Understanding Semantic Uncertainty in Volunteered Geographic Information

Rui Zhu
University of California, Santa Barbara

AAG 2019 Session: GIScience in the Post-Truth Era I
Outline

- Motivations
- Methodology
- Experiments
- Summary & Discussions
Motivations
Motivation

Uncertainty of VGI:

- Positions: coordinates, addresses, postcodes, …
- Geometric representations and topological relations
- Attributes: number of check-ins, number of employees, ...
- Semantics: place types
Motivation

Semantic Uncertainty of VGI

- *Restaurants* in Foursquare and Google places might not be the same.
- *Mountains* in DBpedia Places are different from *Mountains* in GeoNames.
- A place should be labeled more likely as a restaurant or as a bar?
- Will spatial contexts help to reduce such uncertainties?
Methodology

Conventional approach for uncertainty analysis

- Numeric data
  - Distribution of the data → variance, entropy, Bayes theorem
- Categorical data
  - Indicator statistics
  - **Semantic signatures** → semantic distribution of place types
Methodology

Methodology

Spatial Signatures

- **Spatial structure** of the data belonging to a place type is used to quantify its semantics.
- **Spatial statistics** are applied to describe such spatial structure.
- Spatial point patterns, Spatial autocorrelation analysis, spatial interaction analysis with other geographic features, place-based analysis. → **41 statistics**
Methodology

Spatial Signatures - Spatial point patterns

- Intensity-based: local intensity, kernel density estimation
- Distance-based: nearest-neighbor distance, Ripley’s K, and standard deviational analysis
Methodology

Spatial Signatures - Spatial point patterns - Examples

Figure 1: Ripley’s K of Park (left) and Dam (right) from DBpedia Places.

Statistics: mean and std. of the deviation between theoretical can observed K curves
Methodology

Spatial Signatures - Spatial Autocorrelation Analysis

- Moran’s I: how intensities of cells differ from their neighbors
- Semivariogram: measure the variation of cell intensities in a specific distance lag class.

Dams in GeoNames

Cell size: 36 km * 22.2 km

Cell value: number of instances falling in the cell
Methodology

Spatial Signatures - Spatial Autocorrelation Analysis - Examples

Figure 2: Experimental semivariogram of Park (left) and Dam (right) from TGN.

Statistics: mean and std. of the semivariance at first, median and last lag distance
Methodology

Spatial Signatures - Spatial Interaction with Other Geographic features

- Population
- Climate
- Road network

Population (LandScan2014)
Population for each feature point

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

- Minimum
- Maximum
- Mean
- Standard deviation
Methodology

Spatial Signatures - Spatial Interaction with Other Geographic features - Examples

Road suffix distribution of **amusement park** and **restaurant** from Google Places
Methodology

Spatial Signatures - Spatial Interaction with Other Geographic features - Examples

Road suffix distribution of car dealer from Google Places and car dealership from Foursquare
Methodology

Spatial Signatures - Place-based statistics

In contrast to spatial statistics, place-based statistics focus more on describing the *topological* and *hierarchical relations* between places.

- The number (and entropy) of distinct states (or counties) a place type occurs in;
- The number (and entropy) of adjacent states (or counties) that also contain places of the same type;
Methodology

Spatial Signatures - Place-based statistics - Examples

- To distinguish feature types:
  - Glacier: found in eight US-states according to DBpedia
  - River: found in all states
# Methodology

<table>
<thead>
<tr>
<th>Spatial Point Pattern</th>
<th>Spatial Autocorrelations</th>
<th>Spatial Interaction with Other Geographic Features</th>
<th>Place-based statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td></td>
<td>Population</td>
<td>Number of distinct states (or counties)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Entropy of states (or counties)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Number of adjacent states (or counties) that have the same feature type</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Road Network</td>
<td>Number of distinct feature types for nearest neighbor</td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>Climate</td>
<td>LDA-based approach Mean KL Divergence of the topic distribution</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Entropy of the topic distribution</td>
</tr>
<tr>
<td>Intensity</td>
<td>Global Moran’s I</td>
<td>Population</td>
<td>Number of distinct states (or counties)</td>
</tr>
<tr>
<td>Mean distance to nearest neighbor</td>
<td></td>
<td></td>
<td>Entropy of states (or counties)</td>
</tr>
<tr>
<td>std. of distance to nearest neighbor</td>
<td></td>
<td></td>
<td>Number of adjacent states (or counties) that have the same feature type</td>
</tr>
<tr>
<td>Kernel density (range)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel density (bandwidth)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ripley’s K (range)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ripley’s K (mean deviation)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>std. ellipse (rotation)</td>
<td>Semivariogram (first distance lag)</td>
<td>mean precipitation</td>
<td></td>
</tr>
<tr>
<td>std. ellipse (std. along x-axis)</td>
<td></td>
<td>entropy of nearest road types</td>
<td></td>
</tr>
<tr>
<td>std. ellipse (std. along y-axis)</td>
<td></td>
<td>entropy of feature types for nearest neighbor</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>Semivariogram (median distance lag)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

1. Similarity of place types

Case 1

Case 2

Case 3
Experiments

2. Coreference resolution

Have to use the feature types in addition to string and spatial distances.

Which Kobani?
Summary & Discussions

- Semantic uncertainty of VGI has to be understood and quantified
- Semantic signatures are introduced to quantify the semantic uncertainty

In the future:

- Need a framework/guideline of using semantic signatures
- To use semantic uncertainty to infer other types of uncertainties
- From exploratory study to solving emergent VGI challenges:
  - federated geographic information retrieval, place alignment, data cleaning, …
- More advanced spatial /platial statistics could be incorporated into the signature set
Thanks a lot!

Any questions / comments?