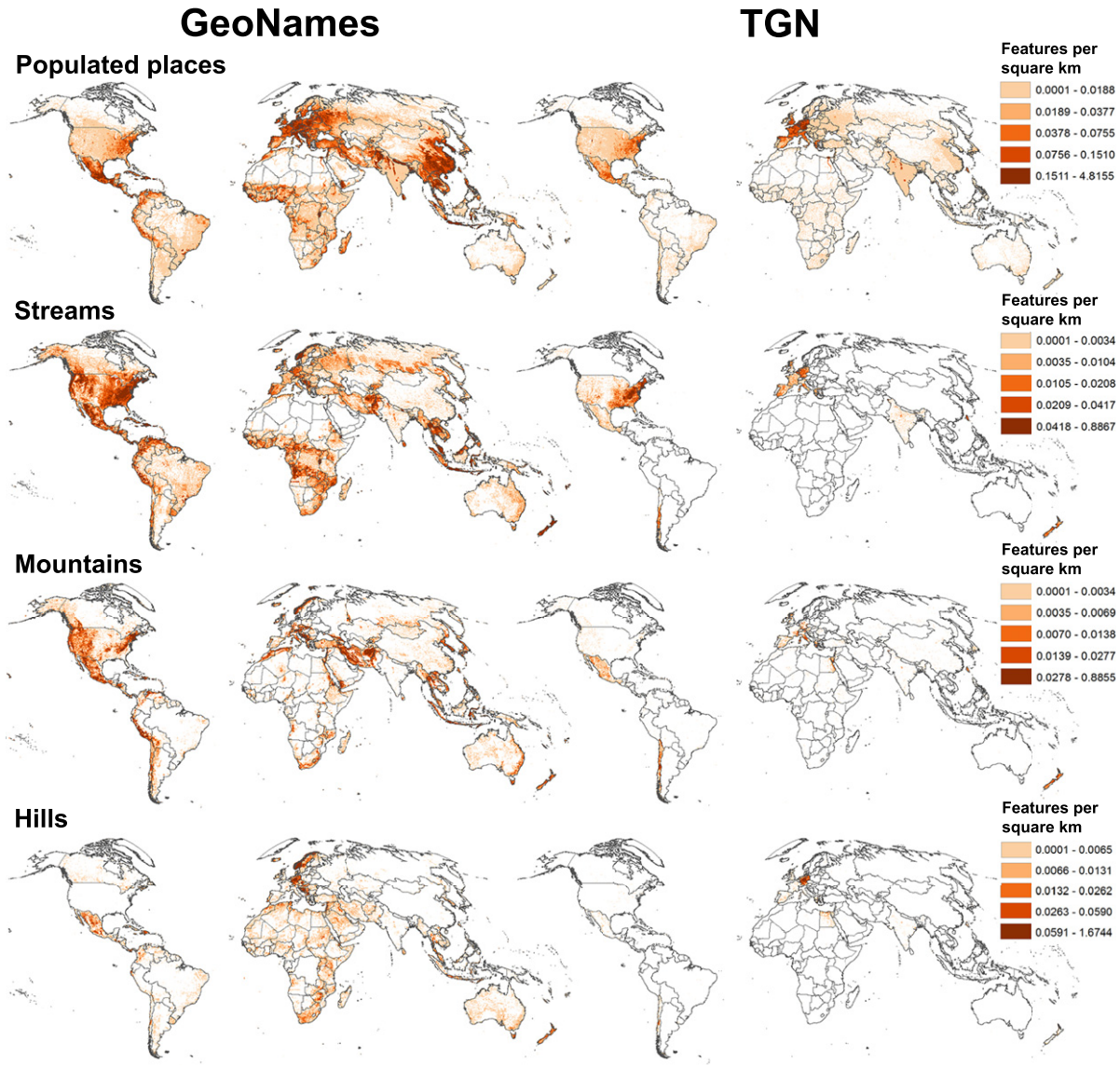


Fig. 4. Point density maps by gazetteer (GeoNames, TGN) and feature type (populated places, streams, mountains, hills), rendered in terms of GeoNames quantiles, Goode Homolosine Land projection.

which overall has about ten times less data than GeoNames in terms of all features, but is also slightly more skewed towards populated places than GeoNames with 39% of its features populated places but only 1.8% hills, compared to 37% and 3.4%, respectively, for GeoNames (Fig. 2). While the TGN populated places map again clearly shows the resource's emphasis on Europe and the United States, the global TGN hills map shows just how few features of this type are catalogued.

In order to address whether the coverage of features in GeoNames and TGN corresponds well to the true distribution of named features in the world, it is helpful to consider feature types in isolation. For the best-represented feature type, populated places, population density may be a reasonable proxy for density of named populated places at our scales of analysis (Graham & De Sabbata, 2015). In both the GeoNames and TGN populated places maps, some areas of visibly high feature concentration correspond with areas of high population density such as continental Europe and the North East of the United States, but other populous regions seem comparatively less well-covered, including India, Brazil, and the African continent in general. TGN in particular appears to have a strong focus on Western Europe and the United States in terms of catalogued populated places, and relative data poverty through China, South-East Asia, and parts of the Middle East compared to GeoNames.

Shifting down from the best-represented feature type to the sparsest type in our analysis, hills, the uneven coverage becomes more extreme, with virtually no hills catalogued in the United States (both in GeoNames and TGN), and a very high concentration of hills in Norway (GeoNames) and Germany (GeoNames and TGN). These maps again show sharp changes in feature type coverage at country borders, such as the US-Mexico border for hills (GeoNames) and mountains (TGN), and the US-Canada border for streams (GeoNames). A closer look at the individual point features shows a dearth of mountains in TGN for Switzerland (Fig. 5b), a country renowned for its mountains, and virtually no hills in GeoNames for the US (Fig. 5c) against an abundance in neighboring Mexico. Thus it appears from these maps that the country unit is an important driver of global coverage by feature type.

To explore this behavior where these feature types were seemingly well sampled, counts by type were plotted for the ten countries with the highest numbers of the given types (Fig. 5a). These bar charts in all cases show that one or a few countries dominate the distribution for a particular feature type, for instance Norway for hills in GeoNames

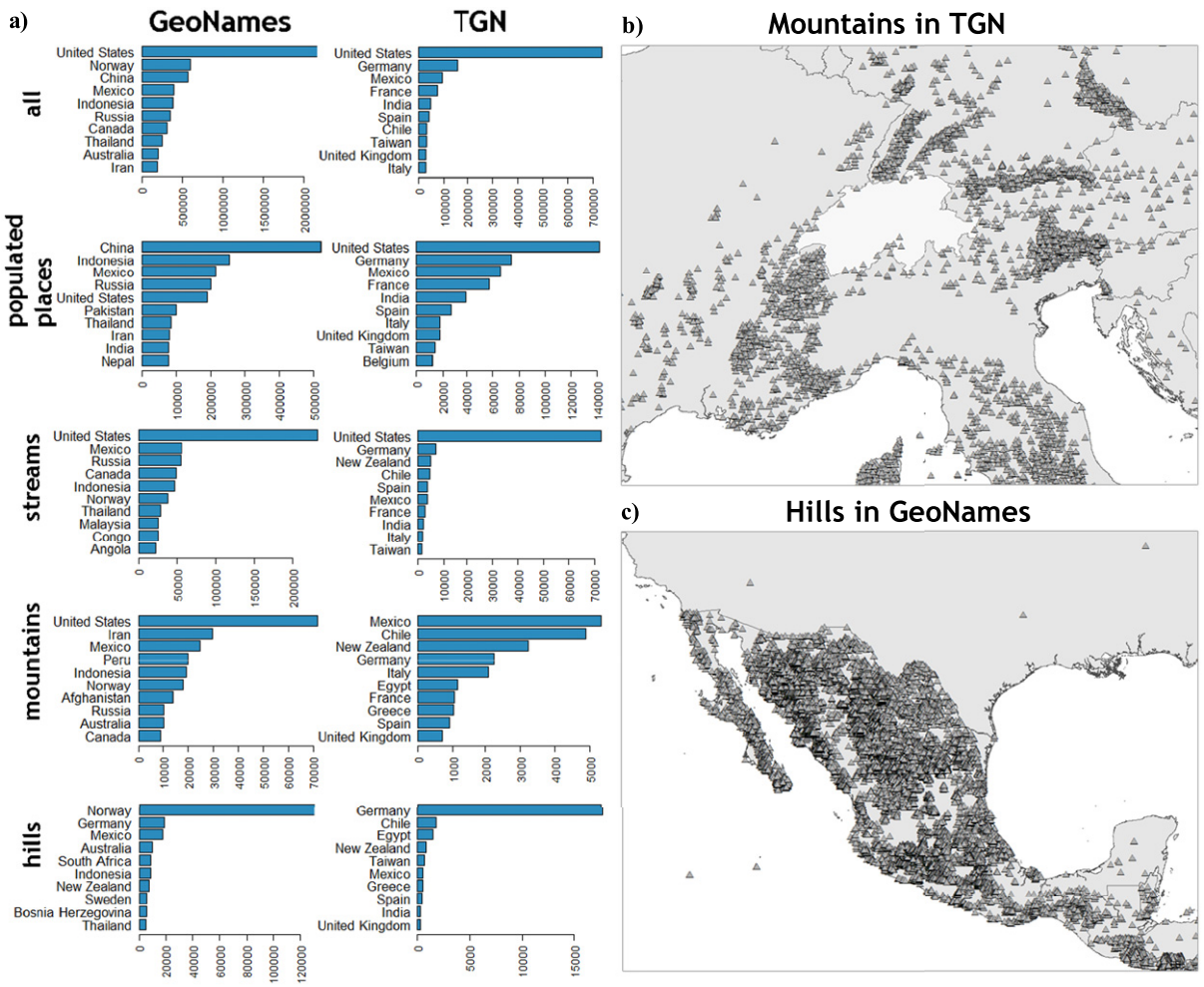


Fig. 5. a) Bar charts of the ten countries with the most features of each type; b) all mountain features in TGN for Switzerland (white) and surrounding countries; c) all hill features in GeoNames for Mexico and the southern United States.

or Germany for hills in TGN. Furthermore, GeoNames and TGN tell different stories about the underlying distribution of data, especially for the natural feature types. Indeed, Norway unequivocally tops the list for hills in GeoNames, yet does not even make the top ten in TGN. Similarly, China is conspicuously absent from the top ten list for populated place features in TGN, but tops the list for populated places in GeoNames. All of these observations again emphasize the lack of balance for all the analyzed feature types in both GeoNames and TGN.

4.2. Correlations between GeoNames and TGN

To compare coverage between GeoNames and TGN systematically, counts of features were spatially aggregated according to the coarser, meaningful unit of countries, and the finer unit of 100 × 100 km cells (created in the equal area Goode Homolosine Land projection). Corresponding counts were analyzed using Kendall's tau rank correlation, with the results shown in Table 3.

Table 3
Kendall's tau correlation coefficients between GeoNames and TGN for countries and raster cells.

Features	Countries (N = 237)		Raster (N = 51,996)	
	M (neither 0)	Kendall's tau	M (not both 0)	Kendall's tau
All	237	0.7138*	20,665	0.6383*
Populated places	235	0.7015*	14,188	0.5377*
Streams	29	0.3004+	13,182	0.2322*
Mountains	159	0.4853*	9868	0.2449*
Hills	74	0.2584^	8704	0.0424*

* $p < 0.00001$.
^ $p < 0.01$.
\+ $p < 0.05$.

For the country counts, a pair was included in the rank correlation calculation when neither country had a count of zero to avoid having artificially high correlation coefficients due to matching pairs with very low or zero counts. For the much larger number of raster cells, a pair was included in the rank correlation calculation when either dataset had a non-zero count, striking a balance between keeping spurious pairs with no data and dropping meaningful pairs.

The correlation coefficients for feature counts by country show relatively strong positive relationships when accounting for both all features and populated places, and a weaker but still positive relationship for mountains. For streams and hills the relationships were much weaker. The correlation coefficients for raster cell counts show similar patterns, with the highest correlations for all features and populated places, and much weaker or no relationships for streams, mountains, and hills. Comparing correlation coefficients for countries and raster cells, we note that the coefficients are greater for countries than raster cells in all five cases, meaning a stronger relationship exists at the country level than for the finer raster cells.

4.2.1. Sensitivity to feature type alignment

In order to ensure our results were robust to changes in feature type alignment, we performed sensitivity tests where we selected different combinations of feature types from GeoNames and/or TGN and

repeated the correlation analysis. We varied GeoNames feature subsets by selecting only a single dominant feature code (PPL, STM, MT) for a type rather than also including its implicit sub-types (PPL*, STM*, MT*). As for TGN, of particular interest was the inclusion of types featured in relatively high numbers in the collection: creeks and peaks.

Table 4 illustrates the results of these sensitivity tests. In all but one case (where streams + creeks from TGN are included) correlations are similar and statistically significant. In the case of creeks, 99.6% of all creeks in the TGN are located in the USA. These results suggest that our choice of feature alignment is robust.

4.3. Linear models

The Kendall rank correlation coefficients indicated the existence of varying degrees of positive relationships between aggregated features counts in GeoNames and TGN, depending on type and unit of analysis (country or raster cells). In order to analyze these relationships considering not only the rank, but also the magnitude of feature counts, linear regression models were used.

Based on the positively skewed distribution for both GeoNames and TGN aggregated counts, we used log-log regression models, arbitrarily using GeoNames counts as the independent variable and TGN counts as the dependent variable. While raw counts suggest that TGN contains only 14% of the amount of features in GeoNames, the first two stages of the analysis suggest that the relationship is more complex. The coefficients (b) of the regression models provide a better estimate of the quantitative relationship, whereas the coefficient of determination (R^2) provides a measure of model fitness.

As such, the linear models presented in this section should not be interpreted as explanatory or predictive, but rather as descriptive. These two data sources are clearly independently produced, but as they are both sampling geographical features from the real world, relating their feature counts can give us insight into how similar their coverage is. The selection of one variable as dependent and the other as independent is purely arbitrary, and does not affect the interpretation of the results.

A first linear model was constructed, using countries as the unit of analysis, based only on the independent and dependent variables mentioned above, but the model did not meet the assumption of homoscedasticity of the residuals. This was confirmed through a Breusch-Pagan test, which was significant ($p < 0.001$). The log-log scatter plot of TGN vs GeoNames for all features showed an interesting pattern where a group of countries showed relatively high counts in TGN compared to the remaining countries, as depicted in Fig. 6. From this scatter plot, we identified this set of 15 countries as 'high coverage' countries: United States, Germany, Mexico, France, India, Spain, Chile, Taiwan, United Kingdom, Italy, Egypt, Greece, Belgium, New Zealand, and the Netherlands. We found that these same countries were also well covered across the four feature types we analyzed. The data point for the Faroe Islands was excluded as a statistical outlier. A dummy variable (i.e., Boolean indicator) was thus introduced, taking the value 1 when the country is a member of the so-called *HighCoverage* set, and 0 otherwise. We then used linear models of the form:

$$\ln(\mathrm{TGN}) = b_0 + b_1 \ln(\mathrm{GeoNames}) + b_2\ \mathrm{HighCoverage} + \varepsilon$$

The linear model then obtained is presented in the first section of Table 5 (Model 1).

Model 1 in Table 5 is robust and fit. The residuals are normally distributed (Shapiro-Wilk test, $W = 0.99, p > .01$), satisfy the homoscedasticity assumption (Breusch-Pagan test, $BP = 4.27$, $p > .05$), and the errors are independent (Durbin-Watson test, $DW = 1.88$, $p > .05$). No statistically influential cases were identified. The number of features in GeoNames, combined with the distinction between high coverage and low coverage countries, accounts for 87% ($F(2,232) = 803.6$, $p < .001$) of the variation in the number of features in TGN, when aggregated by country.

This model illustrates how the number of features in TGN in high coverage countries is of the same order of magnitude as counts in GeoNames, having over 60 ($e^{4.13} = 62.18$) as much content as low coverage countries. Still even in the high coverage group, an increase of 100 features in GeoNames corresponds to an increase of only 71 features in TGN. The model also supports the two assumptions discussed in this section. First, a clear relationship exists between the two datasets. Second, two distinguishable groups of countries are present in TGN, one featuring an amount of content comparable with what can be found in GeoNames, and another group covered in far less detail. The characteristics of the high coverage group are further discussed in the next section.

Based on these results, we investigated the possibility that linear models might hold at a different scale of analysis. We again used the 100×100 km raster cells that we used in the correlation section and relate the number of features in each cell for GeoNames and TGN. We tested a linear model similar to the model presented above, assigning each cell to a country (or no country where necessary) and re-creating the dummy variable *HighCoverage* as above. All cells with a count of zero for either gazetteer were discarded, due to the logarithmic nature of the models. Given the large number of cells remaining (13,910) we also tested the same model with smaller random samples (150 and 1500 cells), which resulted in consistent outcomes and significance levels.

The linear model based on the raster cell counts (Model 2 in Table 5) is very similar to the model based on country counts presented above (Model 1), when disregarding cells not associated with any country. The number of features in GeoNames (combined with the dummy variable) accounts for 82% ($F(2, 11{,}152) = 0.00026$, $p < 0.001$) of the variation in the number of features in TGN. Similarly to the model above, high coverage cells contain about thirty ($e^{3.40} = 29.96$) times as much content as low coverage cells in TGN. The residuals are normally distributed, but they show heteroscedasticity, and errors are not independent. This is most probably due again to a different behavior between countries in the high coverage and low coverage sets. Including the cells which are not associated to any country (Model 3 in Table 5) leads to a lower $R^2 = 0.75$ ($F(3, 13{,}906) = 0.00014$, $p < 0.001$), while the remaining values are stable.

Finally, we tested similar linear models based on number of features per country, for the feature types populated places and mountains. These linear models showed a relatively strong relationship between the two gazetteers, but the residuals of the models were not normally distributed. It is also important to note that these models showed substantial linear relationships only when excluding countries with no features in at least one gazetteer, as TGN in particular contains no features of those two types for many countries (counts of zero are found in 34 countries for populated places, 99 countries for mountains, 172 countries for hills, and 209 countries for streams).

Table 4
Kendall's tau correlation coefficients between GeoNames and TGN for countries using different feature type alignments. Alignments used in the rest of the paper are highlighted.

Feature type name	GeoNames feature codes	TGN place types	M (neither 0)	Kendall's tau
Populated places	PPL*	Inhabited places	235	0.7015^{α}
Populated places	PPL only	Inhabited places	232	0.6996^{α}
Streams	STM*	Streams	29	0.3004^{+}
Streams	STM only	Streams	29	0.3265^{+}
Streams	STM*	Streams + creeks	42	0.1757
Mountains	MT*	Mountains	159	0.4853^{α}
Mountains	MT only	Mountains	159	0.4763^{α}
Mountains	MT*	Mountains + peaks	164	0.4946^{α}

$^{\alpha}p < 0.00001$, $^{\wedge}p < 0.01$, $^{+}p < 0.05$.

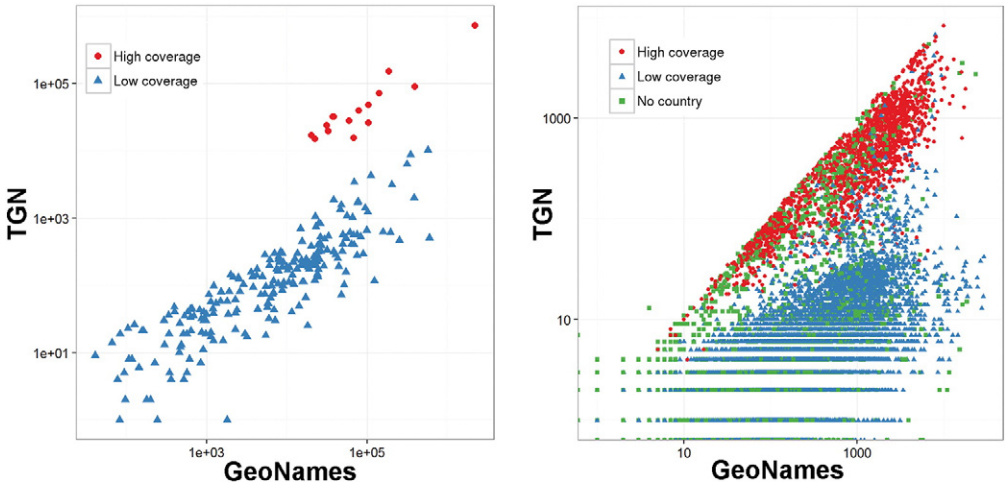


Fig. 6. Log-log scatter plot of feature counts in TGN as a function of counts in GeoNames in matching countries (left) and 100 × 100 km cells (right).

4.4. High coverage countries

The linear models from the previous section showed a systematic pattern where a group of countries were consistently better represented in TGN than the others. We termed this group of countries 'high coverage' and computed the Kendall's tau correlation coefficients for this list of 15 countries in isolation, as we did for all countries in Section 4.2. The resulting values, presented in Table 6, show strong highly significant positive relationships between these high coverage countries in TGN and GeoNames. This is the case not only when considering all features, which was expected based on the models presented in the previous section, but also for the four feature types taken into account.

The correlation coefficients are not only higher in each case for only the list of 15 as compared to the full set of countries with non-zero feature counts, but the values are also much less variable, all ranging from about 0.70 to 0.87. Overall these correlation coefficients show that not only do these high coverage countries have more data in TGN than the rest of the countries, with few exceptions, but also that where TGN coverage is high, the data resembles GeoNames much more, in terms of the feature counts per country.

Table 5
Linear models of feature counts in TGN as a function of feature counts in GeoNames.

Models	Adj. R^2	Coefficients (b)	Std. Error	Std. coef. (β)	p (sign.)
1) Country counts					
TGN	0.87				
Constant		−1.48	0.226		<0.001
GeoNames		0.71	0.025	0.68	<0.001
High coverage (dummy)		4.13	0.218		<0.001
2) Raster counts (countries only)					
TGN	0.82				
Constant		−1.22	0.027		<0.001
GeoNames		0.51	0.005	0.43	<0.001
High coverage (dummy)		3.40	0.022		<0.001
3) Raster counts (all non-zero)					
TGN	0.75				
Constant		−1.26	0.027		<0.001
GeoNames		0.51	0.005	0.48	<0.001
High coverage (dummy)		3.59	0.026		<0.001
No country (dummy)		0.67	0.022		<0.001

5. Discussion

"Are my data fit for purpose?" This is a crucial question in science, and in this paper we set out to illustrate how it applies to the use of two global gazetteers. Coverage and completeness play a fundamental role in the detection of placenames in text, particularly in terms of recall, while balance is important in limiting ambiguity and improving precision. The maps presented in the first part of the previous section illustrate the skewness and idiosyncrasies of two major global gazetteers, GeoNames and TGN. The correlation and regression analysis presented above shows how the two gazetteers do not provide a coherent description of the world's toponyms, and identify regions and scales where similarities do exist.

As TGN is a historically-focused, curated resource possessing about a tenth of the overall quantity of features in GeoNames, differences in placename density must exist. Our methods allowed us to identify a small list of countries whose placenames are catalogued in more detail in TGN than the others - but these still only possess a portion of the quantity of data provided by GeoNames. Among those countries is the United Kingdom, where TGN provides only half the placenames available in GeoNames, which in turn are just a fraction of the data provided by the Ordnance Survey (De Sabbata & Acheson, 2016). Beyond such countries with detailed coverage, the number of features available in TGN drops to two orders of magnitude lower than in GeoNames, consistent with the results reported by Ahlers (2013) for the specific case of Honduras. Thus, TGN coverage is not only sparser overall, but more idiosyncratic than GeoNames and thus also less balanced.

Both GeoNames and TGN differ fundamentally from gazetteers produced by mapping agencies, which adhere to defined data quality standards including completeness and balance, and whose contents can be assumed to vary largely as a function of the true density of named features in the area of interest. Indeed, one of the most challenging issues, which we can only address peripherally in this paper when exploring global gazetteers, is completeness. Our results suggest that GeoNames and TGN are both far from complete given variation in coverage and feature type balance, and based on our results we can suggest regions

Table 6
Kendall's tau correlation coefficients for high coverage countries in TGN.

Features	N	Kendall's tau
All	15	0.6952*
Populated places	15	0.8667*
Streams	15	0.8476*
Mountains	15	0.7905*
Hills	15	0.7656*

* $p < 0.001$.

where the datasets may be particularly incomplete. However, understanding whether toponyms are missing because they have not been mapped, or are simply not used, requires us to also consider both the underlying physical landscape and variation in toponym usage across cultures and languages (Burenhult & Levinson, 2008). As for balance, GeoNames coverage varies partly as a function of the availability of data, with rapid changes in coverage possible overnight when new datasets are integrated, thus affecting the balance of the resource. TGN represents a historically-focused view of the world, but even then some coverage artifacts seem tied to open data integration such as the relative abundance of data in New Zealand and of hills in Germany. Future work studying the lineage of the features could explain some of these observations and perhaps reveal crucial information on common sources between the gazetteers.

Indeed, our results clearly highlight the role of institutional - usually national - open data, as the coverage offered for feature types shows abrupt changes across national borders. Throughout the analysis, countries consistently appeared as the strongest driver of variation. Even at the finest analysis scale, the influence of the country unit was visible in the global maps for both GeoNames and TGN. Therefore, studies assessing the quality of gazetteers through comparison with authoritative datasets cannot simply be generalized to other study areas, particularly across borders. Furthermore, special care should be taken when working in a multi-national study area (for example Europe), as taking gazetteer data as-is across borders will typically result in variations in balance with respect to the sampling of the true distribution of named places, and results of tasks such as geoparsing are more likely to reflect gazetteer properties, rather than true spatial variation. In any work seeking to augment gazetteers, combine gazetteers, or create meta-gazetteers (Gao et al., 2017; Grossner et al., 2016; Smart et al., 2010), an important aim should be not to introduce further bias in coverage or balance.

Another important observation is that coverage and correlations between feature types quickly decreases for all except populated places, which account for over a third of all data in GeoNames or TGN. Our maps show that as overall numbers in one gazetteer for a particular feature type decrease, coverage across the globe becomes not only sparser, but less well correlated, more idiosyncratic and thus less balanced, even for the common natural feature types streams and hills. An analogy can be made with crowd-sourced mapping projects, which have been shown to suffer from biases that are not only geographic, but also thematic. Bégin et al. (2013) notes that natural features tend to suffer from lower positional accuracy than man-made features in VGI, and finds that users in mapping tasks show preferences for mapping certain feature types. Similarly, our results show that the representation of natural features is of a comparatively lower quality than populated places in GeoNames and TGN, the data sources being more focused on populated places globally. Thus, we suggest national data should be used preferentially when dealing with these feature types, particularly since national borders in any case are a strong driver of coverage, removing any advantage of nominally seamless, global, datasets. Furthermore, though our sensitivity tests indicate that our feature type alignment choices for very common features are robust in calculating correlations, the importance of alignment (Zhu et al., 2016) in matching less frequent features types is likely to have a bigger influence on gazetteer quality assessment.

Though our results clearly indicate that balance for natural feature types is poor, quantifying this further requires the use of meaningful proxies for named features. Possible approaches, which might also give insights into completeness, might imply modelling expected densities of named features as a function of morphological properties (Hengl & Reuter, 2008) and relating this to existing authoritative gazetteer data. Finally, current research in information geographies suggests that most datasets are heavily skewed towards the Global North and marginalize the Global South (Graham et al., 2015) – including those used to train machine learning algorithms – rendering any aim for a global, rich, balanced gazetteer a formidable challenge.

6. Conclusions

The present paper has illustrated the important role played by gazetteers in the current data revolution, and argues that fitness for use of global gazetteers has been neglected, despite their very common application to a wide range of tasks. Our results highlight the skewness and idiosyncrasies of these gazetteers whose coverage and balance, especially at the level of feature types, varies widely, and is best predicted by national borders. These results also suggest that the politics and economics of open data (Kitchin, 2014) can have a significant impact on gazetteers, and thus on geographic analysis and automated data processing.

Although making an informed decision on fitness for purpose is straightforward with top-down, authoritative gazetteers through the documented quality criteria, this is in practice much more difficult with global gazetteers such as GeoNames and TGN. In such a situation, the questions that researchers using the gazetteers discussed in this paper should ask themselves are: what components of a pattern extracted by linking text to space reflect meaningful patterns in the underlying data, and what simply reflects skewness and idiosyncrasies of the gazetteer used in the analysis?

Haklay's (2010) conclusion that "places where population is scarce or deprived are, potentially, further marginalised by VGI exactly because of the cacophony created by places which are covered" seems to merit an extension to the realm of global gazetteers. Less connected and less developed countries are currently further marginalized in global gazetteers, as are natural features, drowned out by populated places. The cacophony of information is further intensified by algorithmic data analysis, which through the use of gazetteers produce even more data about the places already covered.

Acknowledgements

This work was supported by an STSM Grant from COST Action IC1203.

References

AAT (2017). Art & architecture thesaurus online. Retrieved from http://www.getty.edu/research/tools/vocabularies/aat/ (Accessed 25.01.2017)

Ahlers, D. (2013). Assessment of the accuracy of GeoNames gazetteer data. *Proceedings of the 7th Workshop on Geographic Information Retrieval* (pp. 74–81). New York, NY, USA: ACM GIR '13 10.1145/2533888.2533938

Alani, H., Jones, C. B., & Tudhope, D. (2001). Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, *15*(4), 287–306. http://dx.doi.org/10.1080/13658810110038942.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging web content. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 273–280). New York, NY, USA: ACM SIGIR '04 10.1145/1008992.1009040

Bégin, D., Devillers, R., & Roche, S. (2013). Assessing Volunteered Geographic Information (VGI) quality based on contributors' mapping behaviours. *Proceedings of the 8th International Symposium on Spatial Data Quality ISSDQ* (pp. 149–154).

Burenhult, N., & Levinson, S. C. (2008). Language and landscape: A cross-linguistic perspective. *Language Sciences*, *30*(2–3), 135–150. http://dx.doi.org/10.1016/j.langsci.2006.12.028.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, *3*(2), 16–19. http://dx.doi.org/10.1145/2047296.2047300.

Campbell, J. C. (1991). Stream generic terms as indicators of historical settlement patterns. *Names*, *39*(4), 333–366. http://dx.doi.org/10.1179/nam.1991.39.4.333.

Cooper, D., & Gregory, I. N. (2011). Mapping the English Lake District: A literary GIS. *Transactions of the Institute of British Geographers*, *36*(1), 89–108. http://dx.doi.org/10.1111/j.1475-5661.2010.00405.x.

De Sabbata, S., & Acheson, E. (2016). Geographies of gazetteers in Great Britain. *Proceedings of the 24th GIS Research UK Conference, GISRUK 2016. Greenwich, UK.*

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)* (pp. 20–24).

Fu, G., Jones, C. B., & Abdelmoty, A. I. (2005). Building a geographical ontology for intelligent spatial search on the web. *Proceedings of IASTED international Conference on Databases and Applications (DBA-2005)* (pp. 167–172). Innsbruck, Austria: ACTA Press.

Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on hadoop. *Computers, Environment and Urban Systems, Geospatial Cloud Computing and Big Data*, *61*(Part B (January)), 172–186. http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.004.

GeoNames (2016). Retrieved from http://www.geonames.org (Accessed 05.10.2016)

GNIS (2016). United States Board on Geographic Names: Domestic names. http://geonames.usgs.gov/domestic/index.html (Accessed 05.10.2016)

GNS (2016). NGA GEOnet Names Server. http://geonames.nga.mil/gns/html/ (Accessed 05.10.2016)

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. http://dx.doi.org/10.1007/s10708-007-9111-y.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*(May), 110–120. http://dx.doi.org/10.1016/j.spasta.2012.03.002.

Graham, M., & De Sabbata, S. (2015). Mapping information wealth and poverty: The geography of gazetteers. *Environment and Planning A*, *47*(6), 1254–1264. http://dx.doi.org/10.1177/0308518X15594899.

Graham, M., De Sabbata, S., & Zook, M. A. (2015). Towards a study of information geographies: (Im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, *2*(1), 88–105. http://dx.doi.org/10.1002/geo2.8.

Grossner, K., Janowicz, K., & Keßler, C. (2016). Place, period, and setting for linked data gazetteers. In J. R. Mostern, & H. Southall (Eds.), *Placing names: Enriching and integrating gazetteers*. Bloomington, IN: Indiana University Press.

Guo, Q., Liu, Y., & Wieczorek, J. (2008). Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, *22*(10), 1067–1090. http://dx.doi.org/10.1080/13658810701851420.

Guptill, S. C., & Morrison, J. L. (1995). *Elements of spatial data quality*. Elsevier Science Limited.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. http://dx.doi.org/10.1068/b35097.

Hastings, J. T. (2008). Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, *22*(10), 1109–1127. http://dx.doi.org/10.1080/13658810701851453.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's Heart: The dynamics of the location field in user profiles. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237–246). New York, NY, USA: ACM CHI '11 10.1145/1978942.1978976

Hengl, T., & Reuter, H. I. (Eds.). (2008). *Geomorphometry: Concepts, software, applications. Developments in Soil Science*, *vol. 33*. Elsevier 772 pp.

Hess, B., Magagna, F., & Sutanto, J. (2014). Toward location-aware web: Extraction method, applications and evaluation. *Personal and Ubiquitous Computing*, *18*(5), 1047–1060. http://dx.doi.org/10.1007/s00779-013-0718-3.

Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. *Research and advanced technology for digital libraries* (pp. 280–290). Springerhttp://link.springer.com/chapter/10.1007/3-540-45268-0_26

Hill, L. L. (2006). *Georeferencing: The geographic associations of information*. The MIT Press.

Janowicz, K., & Keßler, C. (2008). The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, *22*(10), 1129–1157. http://dx.doi.org/10.1080/13658810701851461.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Kessler, C., Maué, P., Heuer, J. T., & Bartoschek, T. (2009). Bottom-up gazetteers: Learning from the implicit semantics of geotags. *GeoSpatial semantics* (pp. 83–102). Springerhttp://link.springer.com/chapter/10.1007/978-3-642-10436-7_6

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Vancouver, Canada: Sage.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. *Workshop on geographic information retrieval, Sheffield, UK*.

Leidner, J. L., & Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), 5–11.

Leveling, J. (2015). Tagging of temporal expressions and geological features in scientific articles. *Proceedings of the 9th Workshop on Geographic Information Retrieval* (pp. 6: 1–6:10). New York, NY, USA: ACM GIR '15 10.1145/2837689.2837701

Moncla, L., Gaio, M., & Mustière, S. (2014). Automatic itinerary reconstruction from texts. *Automatic itinerary reconstruction from texts. 8728*. (pp. 253–267). Vienna, Austria: Springer International Publishinghttp://link.springer.com/10.1007/978-3-319-11593-1

Mostern, R., Southall, H., & Berman, M. L. (Eds.). (2016). *Placing names: Enriching and integrating gazetteers*. Indiana University Press.

de Oliveira, M. G., Campelo, C. E. C., de Souza Baptista, C., & Bertolotto, M. (2016). Gazetteer enrichment for addressing urban areas: A case study. *Journal of Location Based Services*, *10*(2), 142–159. http://dx.doi.org/10.1080/17489725.2016.1196755.

OS 1:50k gazetteer (2016). Ordnance survey 1:50 000 scale gazetteer. Retrieved from https://www.ordnancesurvey.co.uk/business-and-government/products/50k-gazetteer.html (Accessed on 05.10.2016)

Overell, S., & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, *22*(3), 265–287. http://dx.doi.org/10.1080/13658810701626236.

Popescu, A., Grefenstette, G., & Moëllic, P. A. (2008). Gazetiki: Automatic creation of a geographical gazetteer. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 85–93). New York, NY, USA: ACM JCDL '08 10.1145/1378889.1378906

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Gaihua, F., et al. (2007). The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science*, *21*(7), 717–745. http://dx.doi.org/10.1080/13658810601169840.

Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim. 2014. "Adoption of the linked data best practices in different topical domains." In *The Semantic Web – ISWC 2014*, 245–60. Lecture Notes in Computer Science 8796. Springer International Publishing. http://link.springer.com/chapter/10.1007/978-3-319-11964-9_16.

Sehgal, V., Getoor, L., & Viechnicki, P. D. (2006). Entity resolution in geospatial data integration. *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems* (pp. 83–90). New York, NY, USA: ACM GIS '06 10.1145/1183471.1183486

Smart, P. D., Jones, C. B., & Twaroch, F. A. (2010). Multi-source toponym data integration and mediation for a meta-gazetteer service. *Geographic Information Science* (pp. 234–248). Springerhttp://link.springer.com/chapter/10.1007/978-3-642-15300-6_17

Smith, B., & Mark, D. M. (2001). Geographical categories: An ontological investigation. *International Journal of Geographical Information Science*, *15*(7), 591–612. http://dx.doi.org/10.1080/13658810110061199.

Smith, D. A., & Mann, G. S. (2003). Bootstrapping toponym classifiers. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1* (pp. 45–49). Stroudsburg, PA, USA: Association for Computational Linguistics HLT-NAACL-GEOREF '03 10.3115/1119394.1119401

Southall, H., Mostern, R., & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, *5*(2), 127–145. http://dx.doi.org/10.3366/ijhac.2011.0028.

SwissNames 3D (2016). Retrieved from http://www.mont-terri.ch/internet/swisstopo/en/home/products/landscape/swissNAMES3D.html (Accessed 05.10.2016)

TGN (2015). Getty Thesaurus of geographic names: About the TGN. Retrieved from http://www.getty.edu/research/tools/vocabularies/tgn/about.html (Accessed 05.10.2016)

TGN (2016). Getty thesaurus of geographic names online. Retrieved from http://www.getty.edu/research/tools/vocabularies/tgn/ (Accessed 05.10.2016)

Van Oort, P. (2005). *Spatial data quality: From description to application*. Delft: Netherlands Geodetic Commission.

Yin, J., Karimi, S., & Lingad, J. (2014). Pinpointing locational focus in microblogs. *Proceedings of the 2014 Australasian Document Computing Symposium* (pp. 66:66–66:72). New York, NY, USA: ACM ADCS '14 10.1145/2682862.2682868

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, *9*(December). http://dx.doi.org/10.5311/JOSIS.2014.9.170.

Zhu, R., Hu, Y., Janowicz, K., & McKenzie, G. (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*.