Adopting Semantic Signatures to Quantify the Uncertainty of Volunteered Geographic Information

Rui Zhu
University of California, Santa Barbara
Outline

● Motivation
● Methodology
● Experiment
● Discussion
Motivation
Motivation

Uncertainty of VGI:

- Positions: coordinates, address, postcode …
- 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 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 Signature -- Spatial point patterns

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

Spatial Signature -- 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 Signature -- 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 Signature -- Spatial point patterns -- 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 Signature -- 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 Signature -- Spatial Interaction with Other Geographic features -- Examples

Figure 3: Histogram of road types for *Amusement Park* (left) and *Restaurant* from Google.
Methodology

Spatial Signature -- Place-based statistics

In contrast to spatial statistics, they 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 features of the target type;
Methodology

Spatial Signature -- Place-based statistics -- Examples

- Distinguishing feature types such as Glacier (which occur in eight US-states according to DBpedia) and River (which occur 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><strong>Local</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean distance to</td>
<td>Global Moran’s I</td>
<td>Population</td>
<td>Number of distinct</td>
</tr>
<tr>
<td>nearest neighbor</td>
<td></td>
<td></td>
<td>states (or counties)</td>
</tr>
<tr>
<td>std. of distance</td>
<td></td>
<td>min</td>
<td></td>
</tr>
<tr>
<td>to nearest neighbor</td>
<td></td>
<td>max</td>
<td></td>
</tr>
<tr>
<td>Kernel density</td>
<td></td>
<td>mean</td>
<td>Entropy of states</td>
</tr>
<tr>
<td>(range)</td>
<td></td>
<td></td>
<td>(or counties)</td>
</tr>
<tr>
<td>Kernel density</td>
<td></td>
<td>std.</td>
<td>Number of adjacent</td>
</tr>
<tr>
<td>(bandwidth)</td>
<td></td>
<td></td>
<td>states (or counties)</td>
</tr>
<tr>
<td>Ripley’s K</td>
<td></td>
<td></td>
<td>that have the</td>
</tr>
<tr>
<td>(range)</td>
<td></td>
<td></td>
<td>same feature type</td>
</tr>
<tr>
<td>Ripley’s K</td>
<td>Semivariogram (first</td>
<td>Road Network</td>
<td>Number of distinct</td>
</tr>
<tr>
<td>(mean deviation)</td>
<td>distance lag)</td>
<td></td>
<td>feature types for</td>
</tr>
<tr>
<td>std. ellipse</td>
<td></td>
<td></td>
<td>nearest neighbor</td>
</tr>
<tr>
<td>(rotation)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>std. ellipse</td>
<td>Entropy of nearest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(std. along x-axis)</td>
<td>road types</td>
<td></td>
<td></td>
</tr>
<tr>
<td>std. ellipse</td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(std. along y-axis)</td>
<td>precipitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global</strong></td>
<td>Semivariogram (median</td>
<td>Climate</td>
<td>LDA-based approach</td>
</tr>
<tr>
<td>Intensity</td>
<td>distance lag)</td>
<td></td>
<td>Mean KL</td>
</tr>
<tr>
<td>Kernel density</td>
<td>Semivariogram (last</td>
<td></td>
<td>Divergence of the</td>
</tr>
<tr>
<td>(range)</td>
<td>distance lag)</td>
<td></td>
<td>topic distribution</td>
</tr>
<tr>
<td>Kernel density</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(bandwidth)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiment

1. Similarity of place types
Experiment

2. Coreference resolution

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

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

However ...

- 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 complex spatial patterns could be quantified and incorporated into the signature set;
- To design new spatial-semantic interaction statistics: modified LISA, Ripley’s K …
Thanks a lot!

Any questions / comments?