MAXIMA REU in Interdisciplinary Mathematics

Spatial Knowledge and Relatedness

Rebecca Harper[1], Huy Mai[2], Benjamin Mathers[3], Samuel Horlbeck Olsen[4] and Laura Vonessen[5]

1: Willamette University, 2: Brandeis University, 3: Macalester College, 4: Macalester College, 5: University of Arizona

1 Introduction

Waldo Tobler's First Law of Geography: Everything is related to everything else, but near things are more related than distant things.

Using an online survey, we compared human judgments of relatedness to several measures of "distance." We used this "gold standard" data to empirically study three research questions:

RQ1: Does Tobler's First Law hold?

We also studied whether an individual from California perceives the relationship between "Burbank, CA" and "Compton, CA" differently from somebody who is not "a local."

RQ2: Does spatial familiarity affect perceived relatedness?

Finally, we studied the effectiveness of state-of-the-art Semantic Relatedness (SR) algorithms from natural language processing (NLP).

Semantic relatedness (SR): A function of how related two concepts are.

$$SR : Concepts \times Concepts \to [0,1]$$

e.g. $SR(MN, WI) = 0.8$ but $SR(MN, FL) = 0.3$

RQ3: How effective are SR algorithms for geospatial concepts?

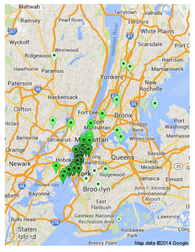
2 Data Extraction

WikiBrain: An open-source software library that supports algorithmic, geospatial, and multi-lingual data mining of Wikipedia.

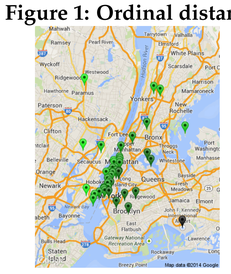
Our research required a variety of datasets. We mined most of these datasets from Wikipedia using WikiBrain.

- **Geodetic distance:** Real distance on the globe, measured in kilometers extracted from Wikipedia coordinate locations entered by humans.
- **Article popularity:** The number of page views for each article for selected days throughout the year, which allowed us to restrict our survey to spatial concepts that were relatively well-known.
- **SR distance:** The estimated relatedness of two spatial concepts. Computed algorithmically by WikiBrain.
- **Scale:** The "type" of concept, i.e. landmark, city, county, state, natural place, or other, classified using a manually curated list of keywords.
- **Ordinal distance:** The smallest number of edges required to travel from one point to another in a directed graph where each location has an edge to the 10 other locations closest to it.

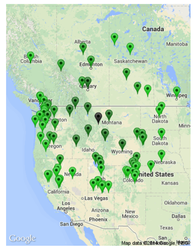
Figure 1: Ordinal distance



Manhattan is in an area densely populated with well-known concepts



The JFK Airport can get to Manhattan in two hops, but Manhattan can't get to JFK at all



Helena, MT is in a much sparser area, so it reaches much farther in only four hops.

3 Question Selection

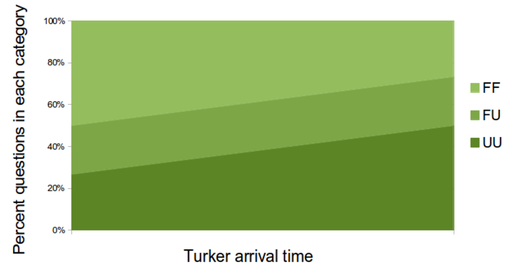
Given a respondent's home, questions are selected that vary in expected familiarity.

Question familiarity: FF (familiar with both locations), **FU** (familiar with one location), and **UU** (unfamiliar with both locations).

The survey balances between UU questions (which are easy to find respondents for), and FF (which require specialized spatial knowledge).

- Give early respondents lots of FF
- Ask later respondents lots of UU previous questions

Figure 2: Question change over time



Balancing the questions:

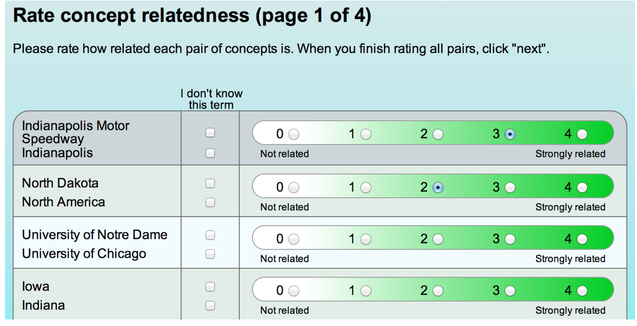
- More questions than necessary are generated.
- Final questions are chosen by varying *scale*, *geodetic distance*, and *SR*.
- Turker gets questions which together get closest to a target distribution.

4 Survey Design

To collect our data, we launched a survey on Amazon Mechanical Turk, crowdsourcing micro-market run by Amazon where users, called Turkers, can complete Human Intelligence Tasks (HITs) for a small amount of pay. We collected 1,000 responses in batches of 100 responses collected throughout the day.

We required Turkers to come from the top nine countries for Turkers according to AppAppeal–in order of percent participation, the United States, India, Pakistan, the United Kingdom, Australia, Brazil, Canada, Spain, and France.

Figure 3: Relatedness Ratings



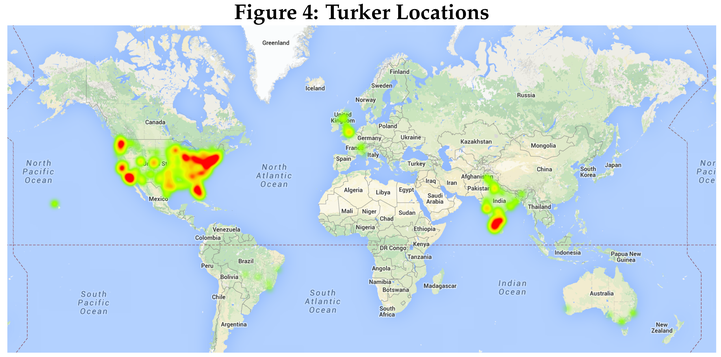
There were four main sections of the survey:

1. **Background information:** Collect basic demographic information and locations where the turker has lived.
2. **Relatedness Ratings:** Use background information to generate questions and collect relatedness ratings
3. **Familiarity Ratings:** Turkers rank how familiar they are with all spatial concepts in the previous section
4. **Valence Ratings:** Turkers answer how much they would like to live in or near each spatial concept. This tries to capture the intrinsic attractiveness of concepts.

5 Results

1,000 Turkers from around the world provided 37,000 SR assessments

Figure 4: Turker Locations



The responses yielded:

- 34,422 answers to spatial questions (excluding validation questions)
- 1,129 distinct spatial questions
- 287 distinct questions with 5 or more FF ratings and 5 or more UU ratings
- For 6,053 answers, the respondent did not recognize at least one location

RQ1: Does Tobler's First Law hold?

Yes. Concept distance and concept relatedness were inversely correlated according to Spearman's correlation for both ordinal distance ($\rho_s = -0.49$) and geodetic distance ($\rho_s = -0.50$).

RQ2: Does spatial familiarity affect perceived relatedness?

Yes. The mean relatedness differed for FF ($\mu = 3.49$) and UU ($\mu = 2.99$). A paired t-test shows this to be significant ($p < 0.0001$). Several questions showed particularly noticeable differences between FF and UU:

Figure 5: Spatial Familiarity vs. Perceived Relatedness

Concept 1	Concept 2	FF	UU	FF - UU
Chhatrapati Shivaji Airport	Maharashtra	0.97	0.29	0.68
Minnesota	Iowa	0.64	0.27	0.37
Oregon	Columbia River	0.89	0.55	0.34
Minneapolis	Chicago	0.47	0.22	0.24
Seattle-Tacoma Airport	Olympic Nat'l Park	0.46	0.46	0.00
Mount Rushmore	Des Moines, IA	0.07	0.23	-0.16
Burbank, CA	Compton, CA	0.29	0.60	-0.32

RQ3: How effective are SR algorithms for geospatial conceps?

Relatively effective. The Spearman Correlation between a linear combination of several SR algorithms and the user's reported relatedness is 0.59.

6 References

[1] B. Hecht and E. Moxley, "Terabytes of Tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge," in *Proc. 9th Int. Conf. Spatial information theory* (COSIT '09), K.S. Hornsby, C. Claramunt, M. Denis, and G. Ligozat (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 88-105.

[2] S. C. Hirtle and M. F. Mascolo, "Effect of semantic clustering on the memory of spatial locations," in *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vo. 12, no. 2, 1986, pp. 182-189.

[3] B. Hecht and S. Sen, "'Arafat' and 'peace': cultural communities and algorithmic gold standards," to appear in *Computer Supported Cooperative Work and Social Computing*, 2015.

[4] E. Kittur et al, "Crowdsourcing user studies with Mechanical Turk," in *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI '08), Florence, Italy, 2008, pp. 453-456.

Faculty Advisor: Shilad Sen (Macalester) **Research Mentor:** Matthew Wright (UMN) **Problem Poser:** Brent Hecht (UMN)

Research Supported by the NSF and the IMA