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

Rui Zhu, Yingjie Hu, Krzysztof Janowicz and Grant McKenzie

STKO Lab

Department of Geography, University of California, Santa Barbara

Outline

- Motivations
- Datasets
- Methods
- Case Studies
- Future Work

Motivations

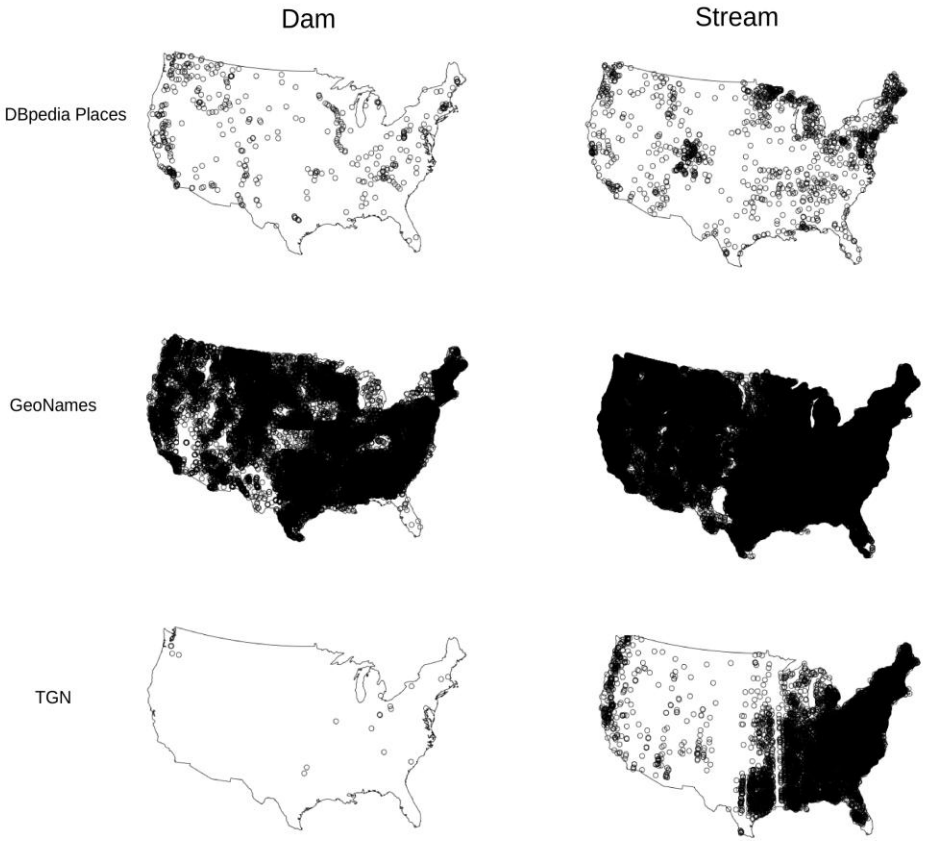
- A wide variety of gazetteers

Motivations (cont.)

- However, these gazetteers differ largely in terms of:
  - Overall coverage
  - Underlying data sources
  - Provided functionalities
  - Geographic feature type ontologies
  - …

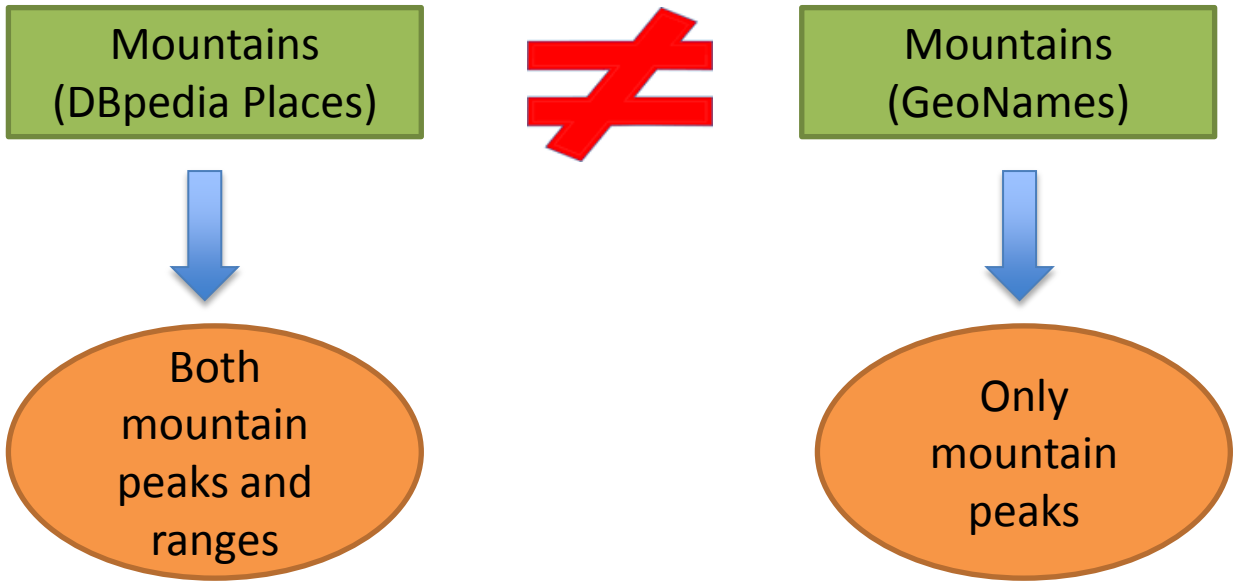
Motivations (cont.)

- Examples:



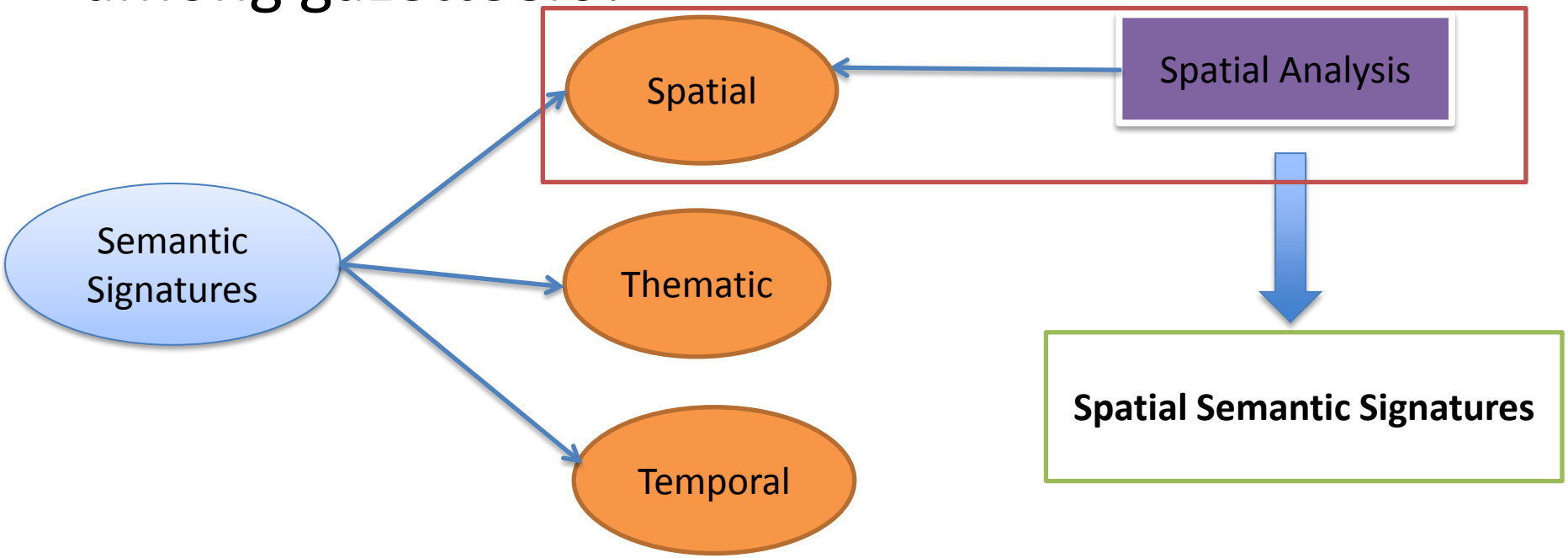
Motivations (cont.)

- Examples



Motivations (cont.)

- How to understand such semantic heterogeneity among gazetteers?



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

Datasets

DBpedia Places

Extracted from DBpedia articles

73 feature types in US

GeoNames

Formal geographical database that contains over eight million place names

198 feature types in US

Getty Thesaurus of Geographic Names (TGN)

Focus on places that are culturally or historically significant

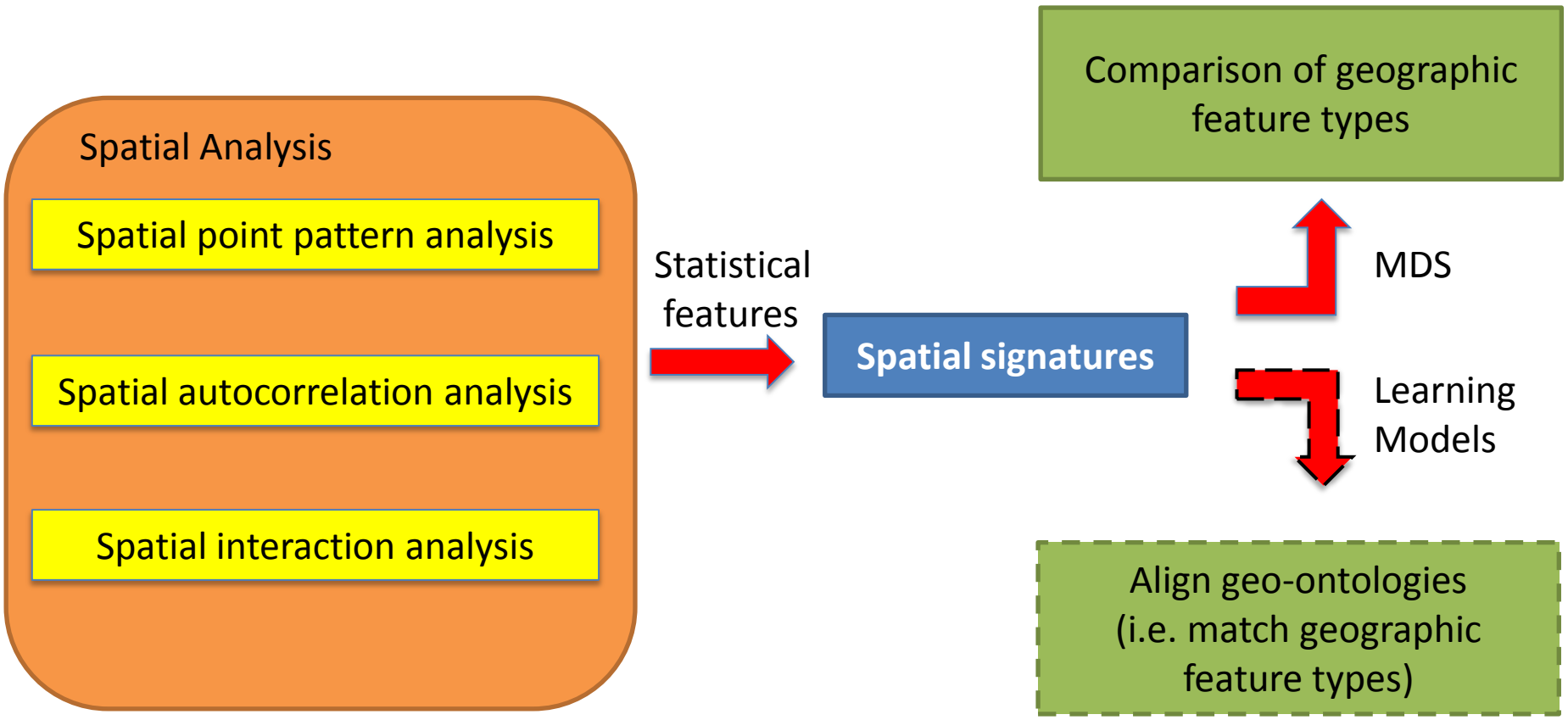
285 feature types in US

Datasets (cont.)

- Common information
  - Toponyms/Place names
  - Geographic feature type
  - Spatial footprints

Dams in GeoNames

Methods

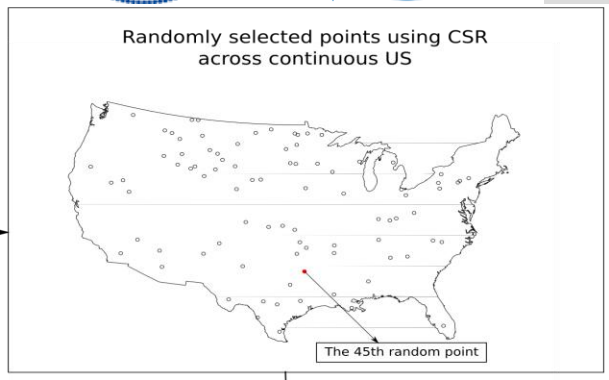

Spatial Analysis
Spatial point pattern analysis
Spatial autocorrelation analysis
Spatial interaction analysis
Statistical features
Spatial signatures
Comparison of geographic feature types
MDS
Learning Models
Align geo-ontologies (i.e. match geographic feature types)

Methods (cont.)

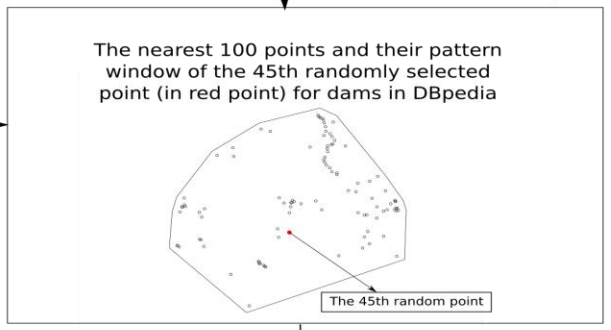
- Spatial point pattern analysis

A sampling technique:

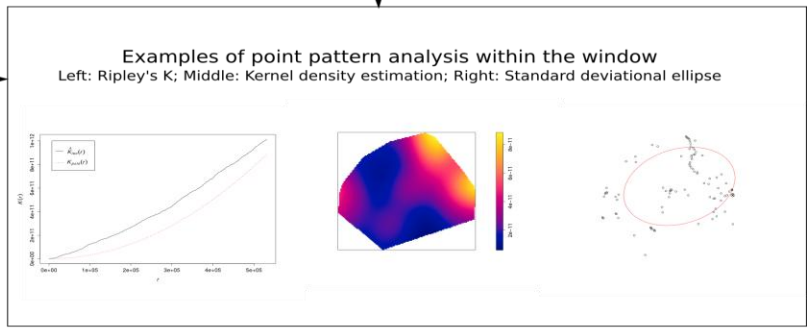
Select random points across study area using complete spatial randomness (CSR) (about 100 points are selcted in our case)



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.



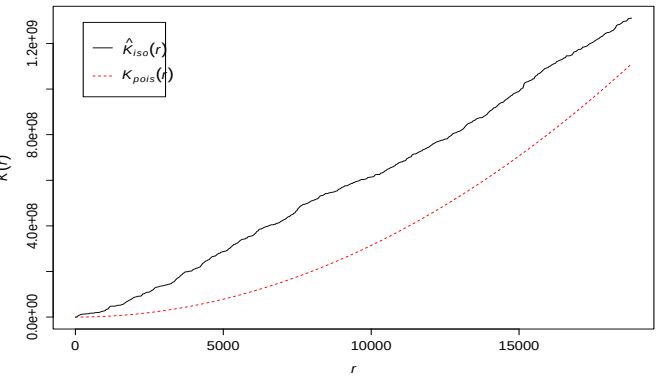
Average each spatial statistics over all random points to obtain the statistical feature for the feature type.

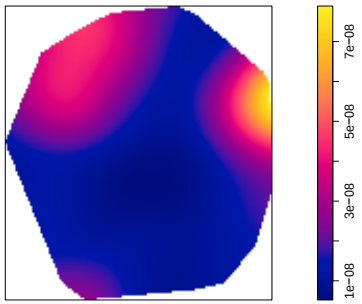
Methods (cont.)

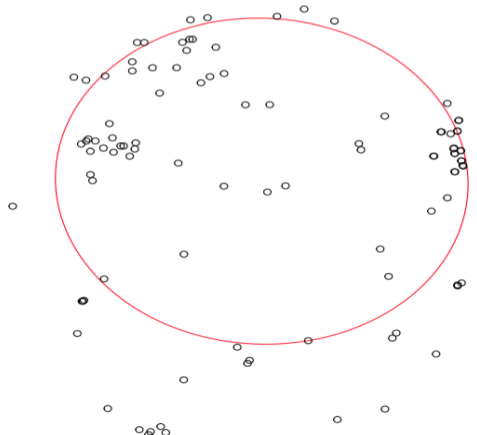
- Spatial point pattern analysis (cont.)

spatial statistics (local)

- **Intensity of point patterns**
- **Distance to nearest neighbor**
- Ripley's K (i.e. **range** and **mean deviation from the theoretical values**)
- Kernel density estimation (i.e. **bandwidth** and **range**)
- Standard deviational ellipse (i.e. **rotation**, **std. along x-axis** and **y-axis**)







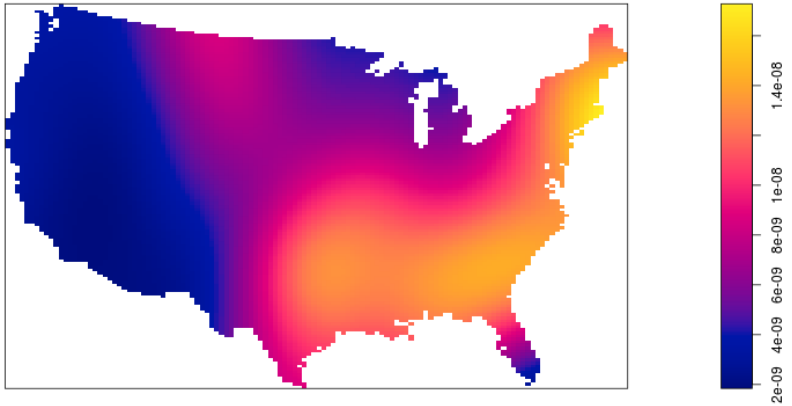
Methods (cont.)

- Spatial point pattern analysis (cont.)

spatial statistics (global)

- **Overall intensity of point patterns**
- Kernel density estimation (i.e. **bandwidth** and **range**)

Dams in GeoNames

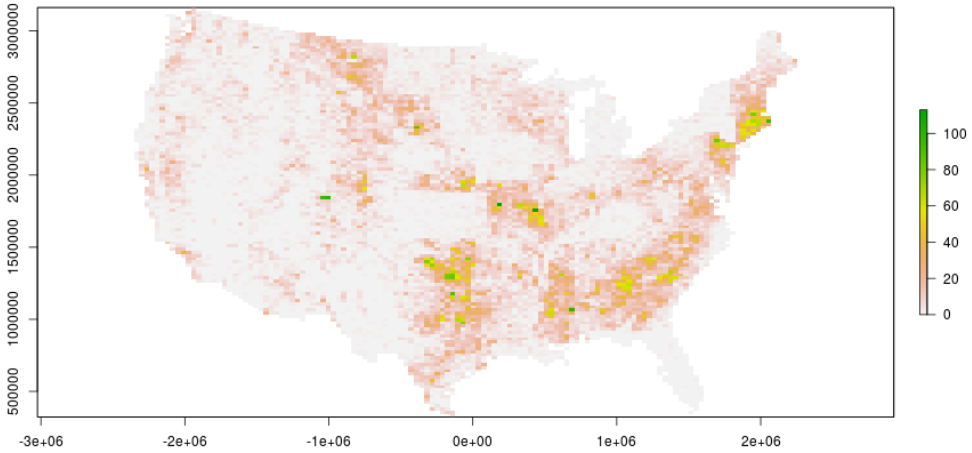


Methods (cont.)

- Spatial autocorrelation analysis

  Conversion: point data → raster map

Dams in GeoNames



Cell size: 36km * 22.2 km

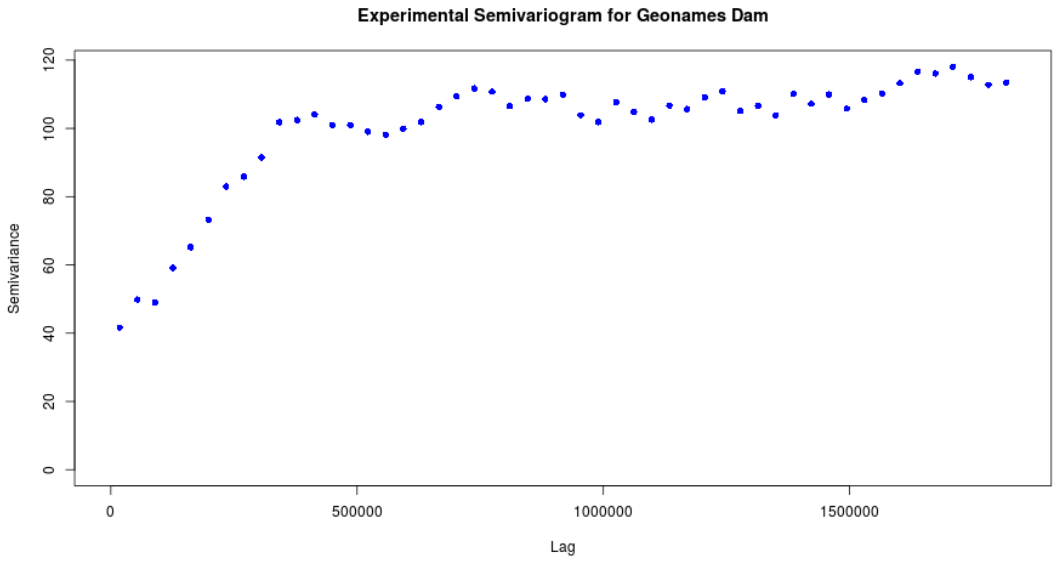
Cell value: number of places falling in the cell

Methods (cont.)

- Spatial autocorrelation analysis (cont.)

Spatial statistics

- **Global Moran's I**
- Sample Semivariogram (i.e. **semivariances at first, median and last lag distances**).



Methods (cont.)

- Spatial interaction with other geographic features

Spatial statistics (Internal)

- **Count of distinct nearest feature types**

$$CountNearest_i = \frac{m_i}{N}$$

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**

$$EntropyNearest_i = -\sum_{j=1}^{N} \frac{n_j}{N} \log(\frac{n_j}{N})$$

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.

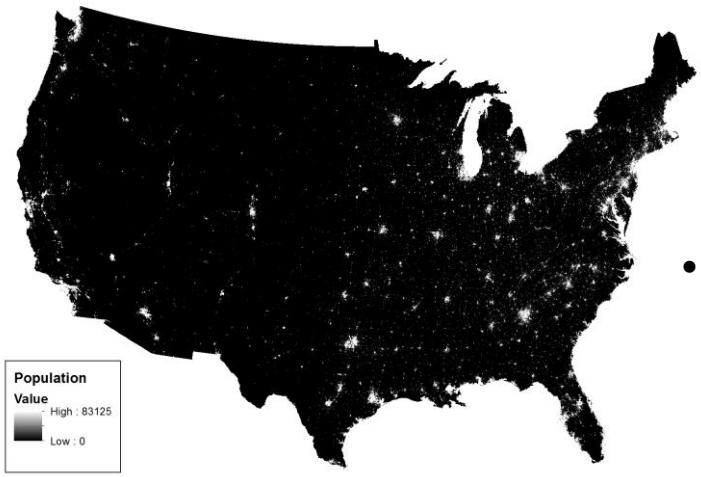
Methods (cont.)

- Spatial interaction with other geographic features

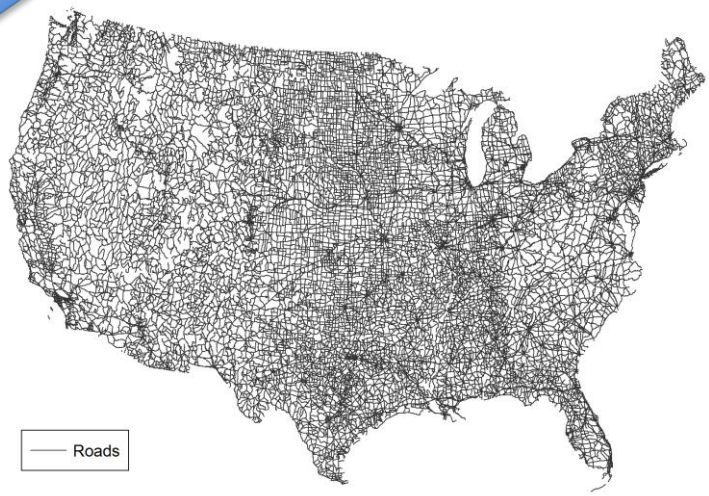
Spatial statistics (External)

Population (LandScan2014)
Population for each feature point

Road Segment (Digital Chart of the World)
Distance to nearest segment for each feature points



- **Minimum**
- **Maximum**
- **Mean**
- **Standard deviation**



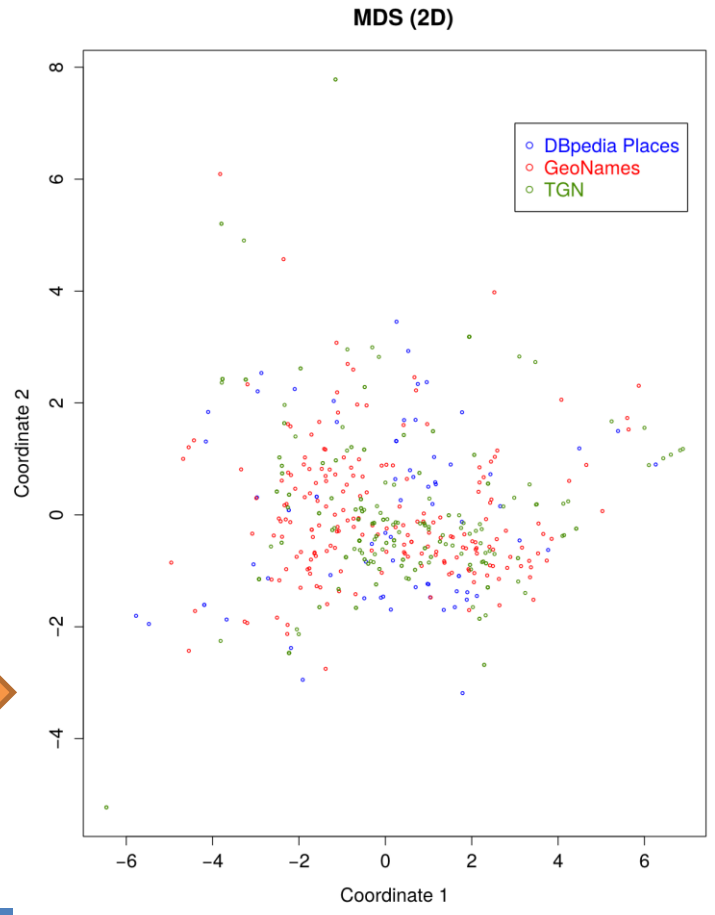
Methods (cont.)

27 statistical features

Transform high-dimensional feature space to 2-dimensional (Multidimensional scaling)

Visualize it in a 2D map

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			Entropy of nearest feature types
	Variance distance to nearest neighbor			
	Kernel density (bandwidth)			Population value (min)
	Kernel density (range)	Semivariogram value		Population value (max)
				Population value (mean)
				Population value (std dev)
				Shortest distance to road (min)
				Shortest distance to road (max)
				Shortest distance to road (mean)
				Shortest distance to road (std dev)

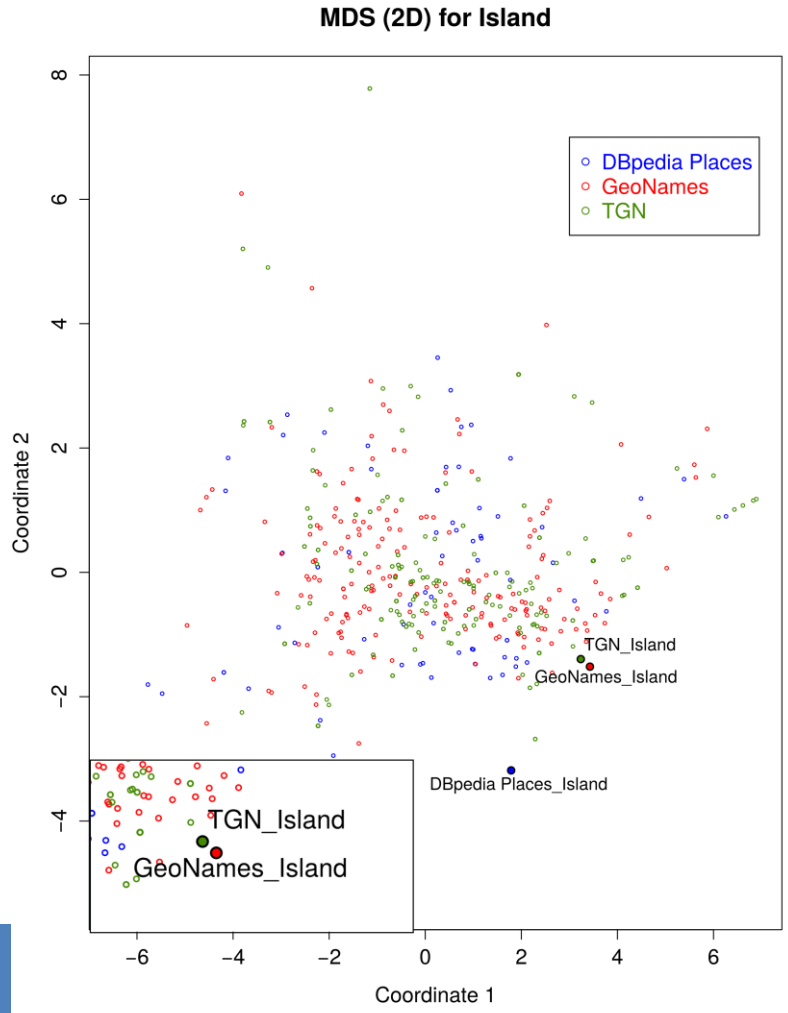


Case Studies

- Overview
  - Same name and similar spatial patterns
  - Same name but different spatial patterns
  - Different names but similar spatial patterns
  - Different names and different spatial patterns

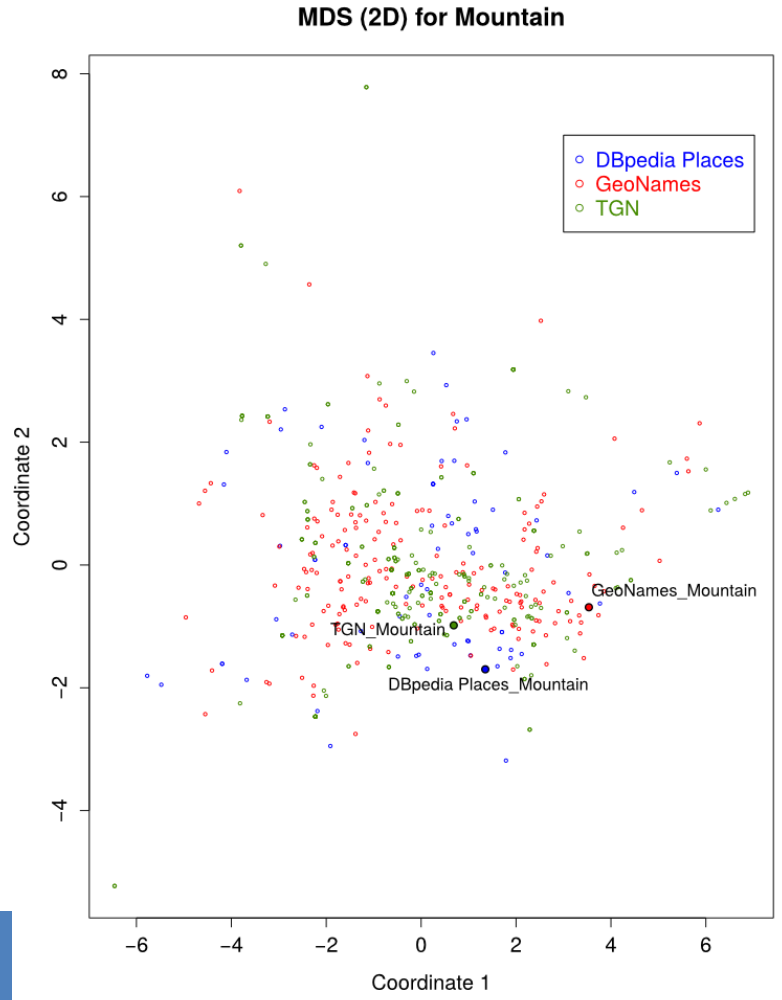
Case Studies

- Same name and similar spatial patterns



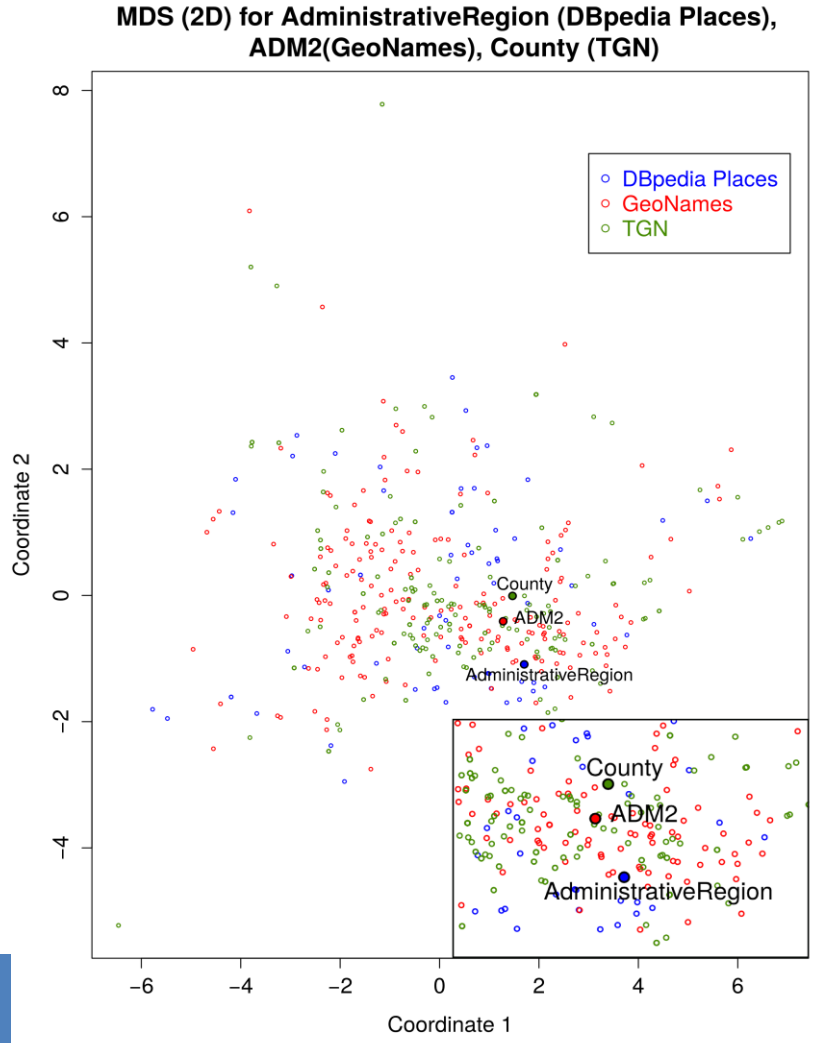
Case Studies

- Same name but different spatial patterns



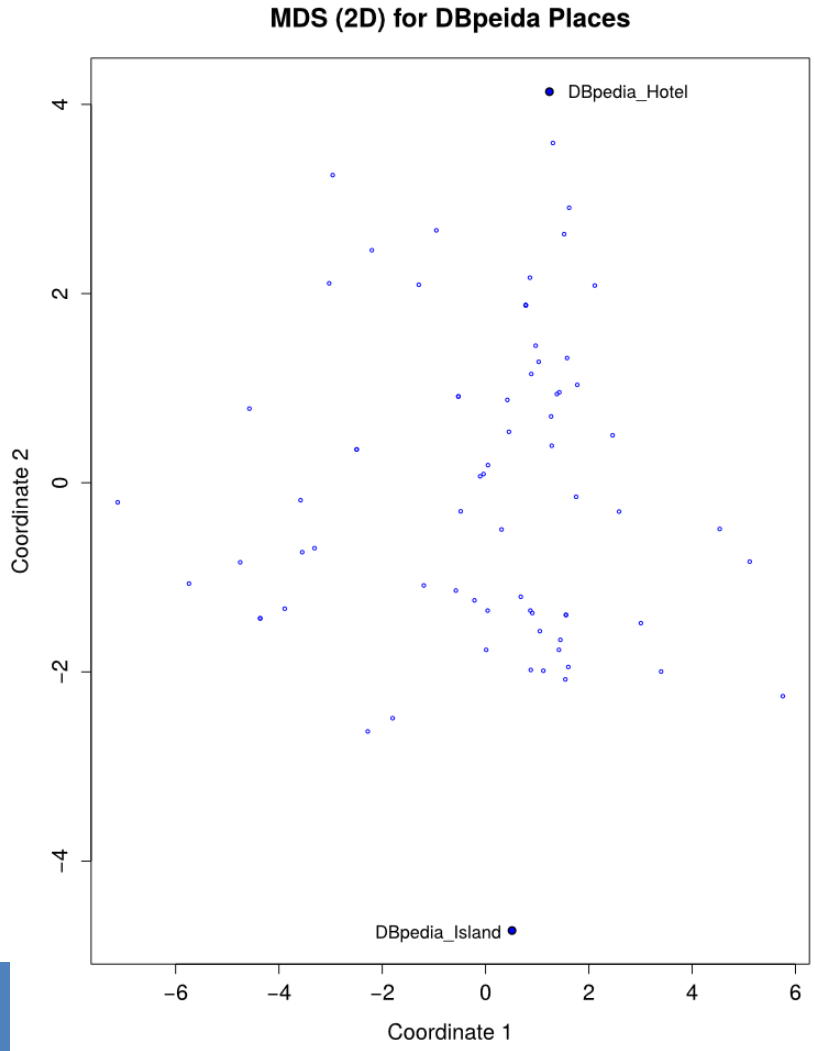
Case Studies

- Different names but similar spatial patterns



Case Studies

- Different names and different spatial patterns



Future work

- Derive **additional statistical** features to represent the spatial signature

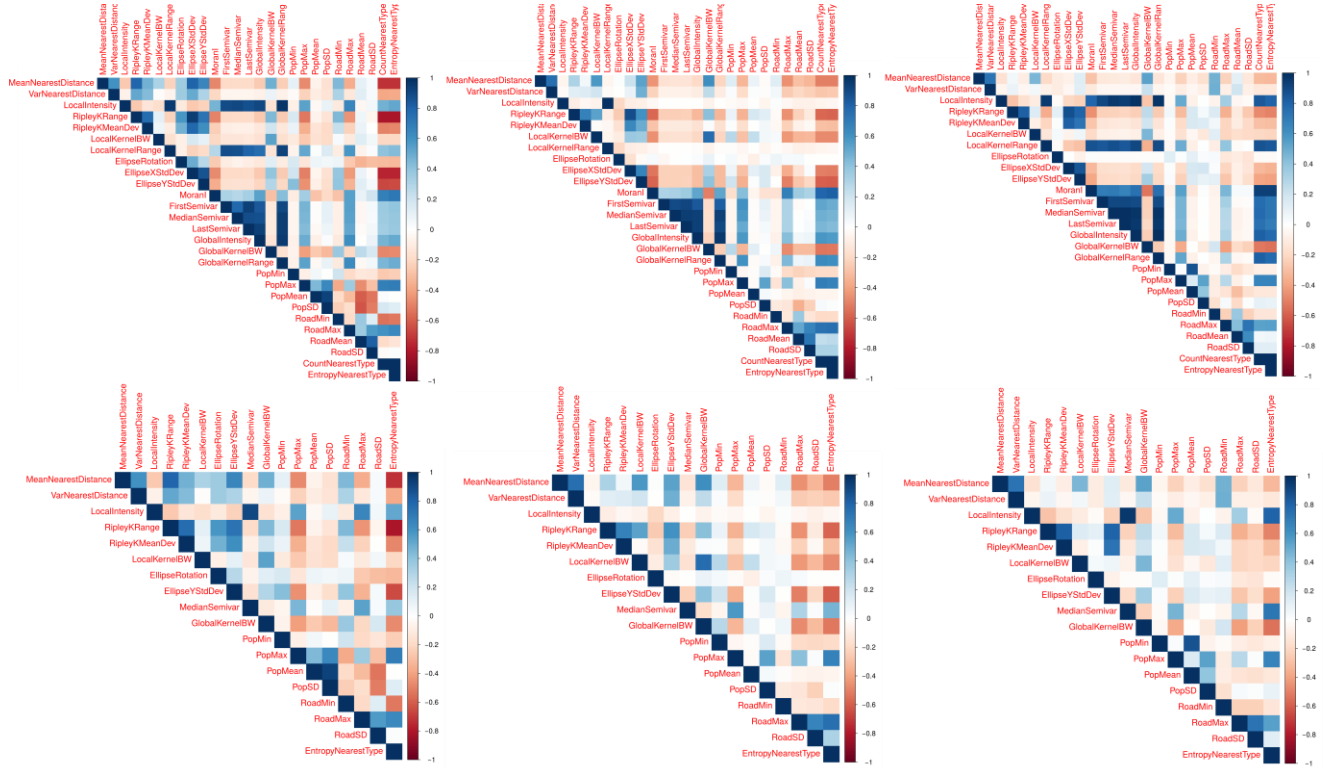
  (e.g. statistics for co-occurrence, topological relations);
- **Quantify the dissimilarity/similarity** of place types using such spatial signatures

  (e.g. supervised/ unsupervised learning algorithms);
- Combine spatial signatures with previously studied **temporal and thematic signatures**;
- Integrate this study (bottom-up) with **classical top-down knowledge engineering**.

Thank you!

Questions and comments?

27 Signatures
(Before reduction)



18 Signatures
(After reduction)

*. 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)

*. 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)

*. 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)

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