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

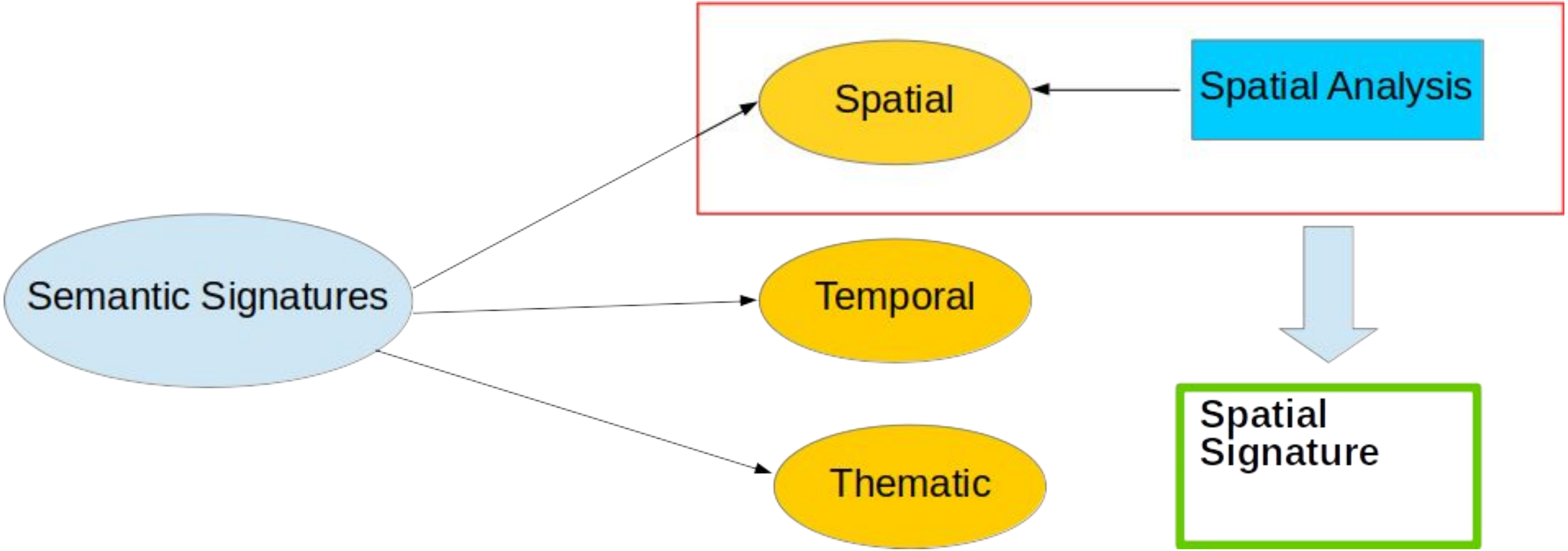
- *Restaurants* in Foursquare and Google places might not be the same.
- *Mountains* in DBpedia Places are different from *Mountains* in GeoNames.
- A place should be labeled more likely as a restaurant or as a bar?
- Will spatial contexts help to reduce such uncertainties?

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



Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., and Gao, So. (2018): Using Semantic Signatures for Social Sensing in Urban Environments. Mobility Patterns, Big Data and Transport Analytics.

Methodology

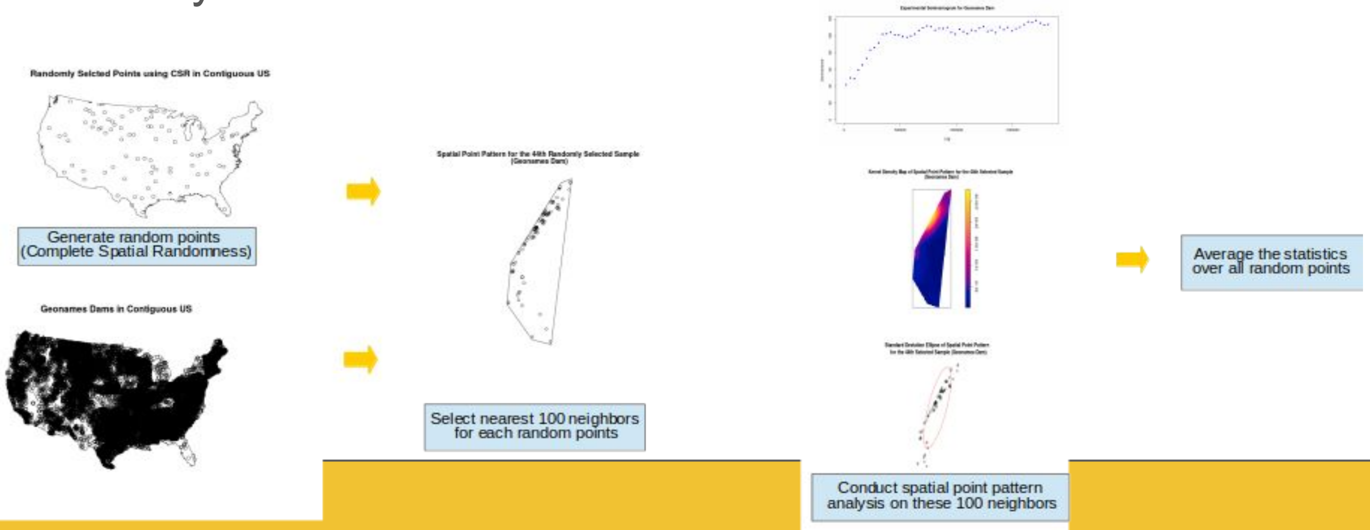
Spatial Signatures

- ***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

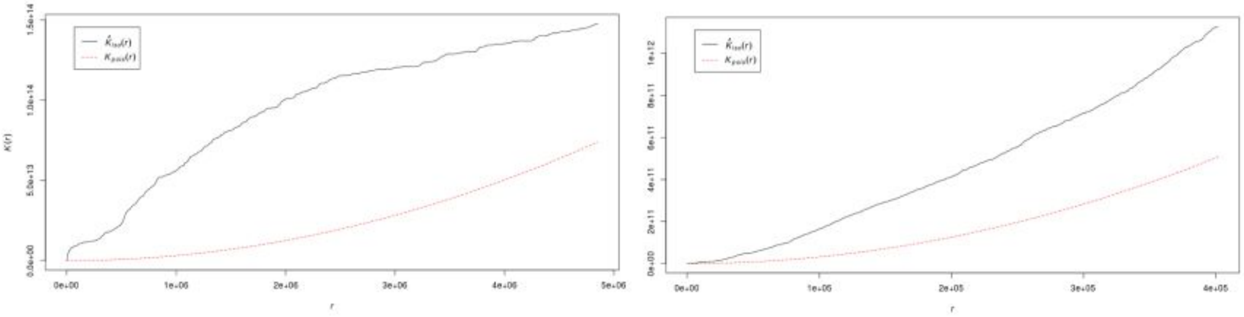


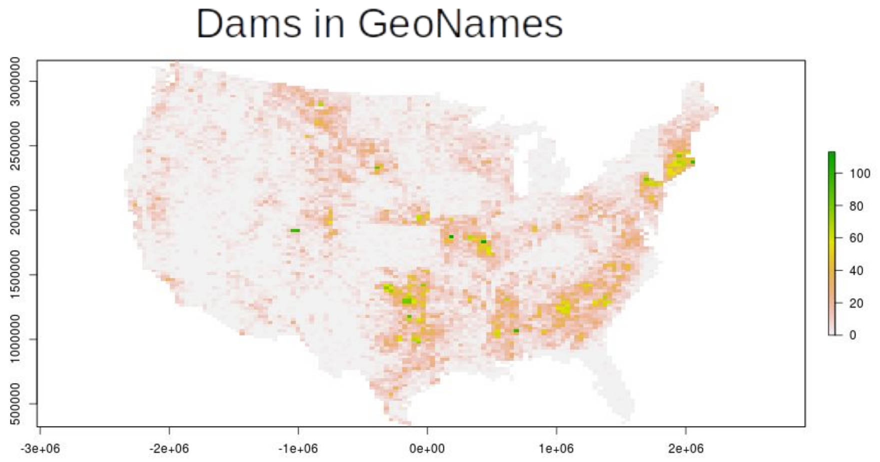
Figure 1: Ripley's K of *Park* (left) and *Dam* (right) from DBpedia Places.

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.



Cell size : 36 km * 22.2 km

Cell value: number of instances falling in the cell

Methodology

Spatial Signatures - Spatial Autocorrelation Analysis - Examples

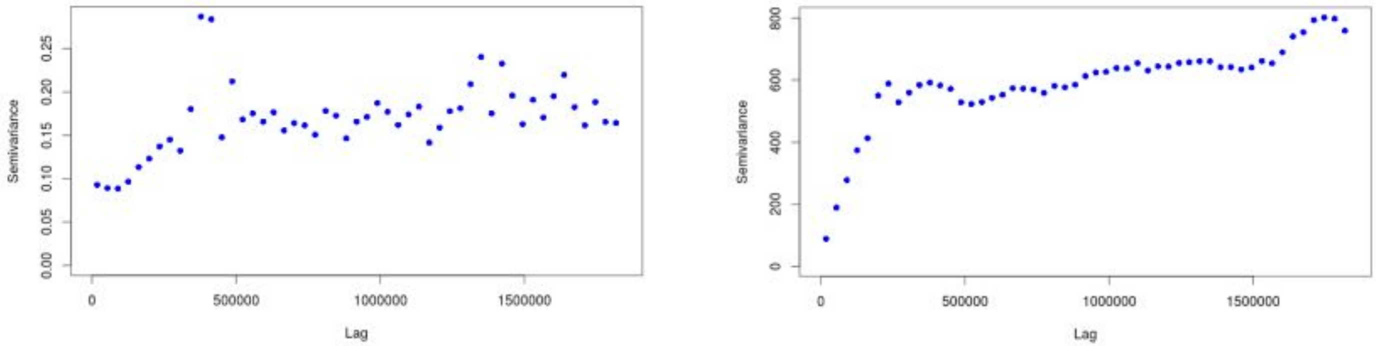


Figure 2: Experimental semivariogram of *Park* (left) and *Dam* (right) from TGN.

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)

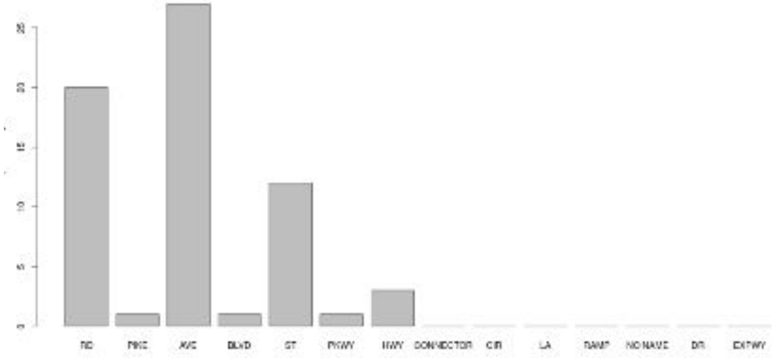
Distance to nearest segment for each feature point

- Mimimum
- Maximum
- Mean
- Standard deviation

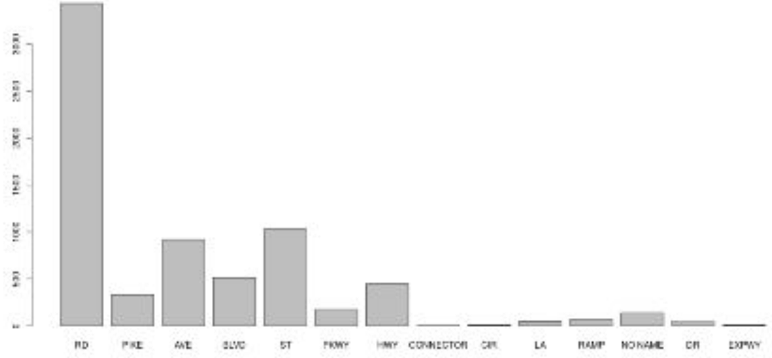
Methodology

Spatial Signatures - Spatial Interaction with Other Geographic features - Examples

amusement park (Google Places)



restaurant (Google Places)

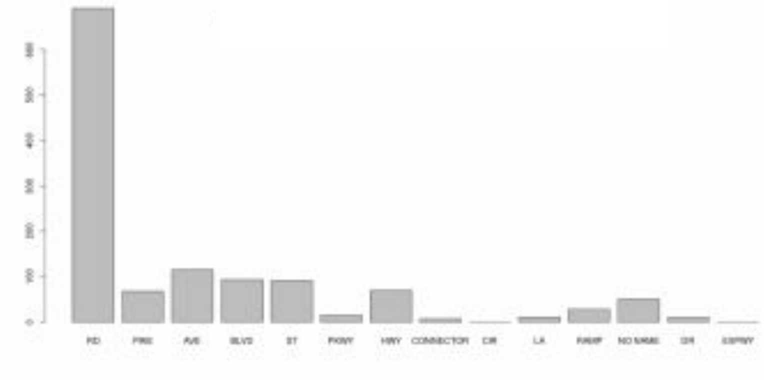


Road suffix distribution of *amusement park* and *restaurant* from Google Places

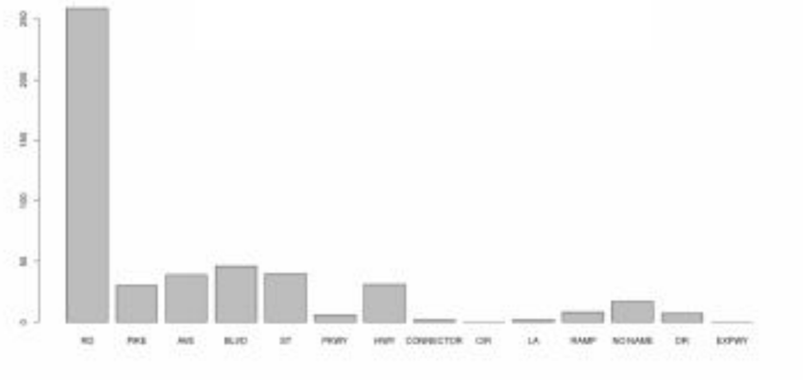
Methodology

Spatial Signatures - Spatial Interaction with Other Geographic features - Examples

car dealer (Google Places)



car dealership (Foursquare)



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

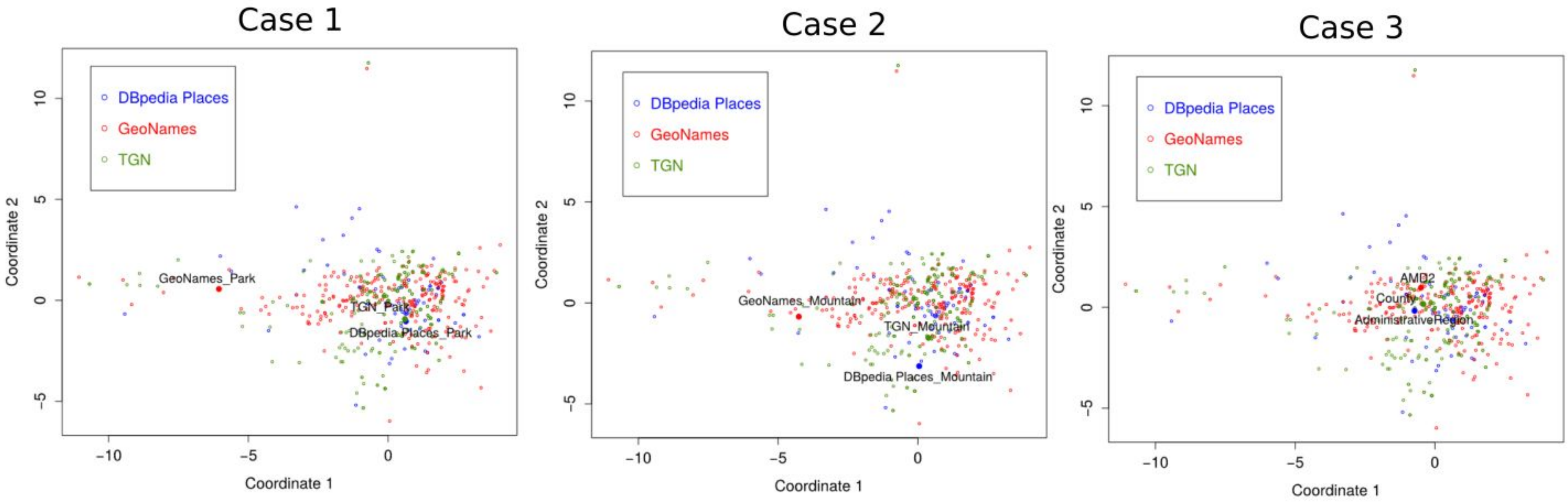
- To distinguish feature types:
    - Glacier: found in eight US-states according to DBpedia
    - River: found in all states

Methodology

<table>
<tr><th colspan="2">Spatial Point Pattern</th><th>Spatial Autocorrelations</th><th colspan="2">Spatial Interaction with Other Geographic Features</th><th colspan="2">Place-based statistics</th></tr>
<tr><td rowspan="10">Local</td><td>Intensity</td><td rowspan="7">Global Moran's I</td><td rowspan="4">Population</td><td>min</td><td colspan="2" rowspan="2">Number of distinct states (or counties)</td></tr>
<tr><td>Mean distance to nearest neighbor</td><td>max</td></tr>
<tr><td>std. of distance to nearest neighbor</td><td>mean</td><td colspan="2">Entropy of states (or counties)</td></tr>
<tr><td>Kernel density (range)</td><td>std.</td><td colspan="2" rowspan="3">Number of adjacent states (or counties) that have the same feature type</td></tr>
<tr><td>Kernel density (bandwidth)</td><td rowspan="5">Road Network</td><td>min of shortest distance</td></tr>
<tr><td>Ripley's K (range)</td><td>max of shortest distance</td></tr>
<tr><td>Ripley's K (mean deviation)</td><td>mean of shortest distance</td><td colspan="2" rowspan="2">Number of distinct feature types for nearest neighbor</td></tr>
<tr><td>std. ellipse (rotation)</td><td rowspan="3">Semivariogram (first distance lag)</td><td>std. of shortest distance</td></tr>
<tr><td>std. ellipse (std. along x-axis)</td><td>entropy of nearest road types</td><td colspan="2" rowspan="2">Entropy of feature types for nearest neighbor</td></tr>
<tr><td>std. ellipse (std. along y-axis)</td><td rowspan="8">Climate</td><td>mean precipitation</td></tr>
<tr><td rowspan="7">Global</td><td rowspan="3">Intensity</td><td rowspan="4">Semivariogram (median distance lag)</td><td>std. precipitation</td><td rowspan="7">LDA-based approach</td><td rowspan="4">Mean KL Divergence of the topic distribtion</td></tr>
<tr><td>mean temperature max</td></tr>
<tr><td>std. temperature max</td></tr>
<tr><td rowspan="2">Kernel density (range)</td><td>mean temperature min</td></tr>
<tr><td rowspan="3">Semivariogram (last distance lag)</td><td>std. temperature min</td><td rowspan="3">Entropy of the topic distribution</td></tr>
<tr><td rowspan="2">Kernel density (bandwidth)</td><td>mean water vapor pressure</td></tr>
<tr><td>std. water vapor pressure</td></tr>
</table>

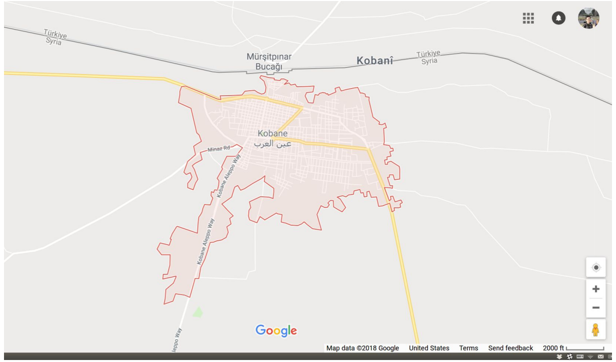
Experiments

1. Similarity of place types



Experiments

2. Coreference resolution



GeoNames Home | Postal Codes | Download / Webservice | About

Kobani — all countries

search | show on map | [advanced search]

13 records found for "Kobani"

	Name	Country	Feature class	Latitude	Longitude
1	Kobani Kobani,Kohani	Mali, Sikasso	intermittent stream	N 11° 5' 23''	W 6° 49' 47''
2	'Ayn al 'Arab Aarab Peunar,Aarab Peunâr,Ain el Aarab,Arab Peunar,Aïn el Aarab,Ein-al-Arab,Kobane,Kobani,`Arab Bina...	Syria, Aleppo	seat of a second-order administrative division population 50,000	N 36° 53' 27''	E 38° 21' 12''
3	Mkoani Kobani,Mkoani	Tanzania, Pemba South Mkoani District > Mbuyuni	populated place	S 5° 22' 0''	E 39° 39' 0''
4	Nāḩiyat Markaz 'Ayn al 'Arab Kobane,Kobani,Kobanê,Kobanî,Kubane,Kubani,Kubanê,Kubanî,Kūbānī,Nahiyat Markaz `Ayn al `Arab,Nāḩiyat ...	Syria, Aleppo	third-order administrative division	N 36° 48' 17''	E 38° 23' 27''
5	Kobani	Ivory Coast, Savanes	intermittent stream		
6	Kobani	Ivory Coast, Denguélé	intermittent stream		
7	Kobani	Ivory Coast, Woroba	intermittent stream		
8	Kobani	Ivory Coast,	intermittent stream		
9	Kobani	Ivory Coast, Savanes	stream		
10	Kobani	Ivory Coast, Santiago Metropolitan Region	stream		
11	Kobani	Ivory Coast, Denguélé	stream		
12	Kobani	Ivory Coast, Denguélé	stream		
13	Razvaliny Kobani Razvaliny Kobani	Georgia,	ruin(s)		

Which Kobani?

DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

About: Kobanî

An Entity of Type : settlement, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Kobanî (Kurdish: کۆبانی pronounced [koˈbaːniː], also rendered Kobanê [koˈbaːne]), also known as Ayn al-Arab (Arabic: عين العرب North Levantine pronunciation: [ʕeːn elˈʕɑrɑb]), is a city in the Aleppo Governorate in northern Syria, lying immediately south of the border with Turkey. As a consequence of the Syrian Civil War, the city has been under control of the Kurdish YPG militia since 2012.

Property	Value
dbo:PopulatedPlace/areaTotal	7.0
dbo:abstract	Kobanî (Kurdish: کۆبانی pronounced [koˈbaːniː], also rendered Kobanê [koˈbaːne]), also known as Ayn al-Arab (Arabic: عين العرب North Levantine pronunciation: [ʕeːn elˈʕɑrɑb]), is a city in the Aleppo Governorate in northern Syria, lying immediately south of the border with Turkey. As a consequence of the Syrian Civil War, the city has been under control of the Kurdish YPG militia since 2012. In 2014, it was unofficially declared to be the administrative center of the Kobanî Canton of Rojava.From September 2014 to January 2015, the city was under siege by Islamic State of Iraq and the Levant. Most of the city was destroyed and most of the population fled to Turkey. In 2015, many returned and reconstruction began.Prior to the Syrian Civil War, Kobanî was recorded as having a population of close to 45,000. The majority of inhabitants were Kurds, with Arab, Turkmen, and Armenian minorities. (en)
dbo:areaTotal	7000000.000000 (xsd:double)
dbo:country	dbr:Syria
dbo:elevation	520.000000 (xsd:double)
dbo:isPartOf	dbr:Ayn_al-Arab_District dbr:Aleppo_Governorate

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

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?