Leveraging Advances in Natural Language Processing to Better Understand Tobler's First Law of Geography

Toby Jia-Jun Li
Dept. of Comp. Sci. and Engineering
University of Minnesota
Minneapolis, MN

Shilad Sen
MSCS Department
Macalester College
St. Paul, MN

Brent Hecht
Dept. of Comp. Sci. and Engineering
University of Minnesota
Minneapolis, MN

ABSTRACT
Tobler's First Law of Geography (TFL) is one of the key reasons why "spatial is special". The law, which states that "everything is related to everything else, but near things are more related than distant things", is central to the management, presentation, and analysis of geographic information. However, despite the importance of TFL, we have a limited general understanding of its domain-neutral properties. In this paper, we leverage recent advances in the natural language processing domain of semantic relatedness estimation to, for the first time, robustly evaluate the extent to which relatedness between spatial entities decreases over distance in a domain-neutral fashion. Our results reveal that, in general, TFL can indeed be considered a globally recognized domain-neutral property of geographic information but that there is a distance beyond which being nearer, on average, no longer means being more related.

Categories and Subject Descriptors
H.1.0 [**Information Systems**]: Model and Principles – *General*

Keywords
Tobler's First Law of Geography, semantic relatedness, distance

1. INTRODUCTION

"Everything is related to everything else, but near things are more related than distant things" – Waldo Tobler [19]

When Waldo Tobler wrote the above words, he did not expect he would be laying part of the "bedrock of geographic thought" and developing "major geographic canon" [18]. Nevertheless, the above passage is now referred to as Tobler's First Law of Geography (TFL), and plays a key role in domains of study ranging from spatial data mining to spatial analysis to regional geography.

While TFL is widely recognized as critical to our understanding of geographic information, it is not well-understood in general, domain-neutral terms. Countless studies have observed TFL in a variety of individual phenomena (e.g. cotton growth, distribution of American palms, conflict in Africa), but it has not been robustly evaluated as a wide-ranging geographic principle. In fact, the literature specifically on TFL is quite narrow, largely consisting of a series of short, qualitative reflections on TFL in an issue of the *Annals of the Association of American Geographers* [18] and an earlier study upon which this work builds [11].

In this paper, we address this gap in the literature by performing the first robust, domain-neutral evaluation of TFL. Our work is enabled by recent advances in the natural language processing (NLP), particularly by the development of *semantic relatedness* (SR) algorithms that achieve near upper-bound performance against benchmark datasets. These algorithms enable us to, for the first time, robustly operationalize and quantify the term "related" in Tobler's First Law. In doing so, they allow us to get a quantitative, domain-neutral understanding of the relationship between distance and relatedness that is the subject of TFL.

At a high-level, our approach is straightforward: since SR algorithms largely rely on Wikipedia as their source of *world knowledge*, we calculate the semantic relatedness between pairs of *geographic Wikipedia articles*, or Wikipedia articles about entities with a geospatial footprint [9] (e.g. the article "Dallas" and the article "Houston"). Comparing the semantic relatedness between articles to the distance between their corresponding geospatial footprints – and doing so for millions of article pairs – allows us to evaluate the extent to which "everything is related to everything else, but near things are more related than distant things".

In order to gain a full understanding of domain-neutral TFL, we execute this high-level approach using a variety of different SR algorithms, each based on a different Wikipedia-based knowledge resource. Finally, research has shown that each language edition of Wikipedia is quite different from the others and reflects the unique understanding of world knowledge held by the speakers of the corresponding language (e.g. [1, 10]). As such, in our evaluation of TFL, we use SR algorithms operating on each of 20 of the largest Wikipedia languages.

No matter the language edition or the SR algorithm, we find strong, domain-neutral support for the negative association between distance and relatedness described in TFL, but only *up to a certain distance.* After that distance, on average two geographic entities that are closer together are not necessarily more related than two entities that are farther apart. This phenomenon occurs regardless of SR algorithm and Wikipedia language edition.

In summary, this work presents three high-level contributions:

(1) We leverage advances in semantic relatedness estimation to perform the first robust, domain-neutral analysis of Tobler's First Law.
(2) We do so using multiple SR algorithms, and multiple conceptions of world knowledge.

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

ACM 978-1-4503-3131-9/14/11...$15.00
http://dx.doi.org/10.1145/2666310.2666493

(3) We show that, in a domain-neutral fashion, Tobler's First Law only applies within a distance threshold. Beyond this distance threshold, two geographic entities (regardless of their domain) that are closer together are not necessarily on average more related than two entities that are farther apart.

2. RELATED WORK

Tobler's First Law made its initial appearance in Tobler's presentation on his computational urban growth model in 1969. TFL was then first published the year after [19]. While TFL has been widely accepted as a fundamental principle of geographic information, when it is explicitly discussed in the literature, it is usually in an applied context for individual studies in specific domains like the growth of cotton [14], the pattern of species richness and species composition [2], and the political conflicts and alliances in Africa [13].

The primary motivation for this work comes from the research of Hecht and Moxley [11], who found that the probability of two Wikipedia pages being directly linked declines over geodesic distance. However, Hecht and Moxley did not establish the *robustness* of this metric by comparing it against benchmark relatedness datasets. More recent work has established that direct Wikipedia links are a relatively weak proxy for semantic relatedness [6]. In addition, Hecht and Moxley only examined TFL at a high-level and did not look at how TFL varies over distance in detail, as we do here.

3. DATA AND METHODS

As noted in the introduction, our high-level approach to evaluating TFL is as follows: First, we calculate the semantic relatedness between many geographic Wikipedia article pairs. Next, we calculate the distance between the geographic entities described by the two articles in each pair. Finally, we compare semantic relatedness estimate for each pair to the distance for each pair.

3.1 Dataset

In order to capture the diverse perspectives of world knowledge present across the different language editions of Wikipedia (e.g. [1, 10]), our experiment was conducted utilizing 20 language editions of Wikipedia, using Wikipedia dump file as of May 2014 (English, Dutch, German, Swedish, French, Italian, Russian, Spanish, Polish, Japanese, Vietnamese, Portuguese, Chinese, Catalan, Norwegian, Finnish, Czech, Korean, Arabic and Hungarian).

In this research, we are interested in geographic Wikipedia articles, which are articles tagged with latitude/longitude coordinates. We use the versions of these tags that have been imported into *Wikidata*, a new Wikimedia project that seeks to serve as a central database for all language editions of Wikipedia.

SR Algorithm	Knowledge Resource	*WordSim353* Accuracy
Ensemble	Linear Combination	0.76
Explicit Semantic Analysis	Article Text	0.69
MilneWitten	Link Graph	0.59
Inlink	Link Graph	0.52
Outlink	Link Graph	0.55
Category	Category Graph	0.48

Table 1. SR algorithms used in our study.

The "Geoweb Scale Problem" (GSP) [7] is an issue that arise from large-area geospatial entities (e.g. Alaska) being represented as single latitude and longitude coordinates. In this context of this work, this would result in, for instance, a geodesic distance measurement of 362km for "Dallas" and "Texas", when in fact the corresponding geospatial entities have a containment relationship (suggesting a geodesic distance of 0). To prevent the GSP from affecting our results in this fashion, we follow standard procedure for the addressing GSP in Wikipedia [8] and remove all countries and first-order administrative districts using the GADM database [5] from our dataset of geographic Wikipedia articles in all languages.

Finally, there are too many geographic articles to tractably include all article *pairs* in each evaluation. For instance, there are 905,617^2 = 8.2×10^11 geographic article pairs in the English Wikipedia. Except where noted, we consider a random sample 100,000 geographic article pairs for each experiment.

3.2 Semantic Relatedness Algorithms

We use semantic relatedness algorithms (SR) to quantify the term "related" in Tobler's First Law. SR algorithms output a single numeric estimate between 0 and 1 of the strength of relationships between any two entities *a* and *b* [10]. For instance, an accurate SR algorithm would output a high value for the entity pair (*GIScience, Michael Goodchild*) but likely output a low value for the entity pair (*GIScience, Kim Kardashian*). In this paper, we limit our attention to geographic entities (e.g. *Houston, Dallas, Calgary*) represented by geographic Wikipedia articles.

Historically, semantic relatedness algorithms used expert-curated world knowledge repositories like WordNet (e.g. [3, 15]). However, since Strube and Ponzetto published their influential paper on using Wikipedia for SR estimation [17], Wikipedia has been the repository of choice among SR researchers. Wikipedia-based SR algorithms have been shown to perform very well against standard SR benchmark datasets like *WordSim353* [4], often matching upper bound performance (human intra-annotator agreement) [6].

This study considered six Wikipedia-based SR measures (Table 1). Three leverage the link graph, including the widely-used MilneWitten algorithm [17] that measures the overlap in both inbound and outbound links to an article. We included two SR metrics that isolated measurements of inbound / outbound overlap (Inlink, Outlink). We also implement a variant of Strube and Ponzetto's algorithm, which utilizes the category graph[24], and Explicit Semantic Analysis [7], which is the most commonly-employed SR algorithm that mines Wikipedia article text. Finally, we developed an ensemble SR algorithm that linearly combines all five other SR measures, trained and tested (with 10-fold cross-validation) on the benchmark ratings in WordSim353.

In order to ensure that our SR implementations replicated published performance, we calculated the Spearman's correlation coefficient between their output and the values in *WordSim353*, the standard means of evaluating SR algorithms. The results in Table 1 are consistent with published results. As expected, the ensemble SR algorithm outperformed all other SR measures. As such, where a single SR algorithm is required in our experiments, we utilize the ensemble algorithm.

All of our SR implementations are contained within our free and open-source Java Wikipedia software library, *WikiBrain* [16]. The code for all of our experiments is available as cookbook examples in the WikiBrain library to better facilitate replication and extension of this work.

4. RESULTS

We begin our evaluation of TFL by examining the semantic relatedness of geographic Wikipedia articles across simple geodesic distance. To do so, we look at all 20 Wikipedia language editions and utilize our ensemble SR algorithm.

Figure 1 contains the results of this analysis. In the figure, all entities pairs are binned by their geodesic distance using the bin size of 50km. The average relatedness is indicated on the y-axis, with relatedness assessed by percentile[1] rather than absolute value in order to easily compare results across language editions. The black line in Figure 1 indicates the average across all language editions, weighted by the number of first-language speakers of each corresponding language as indicated by the Ethnologue dataset [12]. The colored lines indicate specific language editions.

Examining the left-hand side of the graph, it is easy to see that, as predicted by TFL, there is a strong negative association between distance and relatedness for all language editions of Wikipedia as well as their weighted average. Indeed, for the smallest distance bin, the average SR percentile is a very high 0.93, but the average SR percentile is 0.7 at 1000km. In other words, two geographic entities that are separated by less than 50km have an average relatedness in the 93rd percentile, while that for two entities separated by 1000-1050km are in the 70th percentile.

However, moving further to the right in the graph, we see that this negative association decreases and even reverses. In other words, beyond a certain distance, there are many cases where entities in a nearer distance class are *less* related than entities in a farther distance class on average. This is contrary to TFL and suggests that the negative correlation described in TFL may not hold beyond some distance threshold.

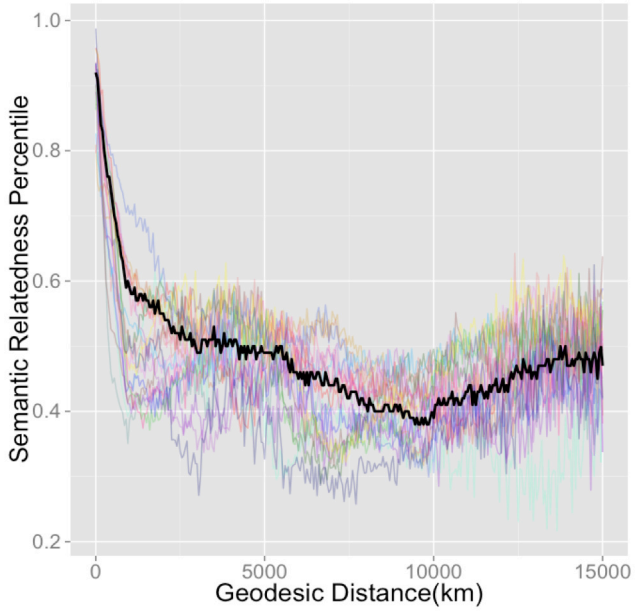

Figure 1. The relationship between distance and semantic relatedness on 20 language editions of Wikipedia using the ensemble SR metric. Each color represents a single language edition and the dark black line is a language speaker-weighted average.

[1] Percentile is assessed across all SR calculates in the sample.

To investigate this phenomenon further, we introduce two metrics *m* and *k*. The former metric, *m*, identifies the distance at which average SR no longer monotonically decreases with distance. If *m* is infinity, TFL would hold for all distance ranges, but if *m* has a finite value, TFL holds only up to *m*. In order to mute the effect of outliers, we define *m* using a window of 5 bins.

Our second metric *k* captures a more global perspective. It describes the first bin at which the average SR is less than the average SR for all entities separated by a distance greater than distance *k* Roughly speaking, *k* is the distance at which there is no longer an *overall* decline in SR.

Tables 2 and 3 show the minimum and maximum values of *m* and *k* for the language editions in our dataset. As can be seen in these tables, *m* and *k* are neither infinity nor zero for any language edition, meaning that there is some distance at which TFL does not hold for all language editions. The lowest *m* value we observed was 850km for the Czech and Catalan Wikipedias. This means that according to our ensemble SR algorithm operating on these language editions, beyond 850km, nearer things are not necessarily on average more related than distant things. Czech and Catalan also had some of the lowest *k* values.

A key trend in Tables 2 and 3 are the differences between the

Rank	Language	*m* value	Rank	Language	*m* value
1	Japanese	2300km	15	Italian	900km
2	Chinese	2250km	18	German	900km
2	Russian	2250km	19	Catalan	850km
…			19	Czech	850km

Table 2. The ranking of *m* values by language edition.

Rank	Language	*k* value	Rank	Language	*k* value
1	Japanese	8850km	17	French	1000km
2	Chinese	8200km	18	Catalan	900km
3	Korean	7950km	18	Dutch	900km
…			20	Czech	700km

Table 3. The ranking of *k* values by language edition.

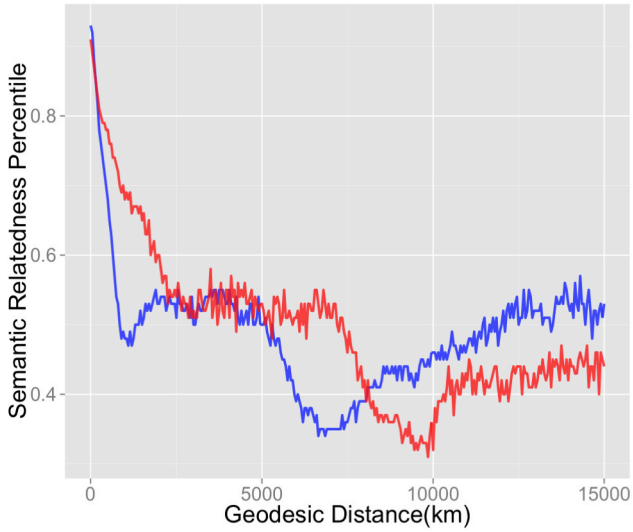

Figure 2. Line plot for the language speaker-weighted average SR of East Asian languages (Red) and that of Western European languages (Blue).

language editions, especially those between East Asian language editions and European language editions. Figure 2 shows a weighted average for these two groups of languages (similar to the black line in Figure 1). It is easy to see that geodesic distance has an uninterrupted negative correlation with semantic relatedness in East Asian language editions for much longer than that for European languages.

Examining the differences between the language editions in more detail, we found that the uneven coverage of geographic entities across the surface of the Earth plays an important factor. Each language edition of Wikipedia is known to exhibit *self-focus bias*, meaning that entities of interest to speakers of a given language are more likely to have a Wikipedia article in the corresponding language edition (among other effects) [8]. For instance, consider the local peak for Western European around 5000km. Our interpretation for this is that, due to the uneven distribution of geographic articles in these language editions, a large portion of geographic entities in this range are located at either Western Europe and U.S. East Coast, which have an extensive shared history and many existing topical relationships (e.g. current events, similar popular culture interests).

4.1 Alternative SR Algorithms

We close our results section with an evaluation of TFL using SR algorithms other than the ensemble algorithm. Even though the ensemble algorithm is the most accurate, each component SR algorithm uses different Wikipedia-based knowledge resources and thus understands different types of relationships [6]. As such, it is useful to examine TFL using each SR measure individually. We repeated above experiment with 5 other SR algorithms described in Table 1. Our result reveals that even though these algorithms utilize different aspect of Wikipedia data, all algorithms agree about the key relationship: at shorter distances, near entities are more related than distance entities, but this breaks down as distances increase.

5. DISCUSSION

While we have discussed and interpreted individual results in the preceding section, here we reflect on a number of important issues we have not yet had an opportunity to address. First, we have assumed a *continuous* interpretation of "near" and "distant" in TFL. Another interpretation – albeit one that not common – is a *discrete* one, in which there are a class of entities that are considered to be "near" and another class considered to be "distant". In this interpretation, the distance that separates near things and distance things can be set to any single value.

To assess whether TFL holds under this discrete definition, we calculated whether there was any geodesic distance separating "near things" and "distant things" in which "near things" were never on average less related than distant things. We found that even with this very liberal definition of "near", for 18 of the 20 language editions there is a geodesic distance at which "near" things are less related than "distant" things, providing additional support TFL only holding within a certain distance range.

Finally, this work has several limitations that we are seeking to address in future research. Most seriously, because we only looked at large language editions, we were not able to include the world knowledge of important cultures around the world, namely those based in sub-Saharan Africa. We hope to look at smaller language editions in future work.

6. CONCLUSION

In this paper, we leveraged recent advances in the domain of natural language processing to perform the first robust evaluation of Tobler's First Law using domain-neutral geographic entities. This evaluation revealed new insight about TFL, a theory that is central to all domains that rely heavily on geographic information. Namely, our findings demonstrate that while as TFL suggests, near things are on average more related than distant things, this is only consistently true within a certain distance range. Beyond that distance, TFL no longer holds.

7. REFERENCES

[1] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M. and Gergle, D. 2012. Omnipedia: Bridging the Wikipedia language gap. *CHI '12*.

[2] Bjorholm, S., et al. 2008. To what extent does Tobler's 1st law of geography apply to macroecology? A case study using American palms (Arecaceae). *BMC ecology*. 8, 1 (2008), 11.

[3] Budanitsky, A. and Hirst, G. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*. 32, 1 (2006), 13–47.

[4] Finkelstein, L. et al., E. 2002. Placing Search in Context: The Concept Revisited. *ACM TOIS*. 20, 1 (2002), 116–131.

[5] Global Administrative Areas: http://www.gadm.org. Accessed: 2014-07-01.

[6] Hecht, B., et al. Explanatory semantic relatedness and explicit spatialization for exploratory search. *SIGIR '12*.

[7] Hecht, B. and Gergle, D. A Beginner's Guide to Geographic Virtual Communities Research. *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena*. IGI Global.

[8] Hecht, B. and Gergle, D. 2009. Measuring self-focus bias in community-maintained knowledge repositories. *C&T '09*.

[9] Hecht, B. and Gergle, D. 2010. On The "Localness" of User-Generated Content. *CSCW '10*.

[10] Hecht, B. and Gergle, D. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. *CHI '10*.

[11] Hecht, B. and Moxley, E. 2009. Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. *COSIT '09*.

[12] Lewis, M.P. ed. 2009. Ethnologue: Languages of the World, 16th Edition. SIL International.

[13] O'loughlin, J. and Anselin, L. 1991. Bringing geography back to the study of international relations: Spatial dependence and regional context in Africa, 1966–1978. *International Interactions*. 17, 1 (1991), 29–61.

[14] Ping, J.L., Green, C.J., Zartman, R.E. and Bronson, K.F. 2004. Exploring spatial dependence of cotton yield using global and local autocorrelation statistics. *Field Crops Research*. 89, 2 (2004), 219–236.

[15] Resnick, P. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI* '95.

[16] Sen, S., Li, T.J.-J., WikiBrain Team and Hecht, B. WikiBrain: Democratizing computation on Wikipedia. *WikiSym / OpenSym* '14.

[17] Strube, M. and Ponzetto, S.P. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. *AAAI '06*.

[18] Sui, D.Z. 2004. Tobler's First Law of Geography: A Big Idea for a Small World. *Annals of the Association of American Geographers*. 94, (2004), 269–277.

[19] Tobler, W.R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*. 46, (1970), 234–240.