# Spatial Signatures for Geographic Feature Types: Examining Gazetteer Ontologies using Spatial Statistics

Rui Zhu, Yingjie Hu, Krzysztof Janowicz, Grant McKenzie
STKO Lab
Department of Geography
University of California Santa Barbara

## Abstract

Digital gazetteers play a key role in modern information systems and infrastructures. They facilitate (spatial) search, deliver contextual information to recommender systems, enrich textual information with geographical references, and provide stable identifiers to interlink actors, events, and objects by the places they interact with. Hence, it is unsurprising that gazetteers, such as GeoNames, are among the most densely interlinked hubs on the Web of Linked Data. A wide variety of digital gazetteers have been developed over the years to serve different communities and needs. These gazetteers differ in their overall coverage, underlying data sources, provided functionality, and also their geographic feature type ontologies. Consequently, place types that share a common name may differ substantially between gazetteers, whereas types labeled differently may, in fact, specify the same or similar places. This makes data integration and federated queries challenging, if not impossible. To further complicate the situation, most popular and widely adopted geo-ontologies are lightweight and thus under-specific to a degree where their alignment and matching become nothing more than educated guesses. The most promising approach to addressing this problem and thereby enabling the meaningfully integration of gazetteer data across feature types, seems to be a combination of top-down knowledge representation with bottom-up data-driven techniques such as feature engineering and machine learning. In this work, we propose to derive indicative spatial signatures for geographic feature types by using spatial statistics. We discuss how to create such signatures by feature engineering and demonstrate how the signatures can be applied to better understand the differences and commonalities of three major gazetteers, namely DBpedia Places, GeoNames, and TGN.

**Keywords:** Spatial Signatures, Geo-Ontologies, Gazetteer, Spatial Statistics, Linked Data

# 1 Introduction and Motivation

Digital gazetteers, i.e., structured dictionaries of geographical places, are a crucial backbone component of modern information systems and cyber-infrastructures. They support geographic information retrieval, deliver contextual cues to recommender systems, enrich textual information with geographical references, disambiguate places with similar names, provide insights into historical events, and offer stable and global identifiers to interlink data within and across data hubs (Alani et al., 2001; Rice et al., 2012; Janowicz and Keßler, 2008; Schlieder et al., 2001; Twaroch et al., 2008). This last aspect is of rapidly growing importance for global knowledge graphs such as Linked Data (Bizer et al., 2009; Janowicz et al., 2012) where places act as Nexuses that weave together statements, called (RDF) triples, about actors, events, and objects. To give a concrete example, Figure 1 shows a fragment of an exploratory *follow-your-nose* search linking together Horatio Nelson and Federico Carlos by the Battle of Trafalgar which took place at the Cape of Trafalgar as well as the Assault on Cadiz (which took place at Cadiz). The Cape of Trafalgar is also the place of death of Nelson who died on deck of the HMS Victory which is the oldest naval ship still in commission and located in a dry dock at Portsmouth, England, thereby linking Portsmouth and Trafalgar.

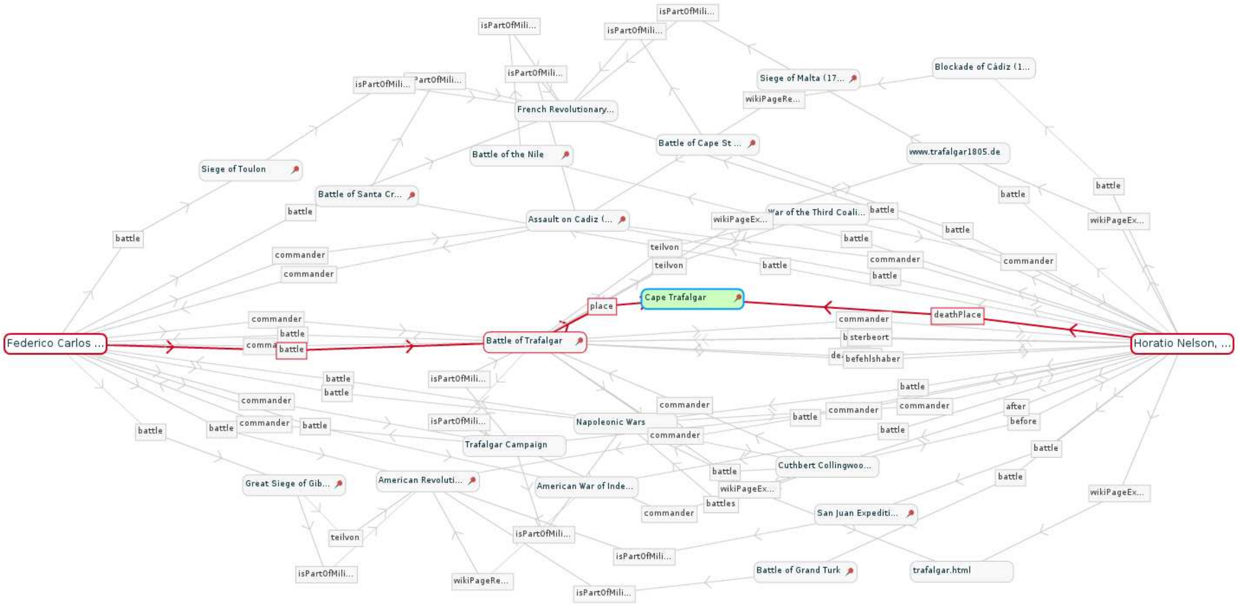


Figure 1: Exploratory *follow-your-nose* search for relations between Horatio Nelson and Federico Carlos.

This role of places for the linkage and integration of entities is also reflected by the popular Simple Event Model (Van Hage et al., 2011) and the fact that the GeoNames gazetteer is the second most interlinked hub[1] on the Web of Linked Data. Furthermore, the most popular hub, DBpedia, contains nearly 1 million places and millions of entities directly linked to these places. Other Linked Data gazetteers include the Pleiades gazetteer for ancient world studies, the Ordnance Sur-

[1] See http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.pdf for graphical representation of popular datasets and their interlinkage.

vey gazetteer of the UK, UCSB's Alexandria Digital Library Gazetteer (ADL), as well as the Getty Thesaurus for Geographic Names (TGN). Finally, as people frequently use place names (instead of coordinates) to refer to places, gazetteers act as an important interface connecting informal human discourse with formal geographic representations in information systems.

While the functionality offered by these gazetteers varies greatly, they all share three core elements, namely toponyms, i.e., (alternative) place names (N), geographic feature types (T), and spatial footprints (F) (Hill, 2000; Goodchild and Hill, 2008). Some gazetteers provide further data, e.g., about temporal scopes, spatial (and platial) containment, related geographic features, population counts, and so forth. A gazetteer's key capabilities can be specified by three common operations: lookup (N $\rightarrow$ F), type-lookup (N$\rightarrow$T), and reverse-lookup (F($\times$T)$\rightarrow$N). The first case, for instance, corresponds to a query for the *spatial footprint* of Cape Trafalgar, the second to the *type* of Cape Trafalgar, and the third one to the names of places at the Strait of Gibraltar that are of type *Cape* (Janowicz and Keßler, 2008).

However, there is no common geographic feature type ontology and thus each gazetteer uses its own typing schema (Keßler et al., 2009). For example, the type *Mountain* may be used by one gazetteer to represent mountain peaks, whereas another gazetteer may use it to refer to mountain ranges. Some gazetteers also group mountain ranges to mountain systems, while others do not. One gazetteer may distinguish between hills and mountains, while another does not support this distinction. Yet another gazetteer may also introduce types such as *Seamount* for mountains that rise from the seafloor without reaching the surface. The most critical cases, however, are those where the type labels used by two or more gazetteers are very similar but the underlying conceptualizations differ dramatically or cases where the labels differ but the types are the same. A well known example are the types *Nation* and *Country* in TGN and ADL. A query for countries yields 165 features in ADL while TGN returns merely 11 results as TGN uses the type *Nation* instead and has reserved *Country* for specific cases such as the divisions of the United Kingdom into Scotland, Britain, and so forth (Janowicz and Keßler, 2008).

Understanding the semantic heterogeneity among gazetteers is a prerequisite for query federation, data integration, conflation, and many other key tasks underlying modern cyber-infrastructures for GIS research and applications. On the one side, global or large-scale geographic studies often require the integration of gazetteers from different countries and authorities, which may have different definitions for the same terms. On the other side, understanding the semantics of places can also help in selecting a suitable gazetteer that fits the particular requirements of an application. The same argument can be made for recommender systems. For example, a historical gazetteer may contain *hotels* that were involved in significant historical events. However, such a gazetteer may not be suitable (or can be incomplete) for a hotel search engine whose objective is to find hotels in which tourists may stay.

The huge variety of geographic feature type definitions introduced by spatial, platial, temporal, cultural, and legal factors is an ideal case for techniques such as ontology alignment and matching (Euzenat et al., 2010). Unfortunately, most geo-ontologies are lightweight and thus under-specific to a degree where their alignment becomes nothing more than educated guesses as existing tools and methods are forced to default back to simple string similarity measures, such as Levenshtein distance, or network similarity measures such as structural equivalence.

In this work, we take a radically different, bottom-up approach and propose using *spatial semantic signatures* to understand the semantics of geographic feature types. *Semantic signatures* are an analogy to spectral signatures in remote sensing (Janowicz, 2012). The underlying idea is that geographic feature types can be characterized by spatial, temporal, and thematic *bands* mined from heterogeneous data sources. In previous work, we have studied temporal and thematic bands and how they jointly form signatures for micro and meso-scale features such as Points Of Interest (McKenzie et al., 2015). Here, we focus on *feature engineering*[2] by means of spatial statistics. Compared with the traditional methods from ontology engineering our approach reveals differences and similarities that cannot be uncovered otherwise.

**The contributions of this paper are twofold:**

- From a methodological viewpoint, this paper presents a statistical framework for understanding the semantics of geographic feature types in gazetteers from a data-driven perspective.
- From an application-centric viewpoint, we engineer a variety of statistical features and showcase their application to type similarity by analyzing and comparing three leading gazetteers: DBpedia Places, GeoNames, and Getty Thesaurus of Geographic Names (TGN).

The remainder of this paper is organized as follows. Section 2 highlights existing work related to the semantics of gazetteers, focusing on the topic of semantic interoperability. Section 3 provides a brief introduction and necessary background on the gazetteers used in this research. Section 4 presents the methodological details, i.e., the engineered features and used statistics. Section 5 applies the proposed approach to the three gazetteers and presents our findings on their semantic differences and similarities. Finally, section 6 summarizes our work and outlines future directions.

## 2 Related Work

One strong motivation for studying the semantics of geographic feature types in digital gazetteers is to facilitate the interoperability between multiple gazetteers. Once the meaning of place types is understood, one can align and integrate multiple gazetteers to support more advanced, federated queries and conflate data from different sources. Gazetteer interoperability is not a new research topic. In fact, it was extensively discussed during a National Center for Geographic Information and Analysis (NCGIA) specialists meeting at Santa Barbara in December 2006 (Goodchild and Hill, 2008).

A simple approach to understanding the meaning of place types is to consider their textual labels. Most gazetteers provide a rich list, taxonomy, or even ontology of feature types which are labeled with natural language terms. For example, the Alexandria Digital Library gazetteer contains place types ranging from *administrative areas* and *wetlands*, to *hills* and *reefs* (Hill et al., 2000). As argued before, however, relying on individual labels alone is dangerous and often misleading. One can try to expand the labels by employing one or multiple external resources. For example, Hess et al. (2006) designed an algorithm, called G-Match, which employs WordNet (Miller, 1995) as an external lexical resource to enhance string similarity matching between geographic terms. Note, however, that such an approach still relies on labels alone and their canonical interpretations.

Semantic interoperability between gazetteers is closely related to topics from knowledge representation and reasoning and more specifically to ontology engineering, ontology alignment, and

[2] *Features* in machine learning are defined as measurable properties of the phenomenon under consideration. We will use the term *statistical feature* to distinguish them from *geographic features*, e.g., places.

semantic similarity measurement. One methodology that has been frequently utilized in the ontology alignment literature is a combination of string similarity and structural similarity (Shvaiko and Euzenat, 2013). Besides considering the (potentially expanded sets of) labels, such approach considers ontologies as graphs and assumes that semantically-similar terms will share similar structures (e.g., with respect to the number of subtypes). Examples for ontology alignment tools that combine these two types of similarity measures include SAMBO (Lambrix and Tan, 2006), RiMOM (Li et al., 2009), and Falcon-AO (Jian et al., 2005). To incorporate expert knowledge in the alignment process, researchers also proposed semi-automatic approaches which, in addition to automatic matching algorithms, allow experts to manually examine and align terms between ontologies. Their methods then consolidate the alignment results from both the experts and the algorithms, and resolve the potential conflicts. This semi-automatic approach has been used in COMA++ (Aumueller et al., 2005) and AgreementMaker (Cruz et al., 2007). A popular alignment API has been developed by Euzenat (2004), while a Linked Data-centric, bootstrapping matcher called BLOOMS has been implemented by Jain et al. (2010).

The studies discussed above mostly focus on types without considering the instances that belong to these types. Brauner et al. (2007) proposed an instance-based approach for gazetteer integration. Their method is based on place instances that are confirmed to be the same in different gazetteers. For example, if one can confirm that place instance $p$ in gazetteer $A$ is the same as the place instance $p'$ in gazetteer $B$, then the feature type $t$ of $p$ should be semantically similar to the type $t'$ of $p'$, even though $t$ and $t'$ may be labeled with different terms. Such approaches, however, do not scale, require manual interaction, are sensitive to various sampling effects and subsumption relations between the considered features.

Our approach can be differentiated from the existing work. First, instead of looking at the textual labels or definitions of place types, we examine the instances belonging to each place type and extract a wide variety of spatial statistical features from them. Second, our approach does not require an agreement on the same place instances across gazetteers. In sum, we propose a statistical and bottom-up driven approach to quantifying the spatial patterns representative of geographic feature type. The derived spatial signatures can help understand similarities and differences underlying those types that cannot be revealed by (lightweight) ontologies alone. The signatures can also be applied to support the alignments of geo-ontologies.

# 3 Gazetteer Datasets

In this work, we focus on three major Linked Data gazetteers, namely DBpedia Places, GeoNames, and Getty Thesaurus of Geographic Names. In the following, we briefly introduce each of them.

**DBpedia Places.** DBpedia is the Semantic Web version of Wikipedia (Lehmann et al., 2015). It was generated by semantically annotating the data extracted and mined from Wikipedia articles. As a result, DBpedia inherits many of the key strengths and also weaknesses of Wikipedia and its geographic data, e.g., a rich amount of user-contributed content. DBpedia Places is a subset of DBpedia focusing specifically on geographic places. It contains feature types such as *AdministrativeRegion*, *Restaurant*, *Stream*, *WineRegion*, and so forth. In total, the DBpedia Places dataset contains $92$ feature types and more than $924,000$ place instances. It is worth noting that DBpedia also contains various other feature types whose instances may have a spatial footprint without being categorized as places themselves. Finally, it also contains places on celestial bodies other than the Earth.

**GeoNames.** GeoNames is a gazetteer that contains over $10,000,000$ places throughout the world. Due to its high coverage, GeoNames has often been used to support geographic information retrieval and enrichment (Passant, 2007; Hu et al., 2015; Pasley et al., 2008). Each place in GeoNames is categorized into one of $9$ major feature types and then further subdivided into one of $645$ minor feature types. The resulting flat hierarchy of so-called feature codes has been often criticized for its arbitrarity and unintuitive choices. Similarly to DBpedia Places, GeoNames contains user-contributed data and therefore has the potential biases and pitfalls that arise from user-generated content.

**TGN.** Getty Thesaurus of Geographic Names is a structured vocabulary that contains approximately $1,106,000$ named places. These places include political entities, such as *counties* and *cities*, as well as physical features, such as *mountains* and *caves*. TGN also contains both current and historical places. Constructed based on national and international standards, TGN follows the terminologies that are warranted for use by authoritative literary sources. TGN focuses primarily on places that are culturally or historically significant and therefore typical Point Of Interest (POI) types (e.g., *restaurants*) are largely missing.

While all of the three gazetteers contain data outside the boundaries of the United States, the data coverage varies from country to country. Therefore, this study focuses on the contiguous United States which has been well covered by these gazetteers. In addition, we removed feature types that contain fewer than $2$ instances, since they cannot be used to produce meaningful statistical results. Table 1 summarizes the number of feature types in each gazetteer used in this study.

Table 1: The number of geographic feature types used in this study.

| | DBpedia Places | GeoNames | TGN |
|---|---|---|---|
| Num of feature types | 73 | 198 | 285 |

# 4 Methods

We investigate three kinds of spatial statistics to understand the similarities and differences in geographic feature type: spatial point patterns, spatial autocorrelations, and spatial interactions with other geographic features. We selected several representative statistics for each kind. We have extracted statistical features for all geographic feature types ($556$ in total) and all gazetteers. For illustration and comparison, we will use the feature types *Dam* and *Stream* as running examples throughout the paper, and will introduce further types such as *Island*, *County*, *Administrative Region*, and *Mountain*, to highlight specific aspects detailing how our signatures help us to make interesting observations about the similarities and differences between the gazetteers. The spatial distribution of *Dam* and *Stream* instances is depicted in Fig. 2

## 4.1 Spatial Point Patterns

Most gazetteers rely exclusively on point coordinates (typically centroids) to represent the spatial footprints of places. In a few cases, however, polygon and polyline representations are available.

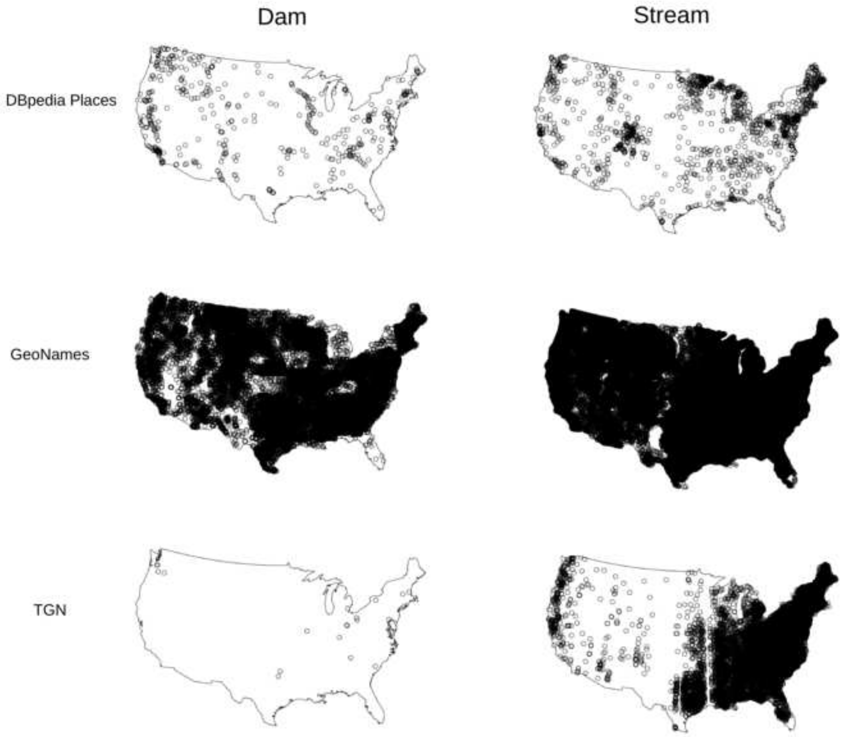


Figure 2: Spatial distributions of *Dam* and *Stream* in contiguous United States.

Thus, we employ spatial point pattern analysis to understand the spatial structure of place instances (of given types). More specifically, we group our analysis into local and global point patterns.

### 4.1.1 Local Point Patterns

Both intensity analysis (i.e. local intensity and kernel density estimation of the sample area) and distance-based analysis (i.e. nearest neighbor analysis, Ripley's K, and standard deviational ellipse) have been performed to examine the local point patterns. Among these analysis, local intensity of the point patterns and nearest neighbor analysis are quantitative measures, but Ripley's K, kernel density estimation, and standard deviational ellipse are curves, maps, or geometries respectively. Therefore, extracting quantitative characteristics from those visually exploratory plots are discussed in this section as well.

Since many geographic feature types have a high number of instances (e.g., GeoNames contains $218,701$ *churches* and DBpedia Places contains $55,969$ *settlements*), calculating some statistics for overall point patterns requires substantial computing resource. For example, calculating Ripley's K for the feature type *Populated Place* in GeoNames requires more than $104$ GB memory, which is beyond the computing capability of a typical workstation. Therefore, we employ a sampling strategy to explore local point patterns in addition to the overall statistics (see Fig. 3).

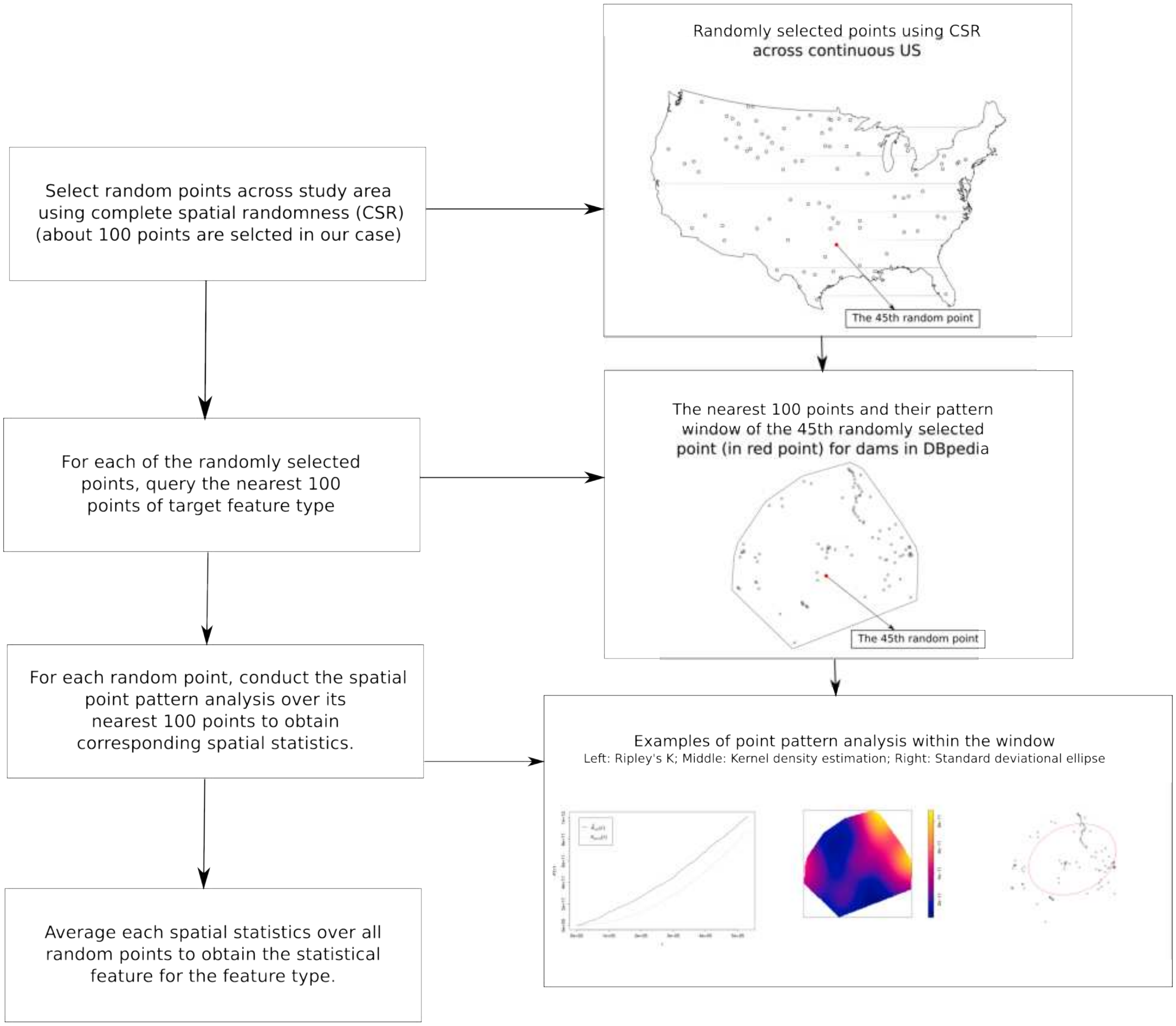

Select random points across study area using complete spatial randomness (CSR) (about 100 points are selcted in our case)
For each of the randomly selected points, query the nearest 100 points of target feature type
For each random point, conduct the spatial point pattern analysis over its nearest 100 points to obtain corresponding spatial statistics.
Average each spatial statistics over all random points to obtain the statistical feature for the feature type.
Randomly selected points using CSR across continuous US
The 45th random point
The nearest 100 points and their pattern window of the 45th randomly selected point (in red point) for dams in DBpedia
The 45th random point
Examples of point pattern analysis within the window
Left: Ripley's K; Middle: Kernel density estimation; Right: Standard deviational ellipse

Figure 3: Sampling strategy for exploring local point patterns. Left: the work flow of sampling strategy; Right: corresponding examples

As shown in Fig. 3, we first generated about 100 random sampling points within the boundary of the contiguous U.S. using a Complete Spatial Randomness (CSR) process. For each random sampling point, we selected its 100 nearest neighbors, and performed statistical analysis on these 100 nearest neighbors. The derived statistics will be assigned to this random point, and this process will iterate for each of the 100 random points (i.e., the sampling areas) in the contiguous U.S. Finally, we averaged the derived statistics from the 100 random points, and used the averaged value to characterize the spatial structure of the place type under consideration, e.g., *Stream*. In the following text, we will describe each of the point pattern statistics. Note that each feature type will use the exactly same set of randomly selected points. Since the 100 nearest points are selected for each of the 100 random point and these random points are independent of geographic locations, we assume that at the end most instances of any feature type are taken into account for calculating the statistics. By using this proposed sampling technique, some instances may be considered more than once, but since the average statistic is finally taken on all randomly selected groups, this effect can be neglected.

**Intensity of the point patterns & distance to nearest neighbor analysis.** Local intensity and the distance to nearest neighbor are selected to reflect the spatial point distribution patterns. Since both are quantitative measurements, they are directly used as statistical features in this work. For distance to nearest neighbor analysis, both mean and variance are extracted. Examples are showed in Table 2

Table 2: Local intensity and distance to nearest neighbor (in meters).

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Local intensity | $4.5 \times 10^{-11}$ | $9.4 \times 10^{-9}$ | $5.0 \times 10^{-13}$ | $2.4 \times 10^{-10}$ | $2.6 \times 10^{-8}$ | $5.2 \times 10^{-9}$ |
| Mean distance to nearest neighbor | $5.8 \times 10^{4}$ | $5.8 \times 10^{3}$ | $1.2 \times 10^{5}$ | $3.5 \times 10^{4}$ | $3.2 \times 10^{3}$ | $1.7 \times 10^{4}$ |
| Variance distance to nearest neighbor | $4.0 \times 10^{9}$ | $5.1 \times 10^{7}$ | $9.0 \times 10^{10}$ | $1.5 \times 10^{9}$ | $1.4 \times 10^{7}$ | $1.1 \times 10^{9}$ |

**Kernel density estimation.** In addition to local intensity, kernel density estimations (KDE) has also been used to analyze point patterns. Figure 4 shows the KDE maps for *Dam* and *Stream* based on one common random sampling point and its 100 nearest neighbors. To understand the kernel density map quantitatively, two characteristics have been extracted to measure the spatial structures of the spatial point patterns. Below are the two quantitative measures:

- *Bandwidth (in meters) of the kernel density map.* For different spatial point patterns, the selections of bandwidth for creating kernel density maps are different. In our work, Berman and Diggle (1989)'s algorithm is used to calculate the optimized bandwidths for various spatial point patterns. This algorithm selects the optimized bandwidth that minimizes the mean-square error defined by Diggle (1985). Based on this algorithm, the bandwidth is likely to be large if the intensity of points varies dramatically in space, and be small otherwise. Subsequently, these bandwidths are selected as characteristics for representing kernel density maps.

- *Range of the kernel density estimation.* The range of KDE is calculated as the difference between the minimum and the maximum density values on the kernel density map. Since each sample area contains a fixed number of points, the range of KDE will be large if the points are clustered. In contrast, if these points are dispersed, the range will be small. Therefore, the range of KDE can also reflect the structure of spatial points. Table 3 shows two examples of these two statistics.

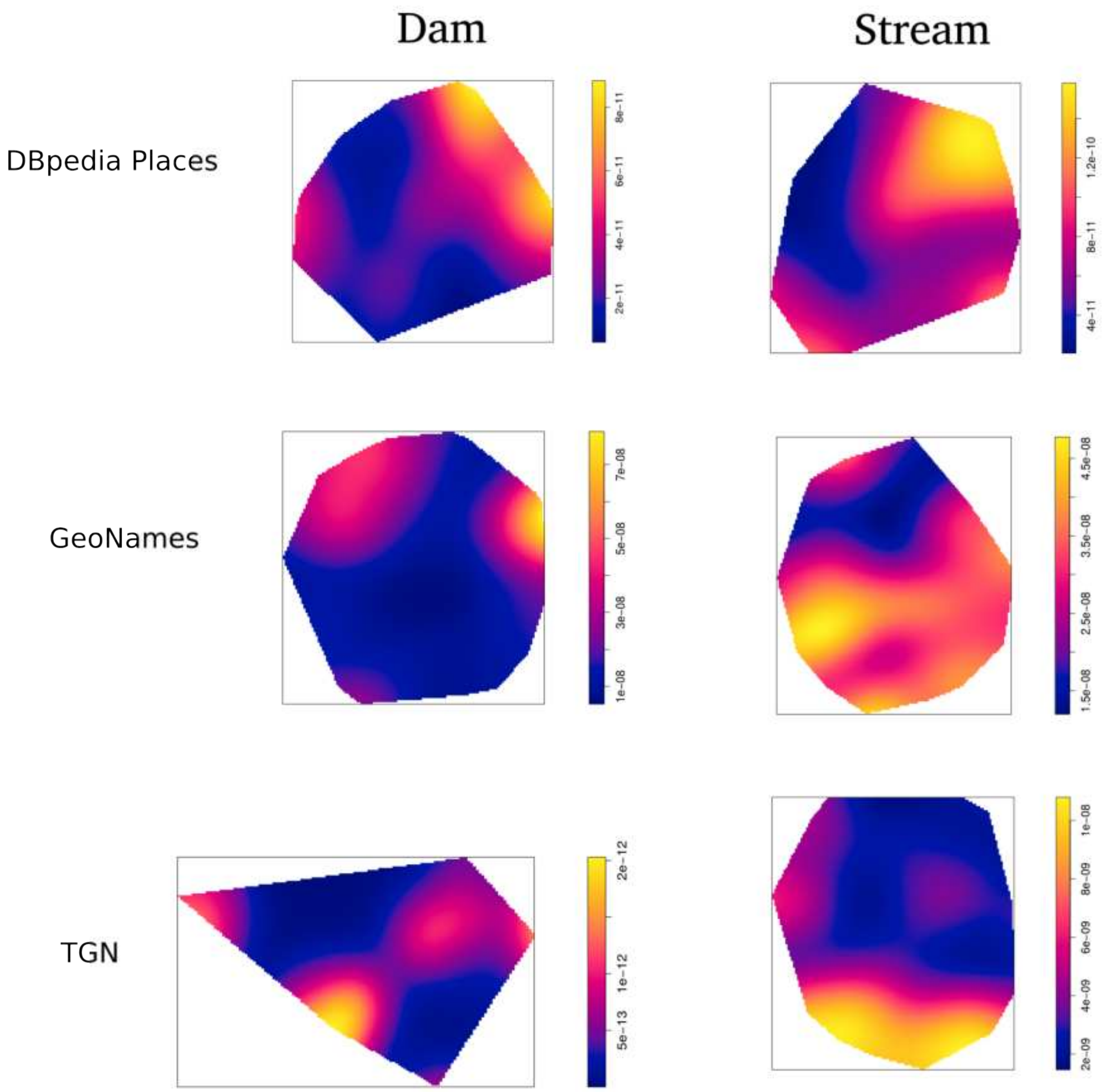


Figure 4: Kernel density estimation for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

Table 3: Bandwidth (in meters) and range of the kernel density estimation.

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Local bandwidth | $4.3 \times 10^4$ | $4.6 \times 10^3$ | $7.2 \times 10^3$ | $2.7 \times 10^4$ | $3.6 \times 10^3$ | $1.0 \times 1.0^4$ |
| Local range | $1.2 \times 10^{-10}$ | $2.3 \times 10^{-8}$ | $2.0 \times 10^{-12}$ | $1.0 \times 10^{-9}$ | $4.9 \times 10^{-8}$ | $1.2 \times 10^{-8}$ |

**Ripley's K.** Ripley's K characterizes spatial point patterns at multiple distance scales. This property distinguishes Ripley's K from many other spatial point pattern measures, such as the distance to nearest neighbor, which are based on a single distance scale. Accordingly, Ripley's K provides a different perspective on quantifying spatial point patterns. However, since multiple distances are used, the computation is intensive, especially for feature types that have many instances. Examples of Ripley's K are depicted in Fig. 5. Two quantitative characteristics are extracted from the Ripley's K graph, which are the range of K and the mean deviation from the theoretical values. Examples are illustrated in Table 4.

- *Range of Ripley's K.* Similar to kernel density maps, the range of K is indicative of the spatial spread of points. When points are dispersed in a large scale, the range will be larger. In our work, the range is calculated from the minimum of Ripley's K rule of thumb, which is one quarter of the smallest side of the bounding rectangle of the studied point pattern.
- *Mean deviation from the theoretical values*. The theoretical K measures are calculated from random spatial points generated by a CSR process. Comparing the K measures from the observed points with the theoretical values can help reveal the pattern of the spatial points. For example, if the observed K measure is larger than the theoretical one, then the place instances of a feature type are more clustered; in contrast, if the observed K measure is smaller, then the points are more dispersed. In this paper, we use the mean of the deviation as the statistical feature.

Table 4: Range and mean deviation from the theoretical Ripley's K.

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Range | $4.4 \times 10^5$ | $4.1 \times 10^4$ | $3.7 \times 10^6$ | $2.3 \times 10^5$ | $2.3 \times 10^4$ | $1.7 \times 10^5$ |
| Mean deviation from the theoretical K measures | $1.2 \times 10^{11}$ | $1.6 \times 10^9$ | $1.7 \times 10^{13}$ | $5.8 \times 10^{10}$ | $3.0 \times 10^8$ | $4.5 \times 10^{10}$ |

**Standard deviational ellipse.** Standard deviational ellipse is used to check the directionality and shape of the spatial point distribution. Two spatial point patterns that are both clustered can

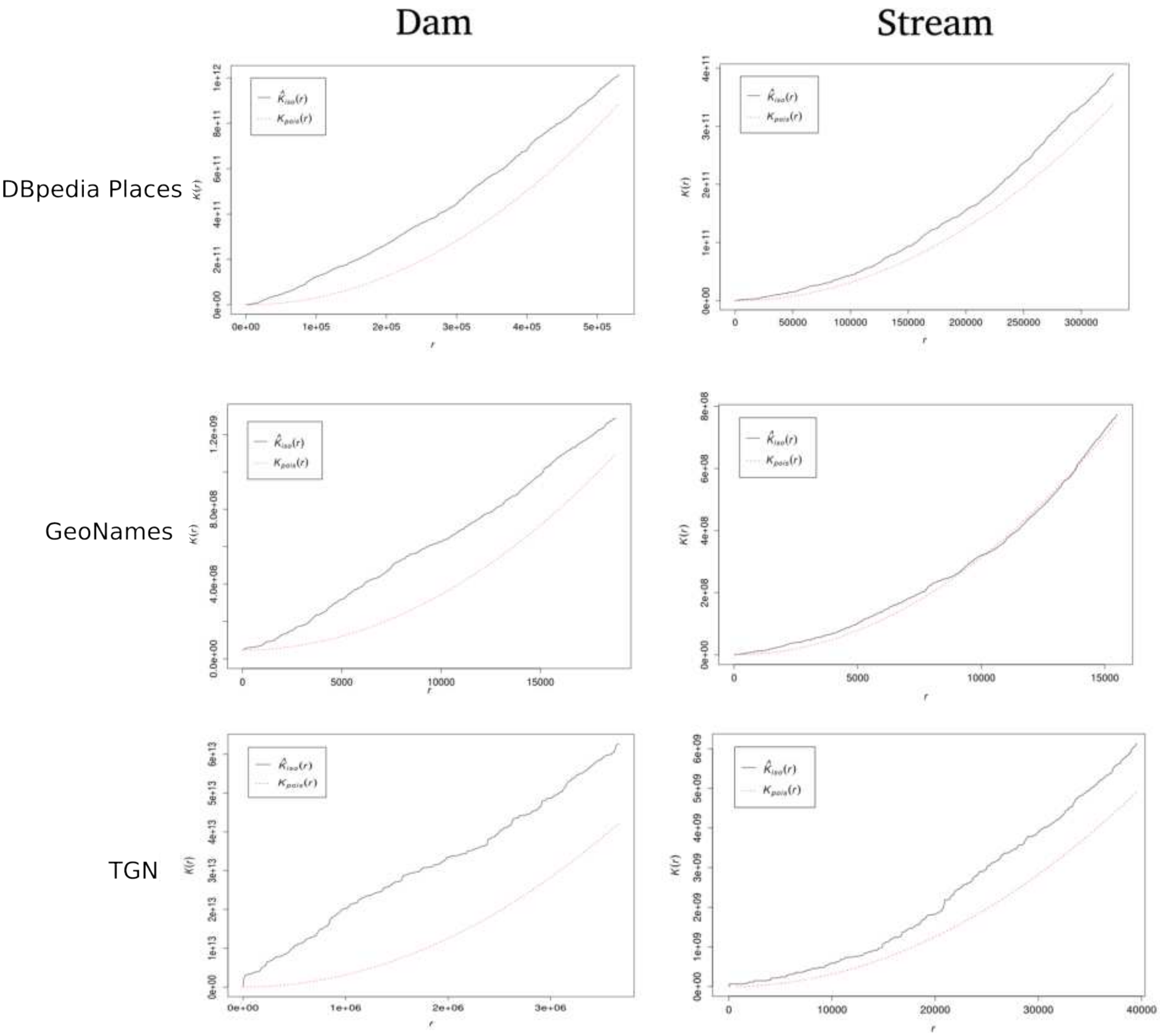


Figure 5: Ripley's K for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

have different directionalities: one is more clustered along the x-axis, while the other may be clustered along the y-axis. Similarly, two clusters can demonstrate distinct geometric shapes as well. The standard deviational ellipse is employed to characterize the directionality and the shape of the spatial points. Figure 6 shows examples of standard deviational ellipse for two types.

We used the rotation of the ellipse, standard deviation along x-axis, as well as standard deviation along y-axis, to capture the shape and directionality of the spatial points. Table 5 provides examples on these statistics.

Table 5: Rotation, standard deviation along x- and y-axis of the standard deviational ellipse.

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Rotation | $-0.23$ | $0.02$ | $0.12$ | $-0.17$ | $0.34$ | $-0.04$ |
| Standard deviation along x-axis | $3.8 \times 10^5$ | $3.6 \times 10^4$ | $3.0 \times 10^6$ | $1.9 \times 10^5$ | $2.1 \times 10^4$ | $1.5 \times 10^5$ |
| Standard deviation along y-axis | $4.5 \times 10^5$ | $5.0 \times 10^4$ | $6.0 \times 10^6$ | $2.6 \times 10^5$ | $2.8 \times 10^4$ | $1.9 \times 10^5$ |

#### 4.1.2 Global Point Patterns

In addition to local patterns, we perform global spatial analysis as well. Specifically, we analyze the overall intensity and the global KDE based on the place instances belonging to a feature type. Figure 7 illustrates the global KDE for two feature types in different gazetteers. Similar to local point patterns, bandwidth and the range of global KDE are extracted from the density map, which are showed in Table 6.

Table 6: Global intensity, bandwidth and range of global KDE.

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Global intensity | $5.4 \times 10^{-11}$ | $7.2 \times 10^{-9}$ | $2.5 \times 10^{-12}$ | $1.8 \times 10^{-10}$ | $2.73 \times 10^{-8}$ | $3.7 \times 10^{-9}$ |
| Global bandwidth | $2.8 \times 10^4$ | $3.5 \times 10^3$ | $1.0 \times 10^5$ | $1.5 \times 10^4$ | $2.6 \times 10^3$ | $4.9 \times 10^3$ |
| Global range | $1.5 \times 10^{-10}$ | $1.6 \times 10^{-8}$ | $1.3 \times 10^{-11}$ | $0.1 \times 10^{-9}$ | $6.5 \times 10^{-8}$ | $2.0 \times 10^{-8}$ |

## 4.2 Spatial Autocorrelations

Spatial autocorrelations capture the interactions between geographic places. Unlike spatial data that have both geographical coordinates and attributes (e.g., precipitations or temperatures), digital gazetteers typically provide only coordinates with place names and types. To understand the spatial autocorrelation patterns of place instances of a feature type, we transfered the point data of each type into a raster map. In order to make the extracted statistics comparable, the resolution of the raster map are set to be same for all feature types, which is about $36.0$ kilometers × $22.2$ kilometers. The value of each cell in the raster map represents the number of instances falling into that cell.

With this data conversion, Moran's I and semivariograms can then be computed to quantify the spatial autocorrelations. Examples of the raster for *Dam* and *Stream* are illustrated in Fig. 8.

**Global Moran's I.** Global Moran's I is calculated on each feature type based on the converted raster maps to check how the intensities of cells differ from their neighbors. In this study, we use Queen's case to define neighborhoods, and therefore all $8$ cells surrounding the target cell are considered as neighbors. Values that are close to $1$ indicate a strong positive autocorrelation, while negative values approaching $-1$ show strong negative autocorrelations. Finally, values around $0$ indicate randomness in the locations of the observed place instances of a given type. Examples are listed in Table 7.

Table 7: Global Moran's I for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Global Moran's I | 0.21 | 0.65 | 0.12 | 0.35 | 0.83 | 0.83 |

**Semivariogram.** We employ semivariograms to explore the intensity of points over a larger spatial scale. To make semivariograms comparable, a fixed distance range, roughly $1130$ miles, and a fixed number of distance lags ($51$) were applied to all types. In addition, due to the complexity of the required computation, $100$ random points were selected for calculating the experimental semivariogram. Using semivariograms, we can check the dissimilarity of cell intensities across multiple distance lags. Observations of *Dam* and *Stream* from the three studied gazetteers are shown in Fig. 9.

Semivariances at the first, the median, as well as the last distance lag, are subsequently extracted as statistical features for understanding the geographic feature types.

## 4.3 Spatial Interactions with Other Geographic Features

The third category of statistics extracted to form spatial signatures examines the interactions between the target place type and other geographic features. These other geographic features could be summarized into two groups: (1) the internal group: the ones that are observed from the same gazetteer but with different feature types, and (2) the external group: the ones that are obtained outside of the gazetteer. This section will discuss these two groups respectively.

**Group 1: Internal.** The interactions of place types from the same gazetteer are studied in this group. Here, two statistics are quantitatively proposed to reflect the spatial interactions of each feature type with its neighbors: the count of distinct nearest feature types and the entropy of nearest feature types.

- *Count of distinct nearest feature types:* instances of one feature type may have different types of neighbors compared with other instances of a different feature type. Take restaurants and mountains as examples, the nearest neighbors of restaurants may belong to a relatively more diverse type set including bars, cinemas, hotels, and so forth, while the set of neighbor types for mountains is more restricted. Therefore, the count of distinct nearest feature types plays a role for determining the semantics of one feature type. In this work, we use the $CountNearest_i = \frac{m_i}{N}$ to quantitatively measure the normalized count of distinct nearest feature types of the $i^{th}$ feature type, where $m_i$ is the number of distinct nearest feature type for feature type $i$ and $N$ is the total number of feature types in one gazetteer.

- *Entropy of nearest feature types:* in addition to the count of nearest feature types, we also use information entropy as an indicator for the diversity of the nearest feature types. The formula of calculating the entropy of $i^{th}$ feature type is given as: $EntropyNearest_i = -\sum_{j=1}^{N} \frac{n_j}{N} \log(\frac{n_j}{N})$ where $N$ is the total number of feature types in one gazetteer and $n_j$ is the number of nearest instances from the $j^{th}$ feature type. Larger entropy values indicate that this target feature type most likely locates in a neighborhood whose feature types vary greatly. Table 8 shows the example of the two proposed signatures.

Table 8: Count and entropy of nearest feature types.

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Count | 0.41 | 0.45 | 0.04 | 0.54 | 0.07 | 0.43 |
| Entropy | 2.86 | 2.86 | 0.37 | 3.90 | 0.45 | 3.69 |

**Group 2: External.** A feature type also interacts with other geographic features that are out of scope of the studied gazetteers. While many kinds of external data could be used, we employ population distribution data and street network data, since these two kinds of data can interact with most feature types in a typical gazetteer, are not part of the gazetteers themselves, and are available from up-to-date, high-resolution, and authoritative datasets. Other examples could include temperature and precipitation data.

- *Population distribution.* The rationale of using population distribution data is that different geographic feature types may have significantly different demographic characteristics. For example, *cities* and *hotels* will most likely only occur in high average population cells, while *streams* may or may not be near human settlements. Finally, *mountains* are unlikely to occur in high population cells (as long as the resolution of the dataset is sufficiently high). We utilize LandScan which is a global-scale population dataset with a spatial resolution of 1 kilometer (Bhaduri et al., 2002, 2007). The newest LandScan data (for 2014), have been retrieved from `http://wms.cartographic.com/LandScan2014/`. Figure 10 shows a map visualization of the used LandScan dataset.

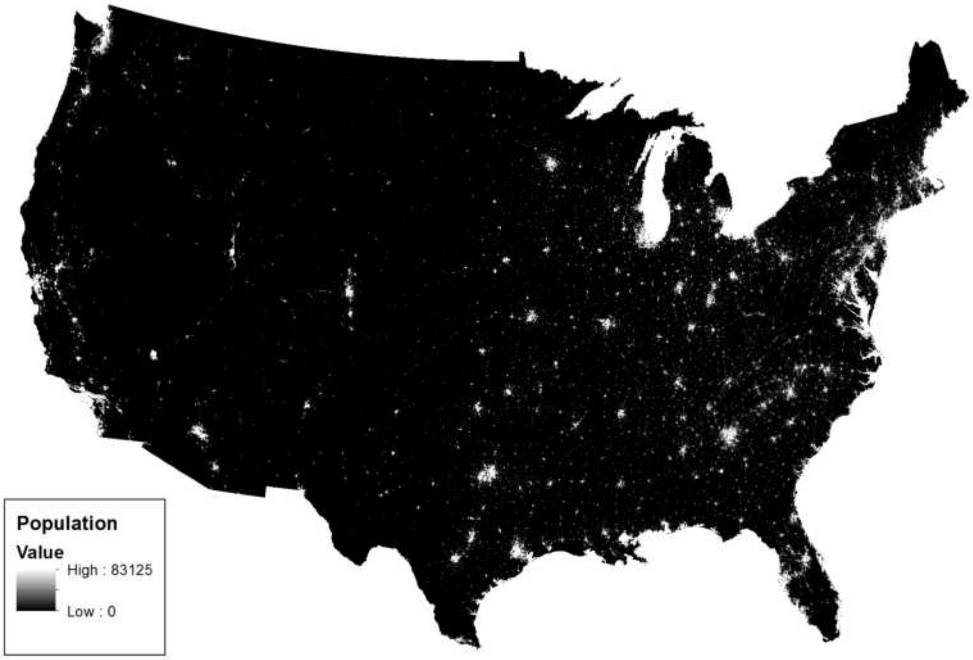


Figure 10: The 2014 LandScan dataset of the US mainland

Based on the LandScan population dataset, we calculated the minimum, maximum, mean, and standard deviation. Table 9 illustrates the statistical features derived from the population distribution data.

Table 9: Spatial signatures based on LandScan 2014 population data (unit: number of persons).

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Min | 0 | 0 | 0 | 0 | 0 | 0 |
| Max | 11806 | 11271 | 405 | 26102 | 25399 | 10901 |
| Mean | 105.107 | 58.860 | 72.083 | 175.744 | 35.484 | 44.411 |
| Std dev | 617.088 | 239.100 | 128.256 | 891.529 | 194.246 | 184.360 |

- *Road network.* Road networks are another type of geographic features used to study the interactions between the target feature type and external datasets. The rationale is similar to including population distribution data: different feature types may show significant differences in terms of their distances to the nearest road segments. We use the road network data from the Digital Chart of the World (DCW) project which provides comprehensive digital map data of the world (Danko, 1992). The data have been retrieved from `http://www.diva-gis.org/gdata`, and are visualized in Fig. 11.

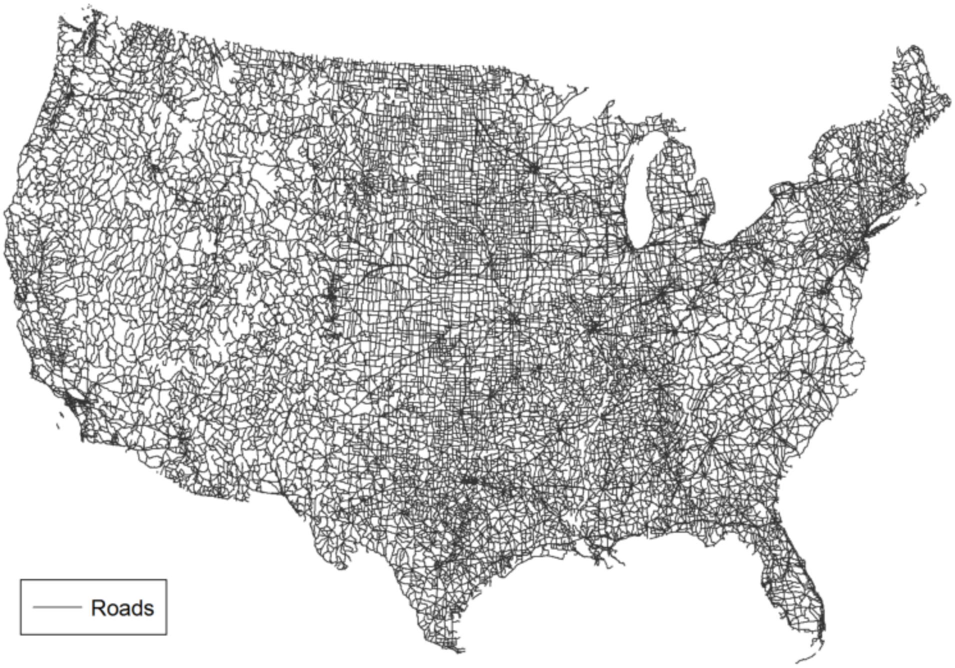


Figure 11: The DCW road network dataset of the US mainland.

For each feature type in the three gazetteers, we also calculate the minimum, maximum, mean, and standard deviations of the shortest distances to transportation infrastructure based on their place instances. Table 10 uses the examples of *Dam* and *Stream* to illustrate these statistical features.

Table 10: Spatial signatures based on DCW road network data (unit: meters).

| | Dam | | | Stream | | |
|---|---|---|---|---|---|---|
| Statistics | DBpedia Places | GeoNames | TGN | DBpedia Places | GeoNames | TGN |
| Min | 3.69 | 0.13 | 164.56 | 1.01 | 0.02 | 0.01 |
| Max | 22386.02 | 35342.17 | 15252.04 | 27079.98 | 77201.22 | 36127.38 |
| Mean | 3013.24 | 3504.90 | 3007.38 | 3653.39 | 4525.16 | 4919.15 |
| Std dev | 3158.05 | 3224.77 | 3728.41 | 4039.65 | 4373.17 | 4173.89 |

## 4.4 Summary of the Statistical Features

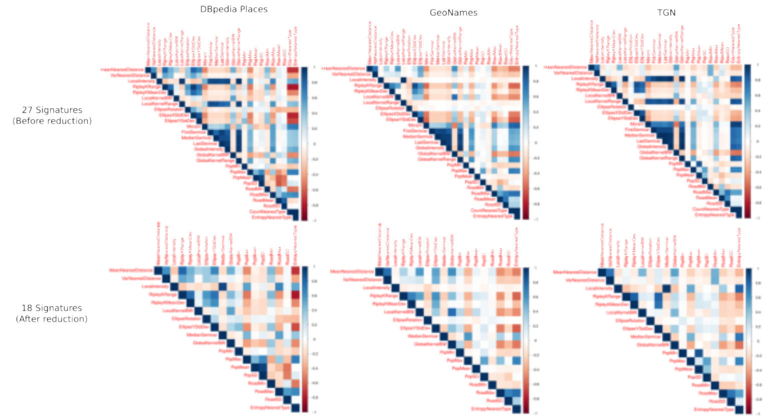
The three categories of spatial analysis reflect spatial point patterns, autocorrelations, and spatial interactions of geographic feature types. In total, we have generated 27 statistical features to characterize the different aspects of geographic feature types in the gazetteers. All these statistical features are calculated in R using packages such as *spatstat*, *gstat* and *sp*. Table 11 shows a summary of these statistics.

Table 11: A summary of the 27 different kinds of statistical features.

| Spatial Point Patterns | | Spatial Autocorrelations | Spatial Interaction with Other Geographic Features | |
|---|---|---|---|---|
| Local | Intensity | Global Moran' I | Internal | Count of distinct nearest feature types |
| | Mean distance to nearest neighbor | | | |
| | Variance distance to nearest neighbor | | | Entropy of nearest feature types |
| | Kernel density (bandwidth) | | | |
| | Kernel density (range) | Semivariogram value (at first distance lag) | External | Population value (min) |
| | Ripley's K (range) | | | |
| | Ripley's K (mean deviation) | | | Population value (max) |
| | Standard deviational ellipse (rotation) | Semivariogram value (at median distance lag) | | Population value (mean) |
| | Standard deviational ellipse (std dev along x-axis) | | | Population value (std dev) |
| | Standard deviational ellipse (std dev along y-axis) | | | Shortest distance to road (min) |
| Global | Intensity | Semivariogram value (at last distance lag) | | Shortest distance to road (max) |
| | Kernel density (bandwidth) | | | Shortest distance to road (mean) |
| | Kernel density (range) | | | Shortest distance to road (std dev) |

# 5 Experiments and Discussion

Having extracted these 27 potential statistical features, the next step is to analyze their correlations. If two statistical features have a significantly high correlation, we only keep one of them to reduce the dimensionality of our feature space. Figure 12 depicts the resulting correlation matrix. From the first row of the matrix, we can clearly identify several features that are highly correlated (i.e., dark blue and dark red cells). For instance, one can see a correlations among the semivarigram values at the first, median and last distance lags, as well as the global intensity value. Therefore, only the semivariogram value at median distance lag, among the four, is kept in the reduced version. Overall, we removed 9 highly correlated statistical features. The final list of these statistical features is shown in Table 12. The second row of Figure 12, shows that there is no commonly significant correlation among all three gazetteers.



*. MeanNearestDistance: Mean distance to nearest neighbor
*. VarNearestDistance: Variance distance to nearest neighbor
*. LocalIntensity: Local intensity
*. RipleyKRange: Ripley's K (range)
*. RipleyKMeanDev: Ripley's K (mean deviation)
*. LocalKernelBW: Local kernel density (bandwidth)
*. LocalKernelRange: Local kernel density (range)
*. EllipseRotation: Standard deviational ellipse (rotation)
*. EllipseXStdDev: Standard deviational ellipse (std dev along x-axis)
*. EllipseYStdDev: Standard deviational ellipse (std dev along x-axis)
*. MoranI: Global Moran's I
*. FirstSemivar: Semivarigram value (at first distance lag)
*. MedianSemivar: Semivariogram value (at median distance lag)
*. LastSemivar: Semivariogram value (at last distance lag)
*. GlobalIntensity: Global Intensity
*. GlobalKernelBW: Global kernel density (bandwidth)
*. GlobalKernelRange: Global kernel density (range)
*. PopMin: Population value (min)
*. PopMax: Population value (max)
*. PopMean: Population value (mean)
*. PopSD: Population value (std dev)
*. RoadMin: Shortest distance to road (min)
*. RoadMax: Shortest distance to road (max)
*. RoadMean: Shortest distance to road (mean)
*. RoadSD: Shortest distance to road (std dev)
*. CountNearestType: Count of distinct nearest feature types
*. EntropyNearestType: Entropy of nearest feature types

Figure 12: Visualization of the correlation matrix. First row: correlation matrix with all statistical features (27); Second row: correlation matrix with reduced statistical features (18)

Table 12: List of spatial signatures after reduction (18 in total)

| Spatial Signatures | |
|---|---|
| Mean distance to nearest neighbor | Global kernel density (bandwidth) |
| Variance distance to nearest neighbor | Population value (min) |
| Local intensity | Population value (max) |
| Ripley's K (range) | Population value (mean) |
| Ripley's K (mean deviation) | Population value (std dev) |
| Local kernel density (bandwidth) | Short distance to road (min) |
| Standard deviational ellipse (rotation) | Short distance to road (max) |
| Standard deviational ellipse (std dev along y-axis | Short distance to road (std dev) |
| Semivariogram (median distance lag) | Entropy of nearest feature type |

After reducing the highly correlated statistical features, the remaining ones are used to derive spatial signatures of geographic feature types for the three gazetteers (i.e. DBpedia Places, GeoNames and TGN). We applied multidimensional scaling, more specifically metric-MDS, to visualize similarities and differences among place types. From Fig. 13, it is clear that place types for the three tested gazetteers have a significant overlap in general. However, how a specific place type in one gazetteer relates to other place types in another gazetteer is not immediately clear. Therefore, we selected three cases to explain the potential of using spatial statistics for understanding differences and similarities in geographic features type specification that are not captured by today's lightweight ontologies. Before diving into details, it is important to remember that we propose the use of spatial signatures in addition to existing alignment methods that use string similarity, structural measures, and so forth. Consequently, considering the signatures alone results in a lower discriminatory power. We discuss them here in isolation nonetheless to ensure that our results are not influenced by a specific alignment technique but merely by the extracted statistical features.

Table 13: Number of instances for selected geographic feature types.

| Number of instances | DBpedia Place | GeoNames | TGN |
|---|---|---|---|
| Island | 102 | 9187 | 2895 |
| Mountain | 2253 | 64316 | 123 |
| AdministrativeRegion/ADM2/County | 3289 | 3098 | 1142 |
| Hotel | 184 | 52692 | NA |

## 5.1 Same Name and Similar Spatial Patterns

The type *Island* is present in the three gazetteers with exactly the same names. Thus, these features would be matched with a high likelihood if structural comparison would not reveal dramatic differences between them. These three types are highlighted in Fig. 13, in which their high dimensional spatial signatures are mapped to 2-dimension space. *Island* signatures in TGN and GeoNames turn out to be very close, and they are actually each other's nearest neighbor in the 2D representation. This observation indicates a high similarity in the (otherwise hidden) conceptualizations of the type *Island* in GeoNames and TGN. However, the distance from DBpedia Places to TGN and DBpedia Places to GeoNames are substantially large. This observation may be surprising at first but can be well explained by revisiting what we know about the datasets. As mentioned in Section 3, DBpedia Places contains places that have been mentioned in Wikipedia articles. Thus only significant islands will be included in the *Island* type of DBpedia Places; see Table 13 for the counts of selected feature types in the US. In contrast, both GeoNames and TGN have a larger coverage. One consequence of this, for instance, is that the average nearest neighbor distance for islands in Dbpedia Places is larger than $50$ kilometers (and there are only $102$ islands) while it is $9$ kilometers and $12$ kilometers for GeoNames and TGN, respectively. The core argument underlying our work is that this substantial difference in the *extension* of the type *Island* is not merely an issue of overall coverage but caused by a different *intension*. Thus, islands in GeoNames and TGN are more alike in terms of their size, accessibility, administration, and so forth. This is not the case for all feature types, e.g., all three gazetteers have similar signatures for administrative divisions; see below.

## 5.2 Same Name but Different Spatial Patterns

Even though two types have exactly the same name, their semantics may vary greatly depending on the source and focus of the gazetteers. This can be demonstrated using the *Mountain* feature type. As depicted in Fig. 14 the MDS distances among DBpedia Places, GeoNames, and TGN are relatively large. Therefore, and with respect to the used statistical features, we can assume that the conceptualization of this class differs substantially among the three datasets. We discussed possible reasons for such differences in the Introduction section. Among other factors, these differences can be caused by a unclear distinction between mountains and hills, mountains and mountain peaks, the fact that mountains have different definitions per country and these differences are often included in Wikipedia and thus DBpedia but not in Geonames, and so forth. A similar observation can be made for the *Monument* type in DBpedia and GeoNames. The type as such is only vaguely defined and contains natural and man made features, ancient and modern features, ranges from small to large scale structures, and so on.

## 5.3 Different Names but Similar Spatial Patterns

The spatial signatures can also be applied to detect similarities among geographic feature types that have very different names/labels. The geographic feature type *County* has different names in the three gazetteers (i.e., it is called *AdministrativeRegion* in DBpedia Places, *ADM2* or *second-order administrative division* in GeoNames, and *County* in TGN). Therefore if only names were used for comparison, those features would hardly be considered as potential matches. However, since counties in United States are administrative regions, their boundaries are officially defined independently from the specific focus area of a gazetteer or its ontology. Thus a county's spatial structures is expected to be similar across gazetteers. Figure 15 justifies such argument by showing a relatively small distance among the three gazetteers and thereby confirms that these types can be aligned despite the different type labels. It is this information, that we see as the key contribution of the signatures (and spatial statistics more broadly) to modern alignment tools.

## 5.4 Different Names and Different Spatial Patterns

Last but not least, one could argue that we have not shown that the spatial signatures have sufficient discriminatory power to differentiate types within a gazetteer that are indeed dissimilar. In other words, types such as *Island* and *Hotel* that should be far apart in statistical feature space could still have ended up close to each other. To demonstrate that this is not the case, consider the example provided in Fig.16. Both DBpedia Places types are far apart indicating that they do not share similar values with respect to their studied statistical features. This is important as we do not want to accidentally match geographic feature types that are neither similar with respect to their names nor their underlying semantics.

Finally, and to show the limitations of our current work, this leads to the question whether all types that are assumed to be dissimilar can be confirmed using the spatial signatures alone. This is not the case and future work will require more feature engineering to find more measurable characteristics that successfully tell apart as many types as possible. To give a concrete example, place types that frequently co-occur such as *Restaurant* and *Bar* cannot be clearly distinguished. In fact, we have defined temporal and thematic signatures that help us differentiate them based on when they are visited and how people communicate about them McKenzie et al. (2015). Put differently, some semantic signatures need to be formed by spatial *and* temporal bands. Other cases include the various kinds of populated places such as *City*, *Town*, *Village*, and so forth.

# 6 Conclusions and Future Work

In this work, we proposed to derive and utilize *spatial semantic signatures* to understand the conceptualization of geographic feature types. Such bottom-up approaches become necessary as existing gazetteer ontologies and vocabularies are lightweight and underspecified to a degree where most of the underlying semantics is communicated by labels and simple (and often unbalanced) type hierarchies alone. This presents a major challenge for ontology alignment techniques and therefore hampers query federation and data integration within cyber-infrastrucutures and knowledge graphs such as the Web of Linked Data. It is believed that a combination of top-down ontology engineering and bottom-up data-driven analysis is required to successfully approach these challenges. For example, one can combine the similarity measures derived from both the top-down textual labels and bottom-up spatial signatures to achieve a higher accuracy in ontology alignment. In fact, we demonstrated this in previous article in which we enrich a top-down ontology using the bottom-up knowledge mined from Linked Data (Hu and Janowicz, 2016).

Semantic signatures are an analogy to spectral signatures in remote sensing where a certain combination of *bands* (wavelengths) uniquely identifies a type, e.g., a land cover class. In previous work, we have used temporal and thematic bands to form signatures for Points of Interest in urban environments. Here, we introduce a variety of *spatial* bands, i.e., statistical features extracted from spatial statistics and aggregated to the level of geographic feature types, to arrive at spatial signatures that enable us to better understand differences and similarities between the typing schemta and ontologies of three major gazetteers.

We have introduced 27 statistical features collected from three groups, those extracted from spatial point patterns, those extracted from measures of spatial autocorrelation, and those extracted from the spatial interaction across feature types. Nine statistical features are later subtracted from the 27 due to their high correlations with others. Next, we have discussed four experiments to show how the resulting spatial signatures can reveal information about the conceptualization of geographic feature types used by the three gazetteers. More specifically, we have discussed (1) cases where the labels and signatures match, i.e., cases in which a common name is indeed indicative of a similar conceptualization, (2) cases where labels match despite substantial differences in the signatures, (3) cases where the labels do not indicate a clear similarity despite the feature types being closely related, and finally (4) cases where both names and signatures differ. Our work is intended to complement modern alignment techniques that mostly rely on string matching and structural measures.

Future work will focus on deriving additional statistical features, especially those that can deal with co-occurring place types as well as types whose instances stand in characteristic topological relations to each other such as malls containing grocery stores and restaurants. In addition, we will combine the spatial signatures presented in this work with our previously described temporal and thematic signatures to increase their discriminative power. Finally, while we focused on the bottom-up part in this work, we will integrate the results with classical top-down knowledge engineering using our ODOE methodology (Janowicz, 2012). Thereby we aim to introduce spatial statistics into the field of geo-ontology engineering.

# References

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.

Aumueller, D., Do, H.-H., Massmann, S., Rahm, E., 2005. Schema and ontology matching with coma++. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, pp. 906–908.

Berman, M., Diggle, P., 1989. Estimating weighted integrals of the second-order intensity of a spatial point process. Journal of the Royal Statistical Society. Series B (Methodological) , 81–92.

Bhaduri, B., Bright, E., Coleman, P., Dobson, J., 2002. Landscan. Geoinformatics 5 (2), 34–37.

Bhaduri, B., Bright, E., Coleman, P., Urban, M. L., 2007. Landscan usa: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal 69 (1-2), 103–117.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts , 205–227.

Brauner, D. F., Casanova, M. A., Milidiú, R. L., 2007. Towards gazetteer integration through an instance-based thesauri mapping approach. In: Advances in Geoinformatics. Springer, pp. 235–245.

Cruz, I. F., Sunna, W., Makar, N., Bathala, S., 2007. A visual tool for ontology alignment to enable geospatial interoperability. Journal of Visual Languages & Computing 18 (3), 230–254.

Danko, D. M., 1992. The digital chart of the world project. Photogrammetric engineering and remote sensing 58 (8), 1125–1128.

Diggle, P., 1985. A kernel method for smoothing point process data. Applied statistics , 138–147.

Euzenat, J., 2004. An api for ontology alignment. In: The Semantic Web–ISWC 2004. Springer, pp. 698–712.

Euzenat, J., Ferrara, A., Meilicke, C., Nikolov, A., Pane, J., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Šváb-Zamazal, O., Svátek, V., et al., 2010. Results of the ontology alignment evaluation initiative 2010. In: Proceedings of the 5th International Conference on Ontology Matching-Volume 689. CEUR-WS. org, pp. 85–117.

Goodchild, M. F., Hill, L. L., 2008. Introduction to digital gazetteer research. International Journal of Geographical Information Science 22 (10), 1039–1044.

Hess, G. N., Iochpe, C., Castano, S., 2006. An algorithm and implementation for geoontologies integration. In: GeoInfo. pp. 109–120.

Hill, L. L., 2000. Core elements of digital gazetteers: placenames, categories, and footprints. In: Research and advanced technology for digital libraries. Springer, pp. 280–290.

Hill, L. L., Carver, L., Larsgaard, M., Dolin, R., Smith, T. R., Frew, J., Rae, M.-A., 2000. Alexandria digital library: user evaluation studies and system design. Journal of the American Society for Information Science 51 (3), 246–259.

Hu, Y., Janowicz, K., 2016. Enriching top-down geo-ontologies using bottom-up knowledge mined from linked data. In: Onsrud, H., Kuhn, W. (Eds.), Advancing Geographic Information Science: The Past and Next Twenty Years. GSDI Association Press, Ch. 13, pp. 183–198.

Hu, Y., Janowicz, K., Prasad, S., Gao, S., 2015. Metadata topic harmonization and semantic search for linked-data-driven geoportals: A case study using arcgis online. Transactions in GIS 19 (3), 398–416.

Jain, P., Hitzler, P., Sheth, A. P., Verma, K., Yeh, P. Z., 2010. Ontology alignment for linked open data. In: The Semantic Web–ISWC 2010. Springer, pp. 402–417.

Janowicz, K., 2012. Observation-driven geo-ontology engineering. Transactions in GIS 16 (3), 351–374.

Janowicz, K., Keßler, C., 2008. The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science 22 (10), 1129–1157.

Janowicz, K., Scheider, S., Pehle, T., Hart, G., 2012. Geospatial semantics and linked spatiotemporal data–past, present, and future. Semantic Web 3 (4), 321–332.

Jian, N., Hu, W., Cheng, G., Qu, Y., 2005. Falcon-AO: Aligning ontologies with falcon. In: Proceedings of K-CAP Workshop on Integrating Ontologies. pp. 85–91.

Keßler, C., Janowicz, K., Bishr, M., 2009. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems. ACM, pp. 91–100.

Lambrix, P., Tan, H., 2006. Samboa system for aligning and merging biomedical ontologies. Web Semantics: Science, Services and Agents on the World Wide Web 4 (3), 196–206.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al., 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web 6 (2), 167–195.

Li, J., Tang, J., Li, Y., Luo, Q., 2009. RiMOM: A dynamic multistrategy ontology alignment framework. Knowledge and Data Engineering, IEEE Transactions on 21 (8), 1218–1232.

McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A., Hu, Y., 2015. Poi pulse: A multi-granular, semantic signature-based information observatory for the interactive visualization of big geosocial data. Cartographica: The International Journal for Geographic Information and Geovisualization 50 (2), 71–85.

Miller, G. A., 1995. Wordnet: a lexical database for english. Communications of the ACM 38 (11), 39–41.

Pasley, R., Clough, P., Purves, R. S., Twaroch, F. A., 2008. Mapping geographic coverage of the web. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. ACM, p. 19.

Passant, A., 2007. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In: International Conference on Weblogs and Social Media.

Rice, M. T., Aburizaiza, A. O., Jacobson, R. D., Shore, B. M., Paez, F. I., 2012. Supporting accessibility for blind and vision-impaired people with a localized gazetteer and open source geotechnology. Transactions in GIS 16 (2), 177–190.

Schlieder, C., Vögele, T., Visser, U., 2001. Qualitative spatial representation for information retrieval by gazetteers. In: Spatial Information Theory. Springer, pp. 336–351.

Shvaiko, P., Euzenat, J., 2013. Ontology matching: state of the art and future challenges. Knowledge and Data Engineering, IEEE Transactions on 25 (1), 158–176.

Twaroch, F. A., Jones, C. B., Abdelmoty, A. I., 2008. Acquisition of a vernacular gazetteer from web sources. In: Proceedings of the first international workshop on Location and the web. ACM, pp. 61–64.

Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G., 2011. Design and use of the simple event model (sem). Web Semantics: Science, Services and Agents on the World Wide Web 9 (2), 128–136.

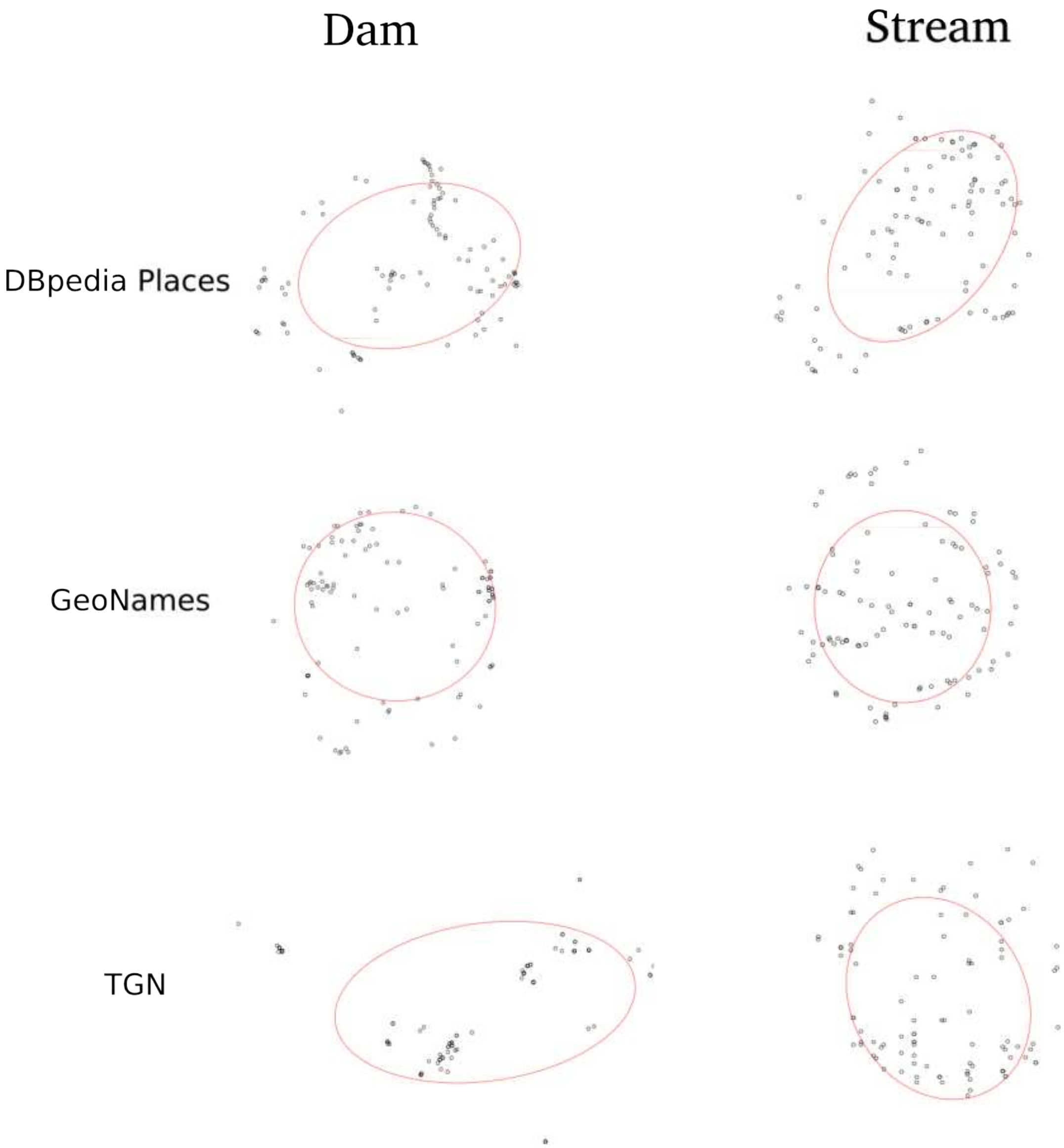


Figure 6: Standard deviational ellipse for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

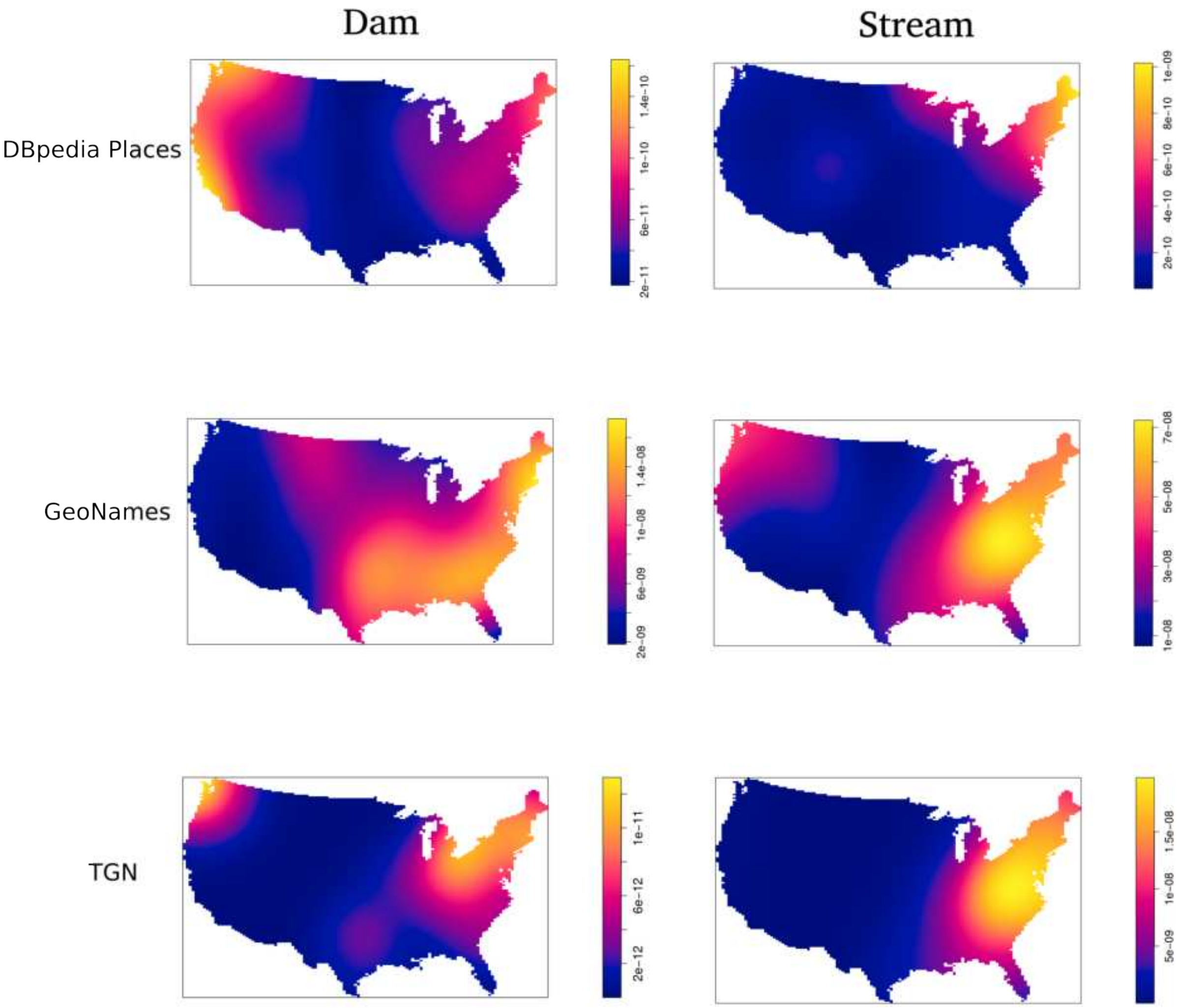


Figure 7: Global kernel density estimation for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

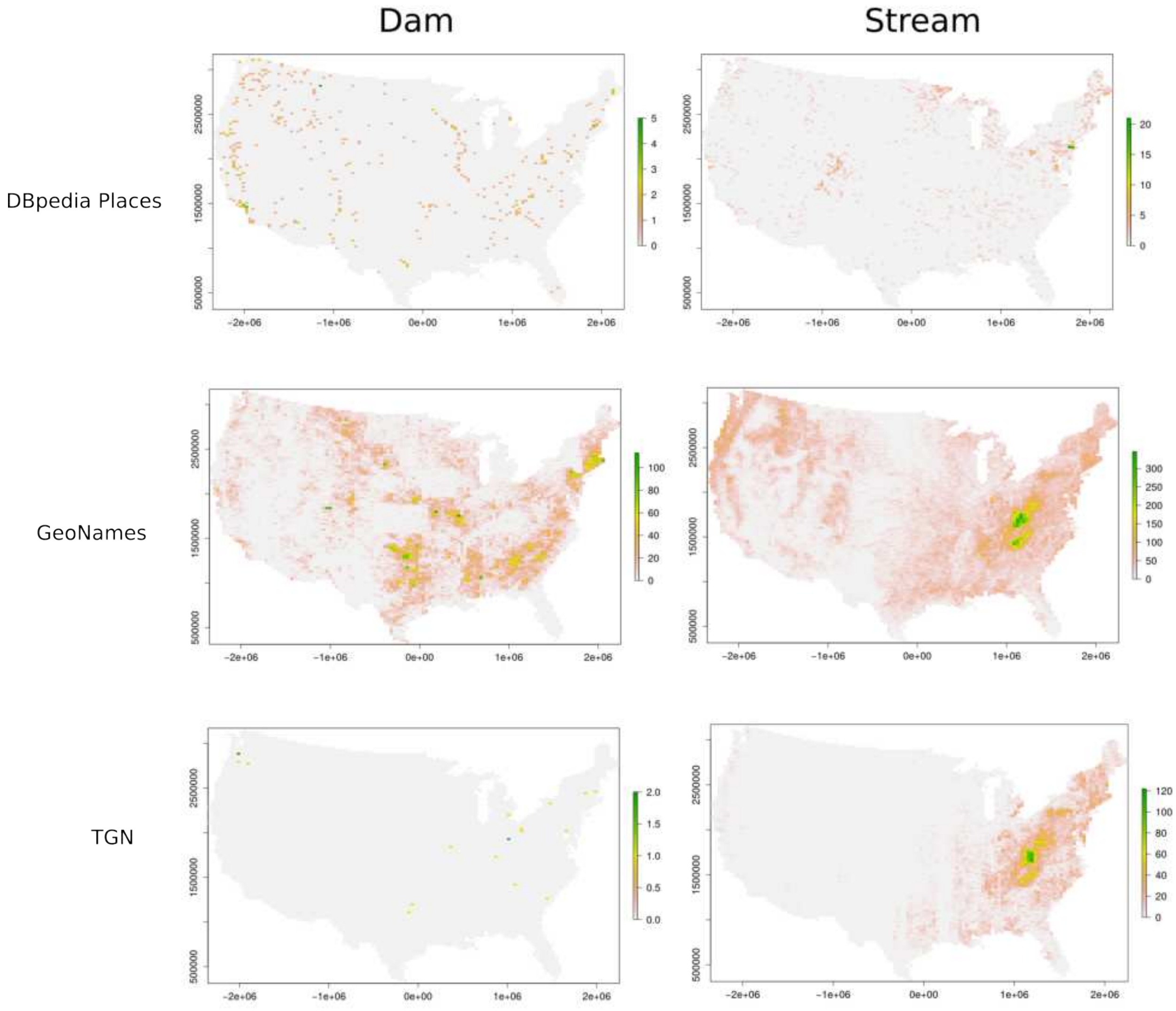


Figure 8: Raster maps for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

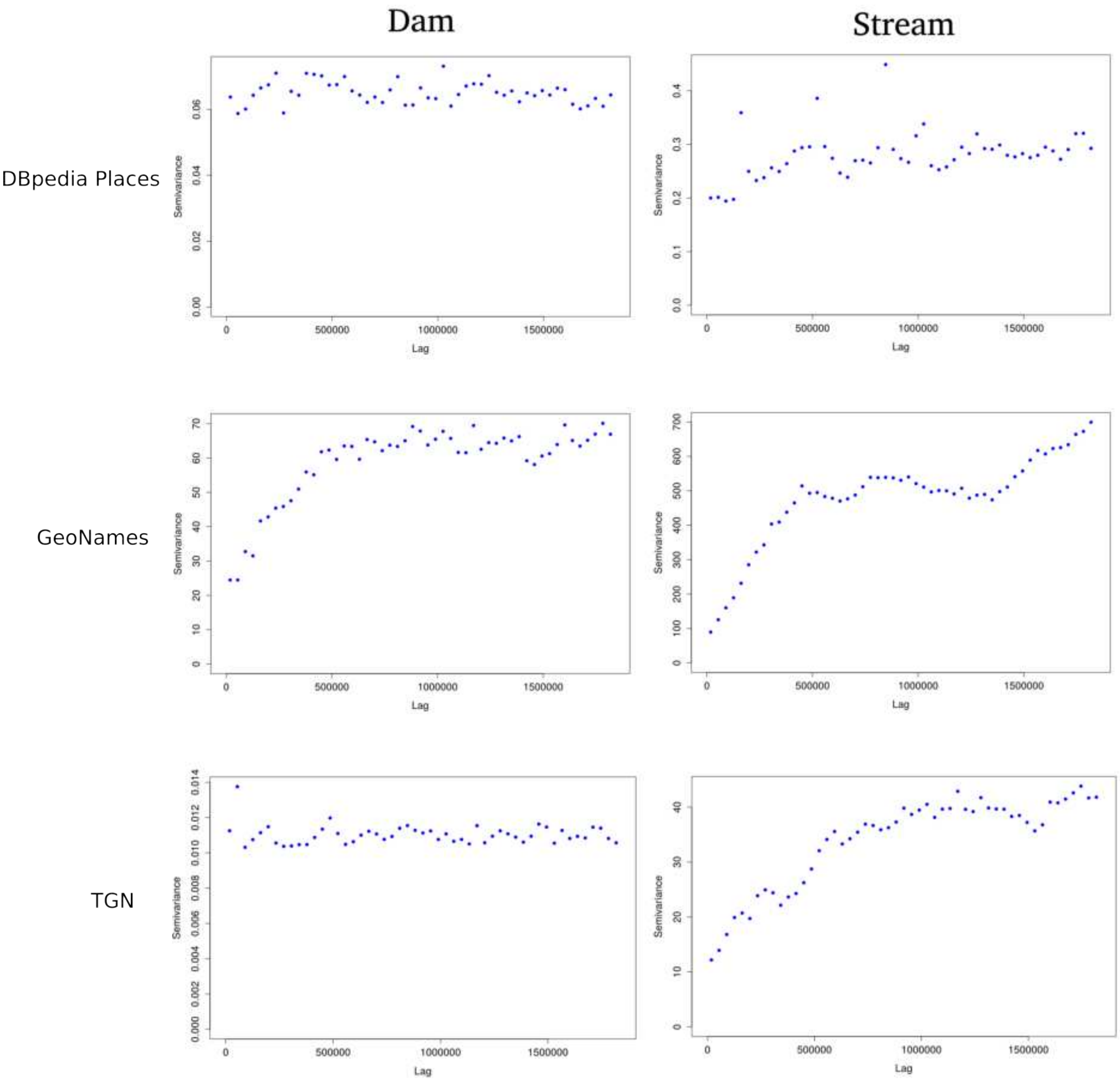


Figure 9: Experimental semivariograms for *Dam* and *Stream* in DBpedia Places, GeoNames and TGN.

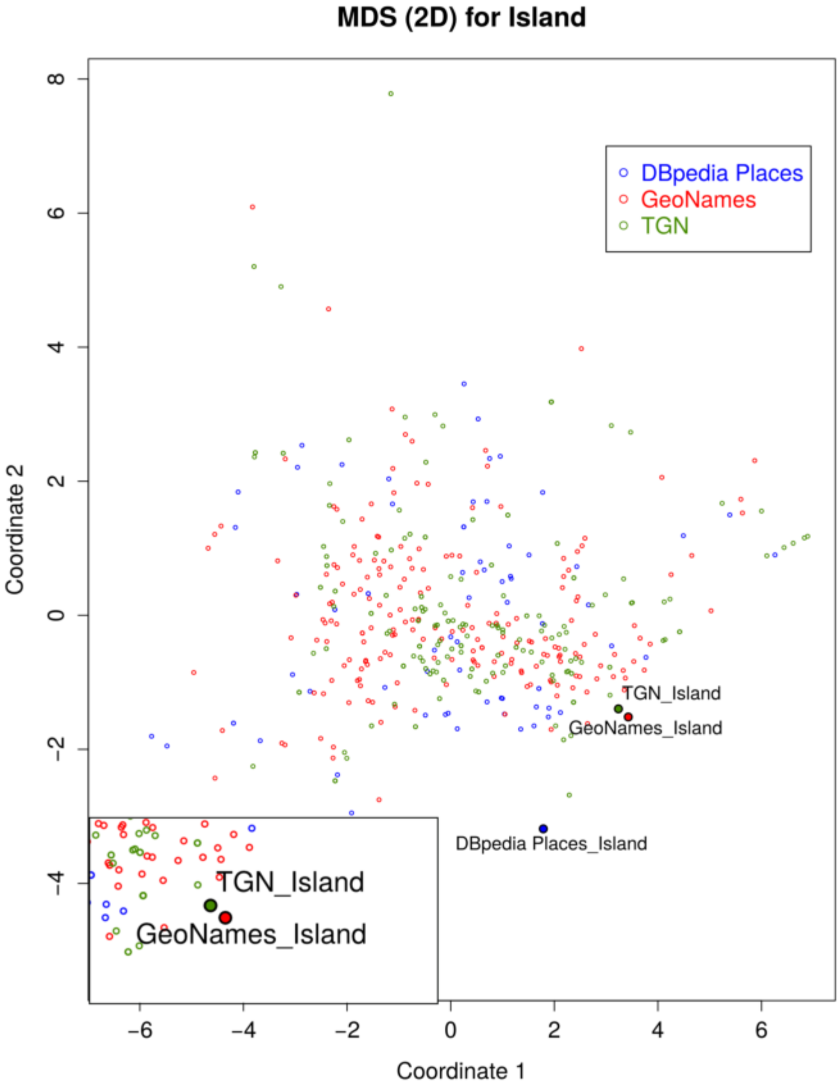


Figure 13: Multidimensional scaling for *Island* in DBpedia Places, GeoNames and TGN.

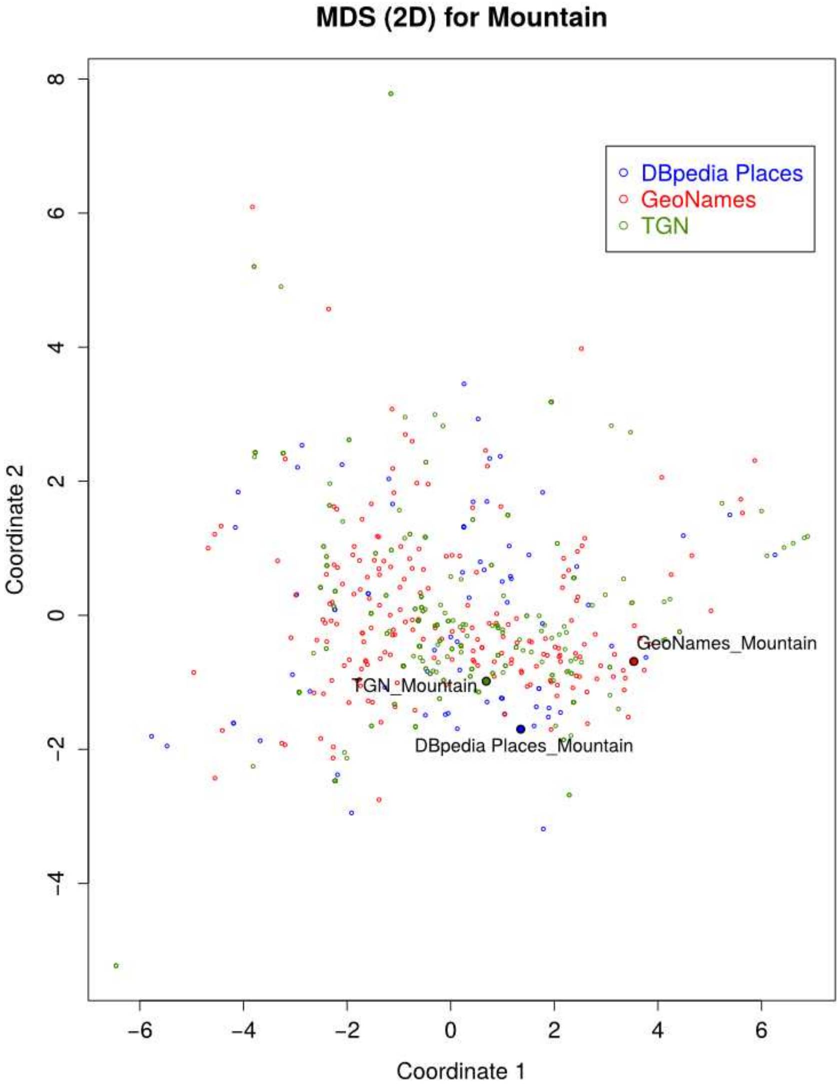


Figure 14: Multidimensional scaling for *Mountain* in DBpedia Places, GeoNames and TGN.

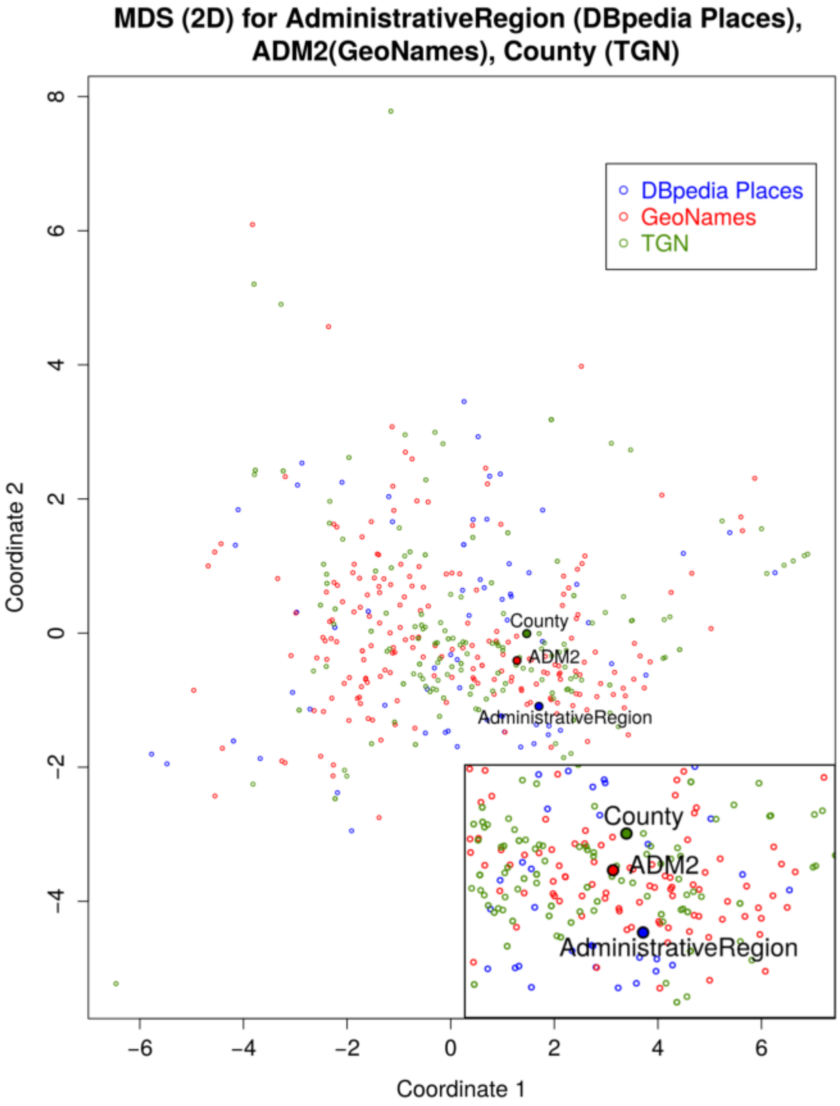


Figure 15: Multidimensional scaling for *County* in DBpedia Places, GeoNames and TGN.

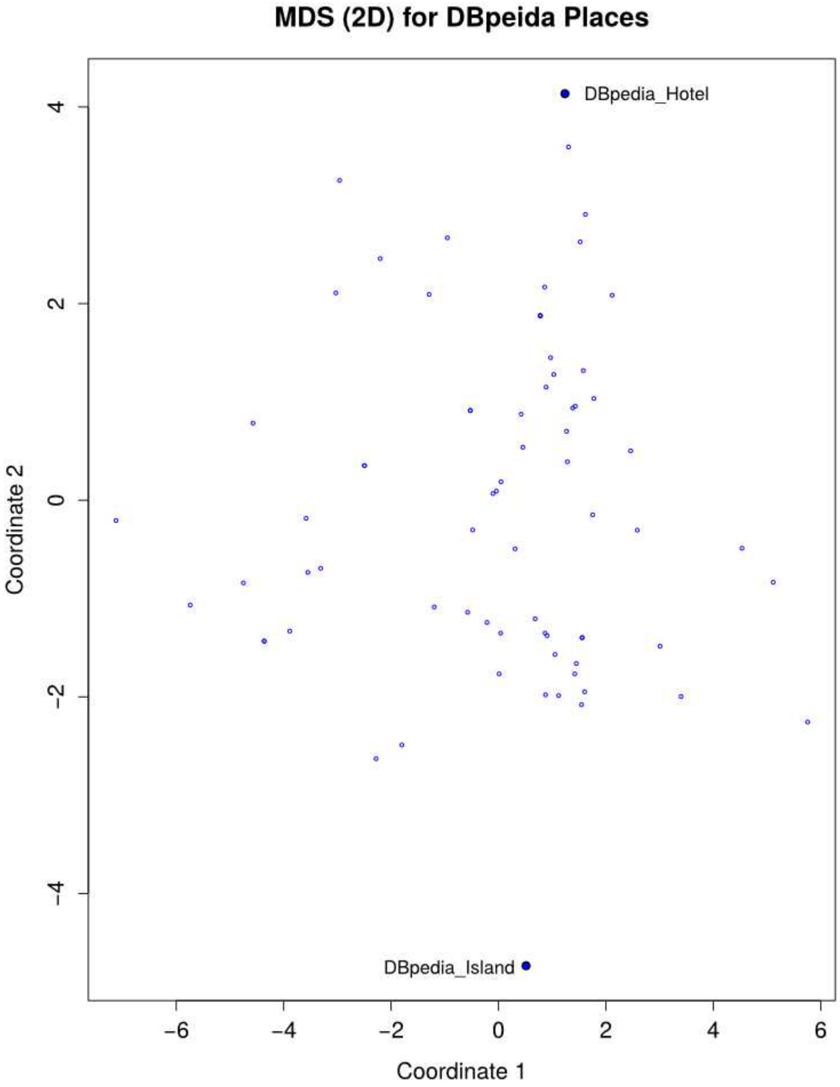


Figure 16: Multidimensional scaling for *Hotel* and *Island* in DBpedia Places